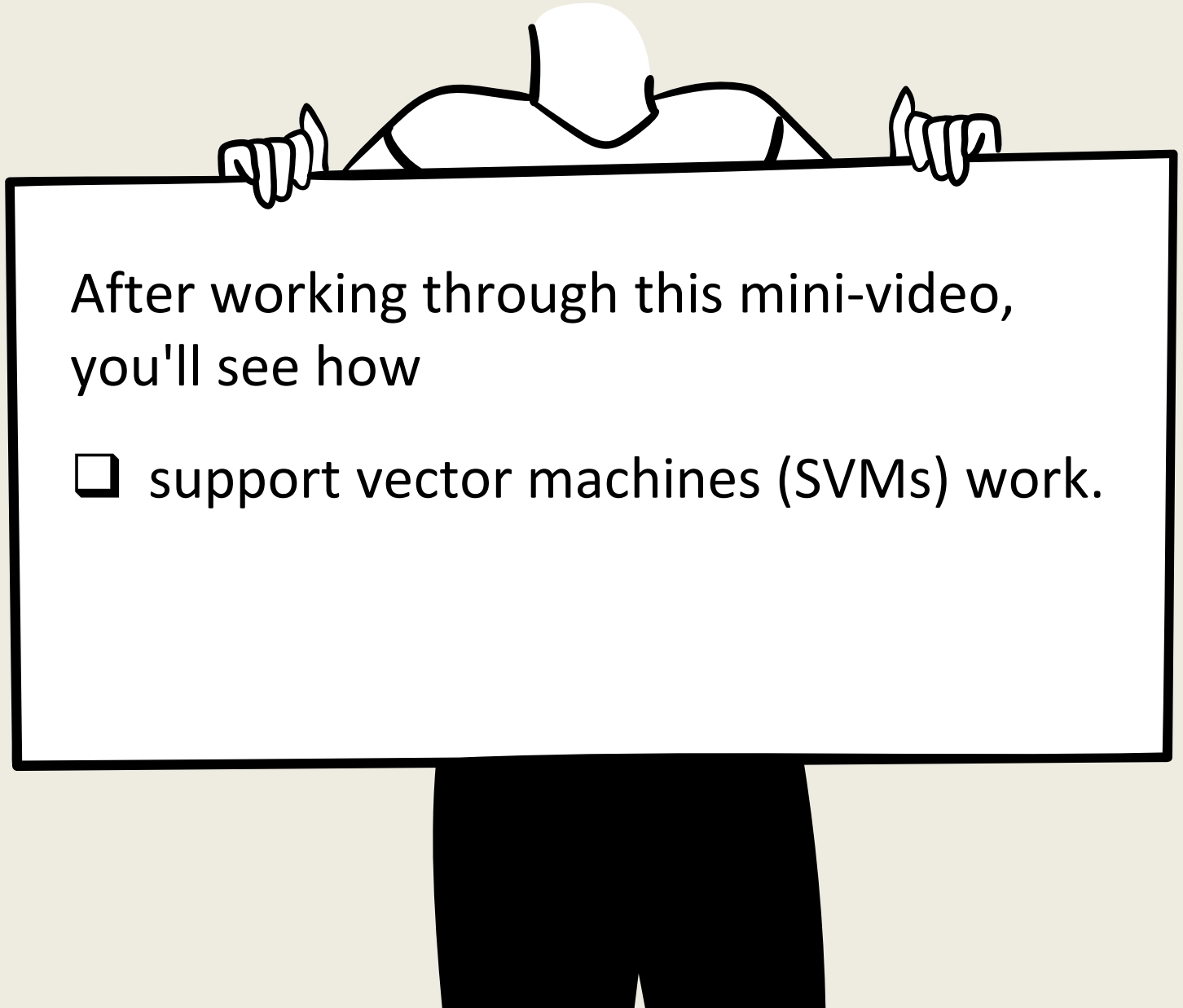
Support vector machines

MACHINE LEARNING

Dr. Temitayo Olugbade

UNIVERSITY
OF SUSSEX

Learning outcome

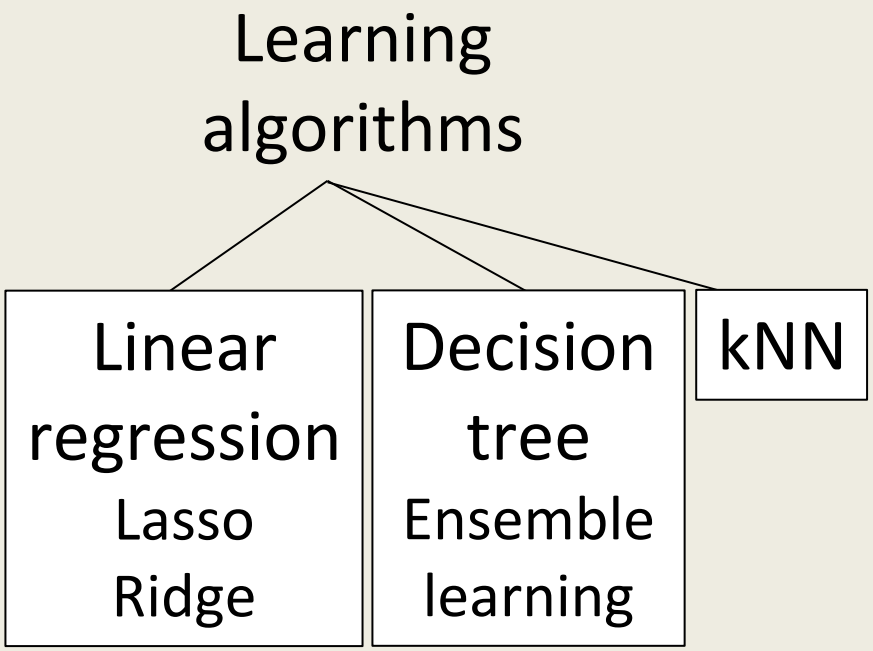


After working through this mini-video, you'll see how

- ❑ support vector machines (SVMs) work.

Recall from Weeks 1 & 2

- The goal of ML is to learn a **model** (from data) that generalizes to unseen data
- Generalizability is trade-off between **weak learning** & **overfitting**.



- A model is characterised by its **learning algorithm**, **parameters**, **loss functions**, **hyperparameters**.

Toy data with numerical categorical labels

labels

$\{y_n\}_{n=1}^{6}, D_y = 1$

features

$\{\boldsymbol{x}_n\}_{n=1}^{6}, D_x = (height, width, channel)$

$channel = 3$ for R,G,B

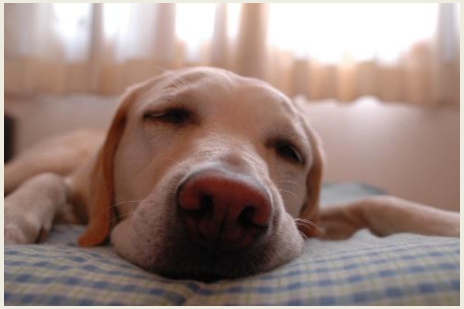
+1

Source: Muhammad Mahdi Karim
https://commons.wikimedia.org/wiki/File:Domestic_cat_felis_catus.jpg

Source: Dimitri Torterat
https://commons.wikimedia.org/wiki/File:Domestic_shorthaired_cat_face.jpg

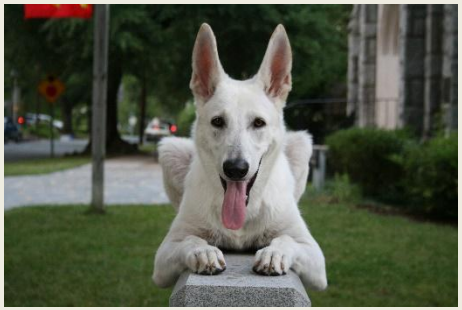
Source: Peter Forster
https://commons.wikimedia.org/wiki/File:Cat_Briciola_with_pretty_and_different_colour_of_eyes.jpg

-1



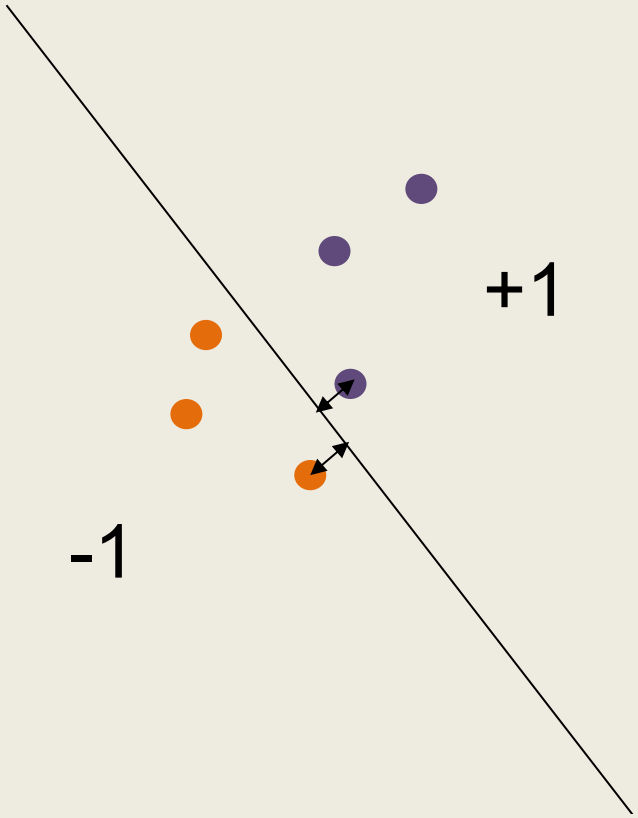
Source: Eugene0126jp
https://en.wikipedia.org/wiki/File:Dog.in.sleep.jpg

Source: Jina Lee
https://commons.wikimedia.org/wiki/File:Pug_dog_nose_face_detail.JPG



Source: IldarSagdejev
https://en.wikipedia.org/wiki/File:2008-06-26_White_German_Shepherd_Dog_Posing_3.jpg

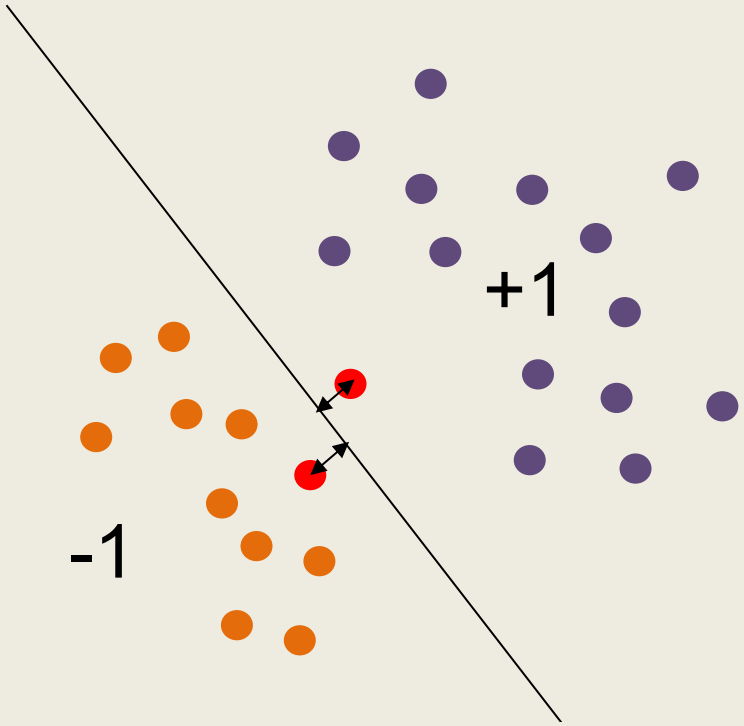
Support vector machine (SVM)



The optimal SVM model is the hyperplane that:

- separates two classes *+1* and *-1*, and
- maximizes its distances (i.e. margin) to them

Support vectors (SVs)



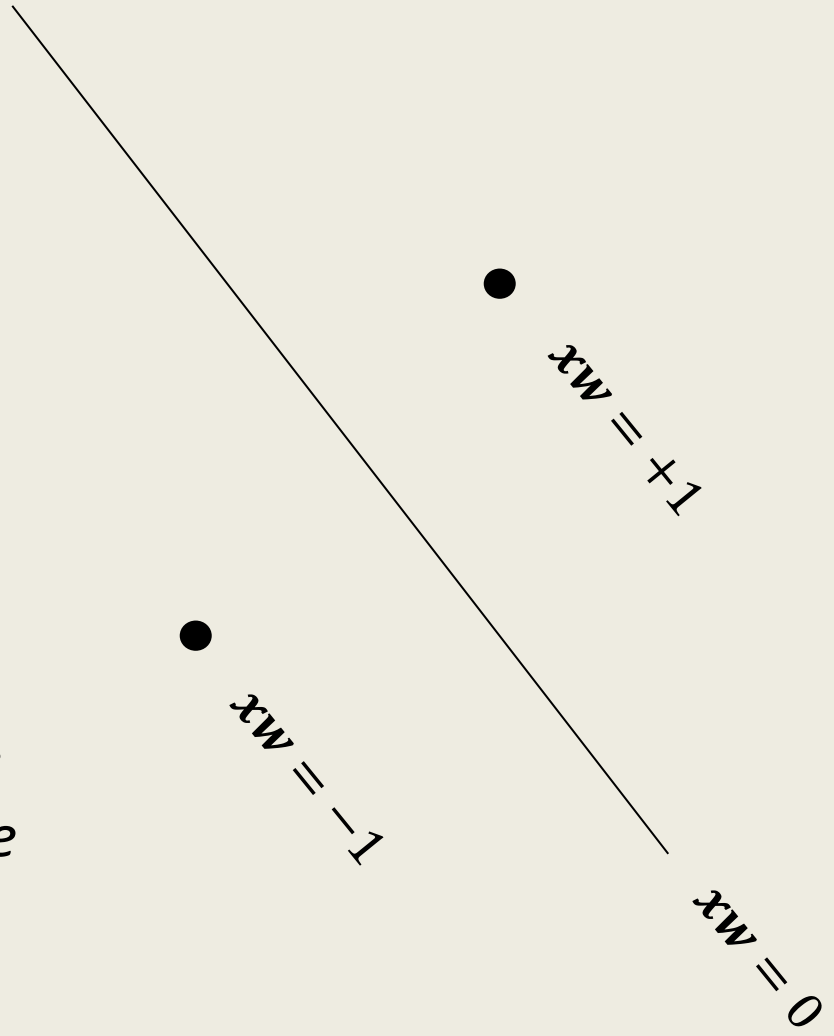
SVs are instances from one class closest to the other class

- so, they determine the maximum margin hyperplane, and
- removing all other data instances except SVs results in the same max margin hyperplane

Maximum margin – distance

The optimal hyperplane is:

- at the maximum margin from both classes
- while still correctly classifying
- *this means that the hyperplane has to be midpoint between the two classes*



Deriving the SVM loss function

The optimal hyperplane is:

1. at maximum margin from both classes
   - i.e. maximising $\frac{2}{\|w\|}$
   - equivalent to minimizing $\frac{\|w\|^2}{2}$
2. while still correctly classifying
   - i.e. such that $y_n(\boldsymbol{x_n w}) \geq 1 \quad \forall n$

see math proof in last slide pages

SVM loss function (primal)

- Thus, the loss/objective function of the SVM is

$$L_{SVM}(\boldsymbol{w}) = -\sum_{n=1}^{N} \beta_n (y_n(\boldsymbol{x}_n \boldsymbol{w}) - 1) + \frac{1}{2}\|\boldsymbol{w}\|^2$$

classification error — *margin to be maximised*

where β_n = Lagrange multipliers (aka coefficients of additional constraints in a function to be optimized)

- This is known as the *primal formulation* of the objective function.

SVM loss function (dual)

- Thus, another formulation of the loss function of the SVM is

$$g(\beta) = \sum_{n=1}^{N} \beta_n - \frac{1}{2} \sum_{n,m=1}^{N} \beta_n \beta_m y_n y_m \boldsymbol{x}_n \boldsymbol{x}_m , \qquad \beta_n \geq 0$$

- This is known as the *dual formulation* of the objective function

see math details in last slide pages derivation from the primal formulation

Finding the optimal β_n and β_m

$$g(\beta) = \sum_{n=1}^{N} \beta_n - \frac{1}{2} \sum_{n,m=1}^{N} \beta_n \beta_m y_n y_m \boldsymbol{x}_n \boldsymbol{x}_m \,, \qquad \beta_n \geq 0$$

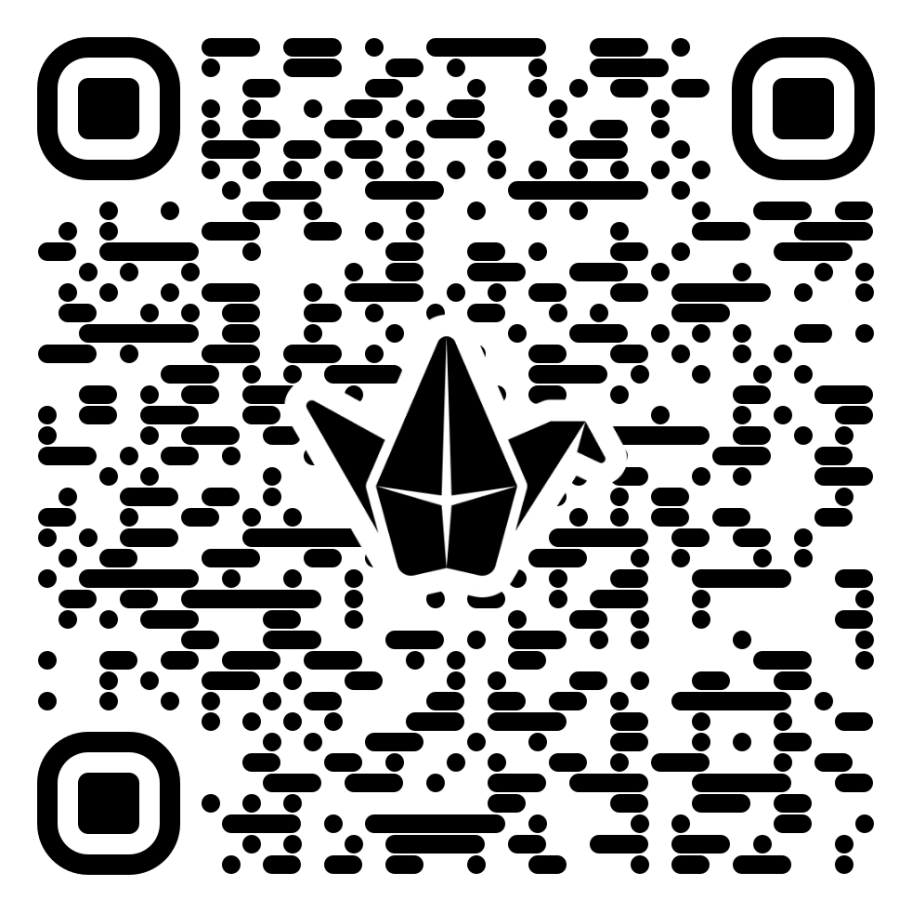
- $\beta_n \beta_m$ makes it a constrained quadratic programming task.
- And it can be solved using quadratic programming solvers

Summary

1. The **SVM** model is a maximum margin hyperplane between two classes.

2. There are two formulations of the loss function that meet this requirement.

Any questions???



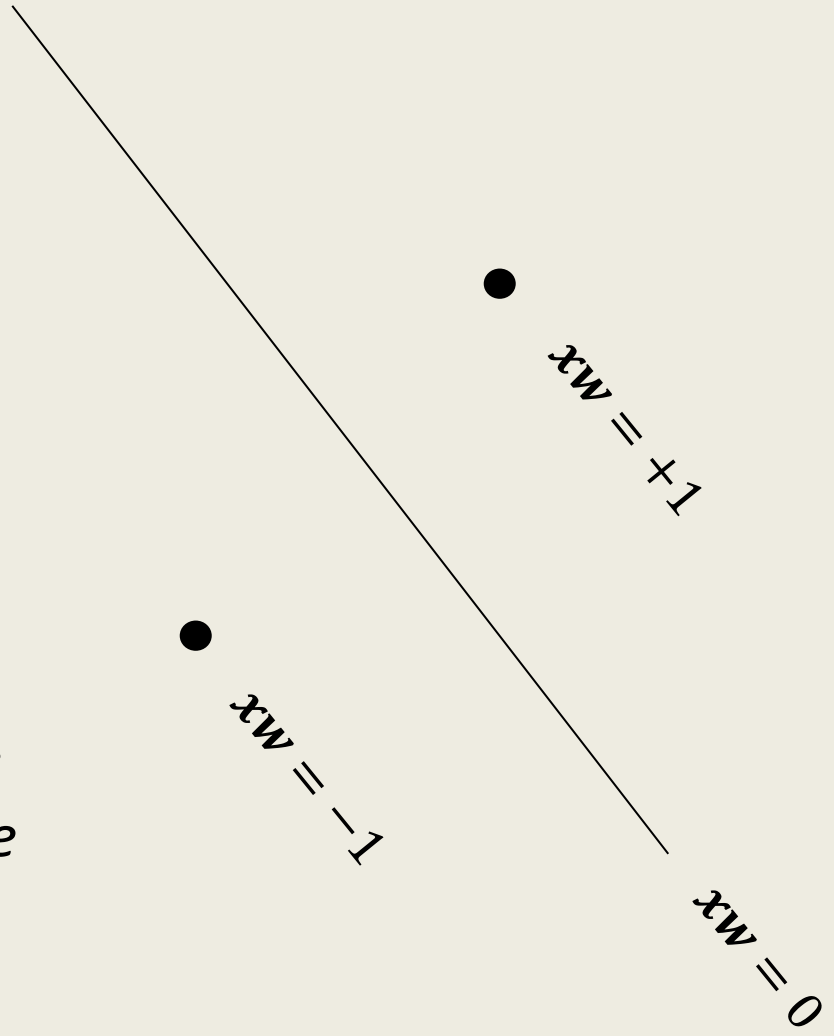
scan the QR code to ask questions

Math details and proofs

Deriving the SVM loss function (0) – MATH

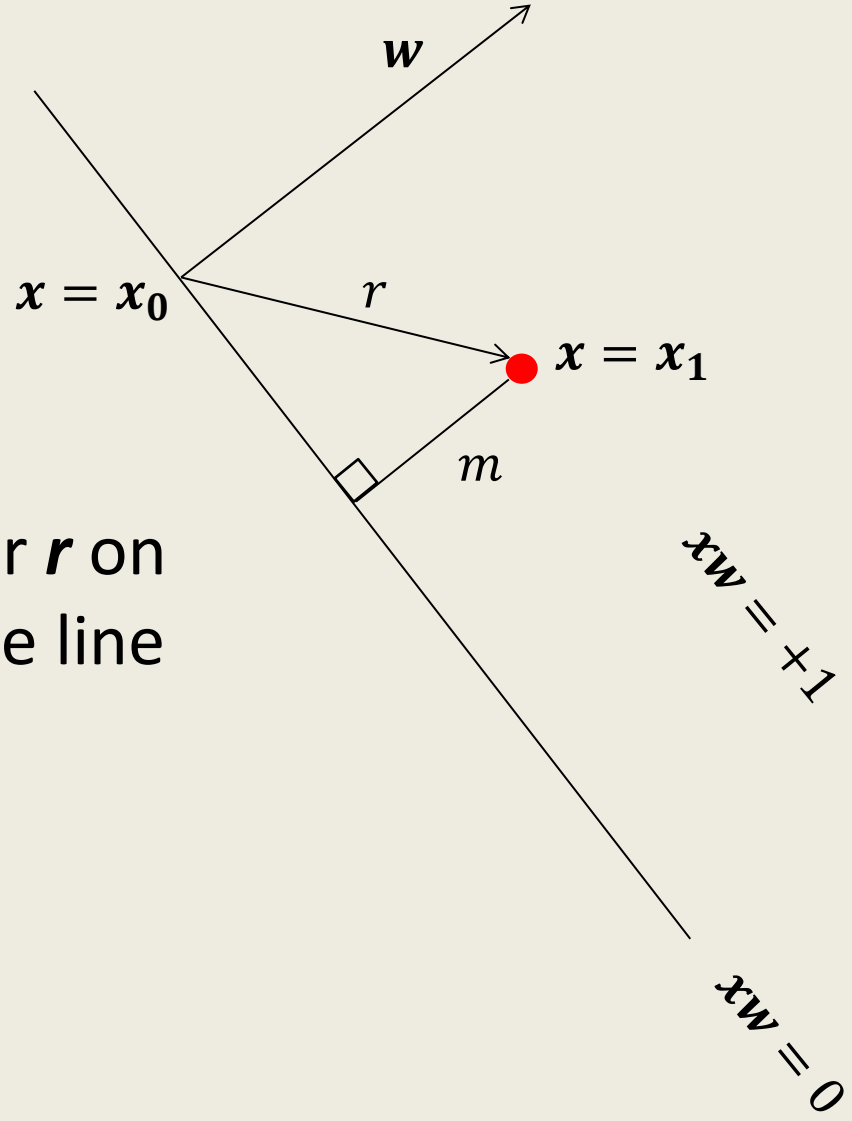
The optimal hyperplane is:

- **at the maximum margin from both classes**
- while still correctly classifying
- *this means that the hyperplane has to be midpoint between the two classes*

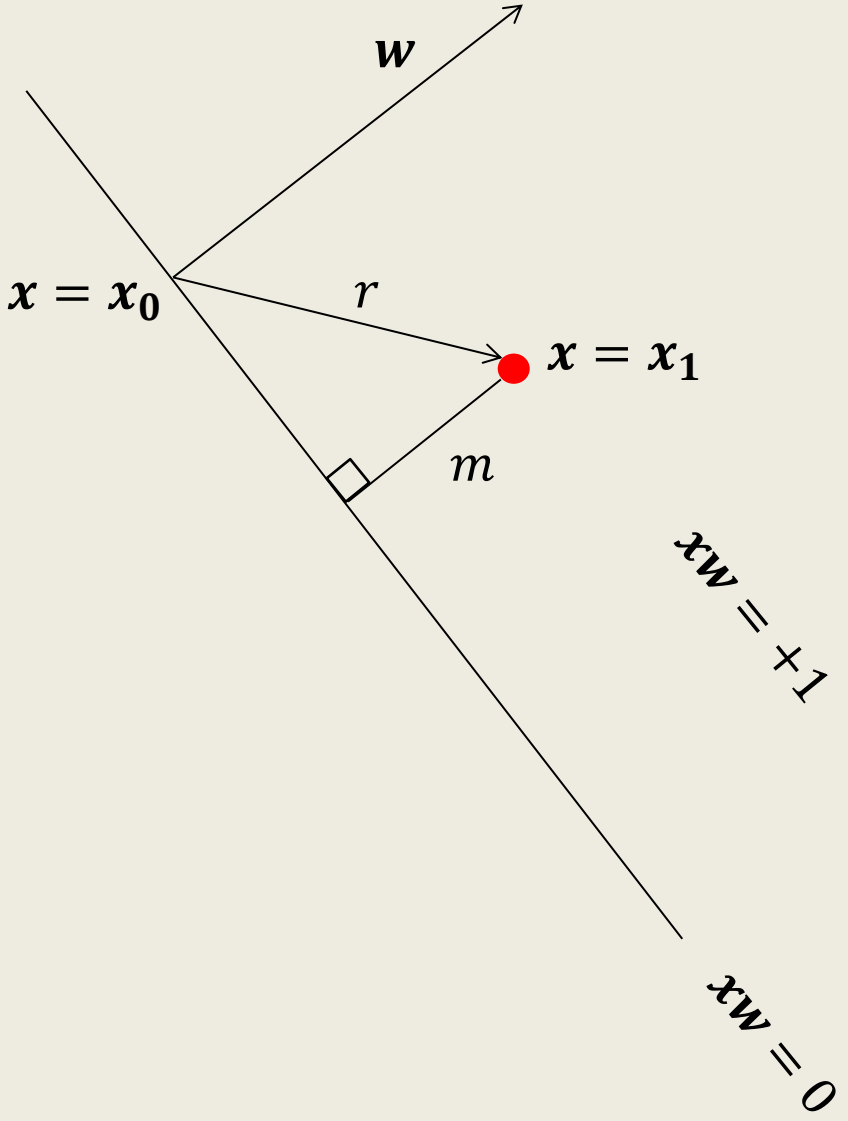


Deriving the SVM loss function (1) – MATH

The shortest distance is the length of projection of vector $\boldsymbol{r}$ on to the normal vector $\boldsymbol{w}$ of the line



Deriving the SVM loss function (2) – MATH



The length of projection of vector ***r*** on to the normal vector ***w*** of the line is

$$\Rightarrow m = \left| \boldsymbol{r} \cdot \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \right|$$

substituting for $r = x_1 - x_0$

$$m = \left| (\boldsymbol{x_1} - \boldsymbol{x_0}) \cdot \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \right|$$

Deriving the SVM loss function (3) – MATH

Remember that:

$$y = \boldsymbol{x_0 w} = 0 \qquad (1)$$

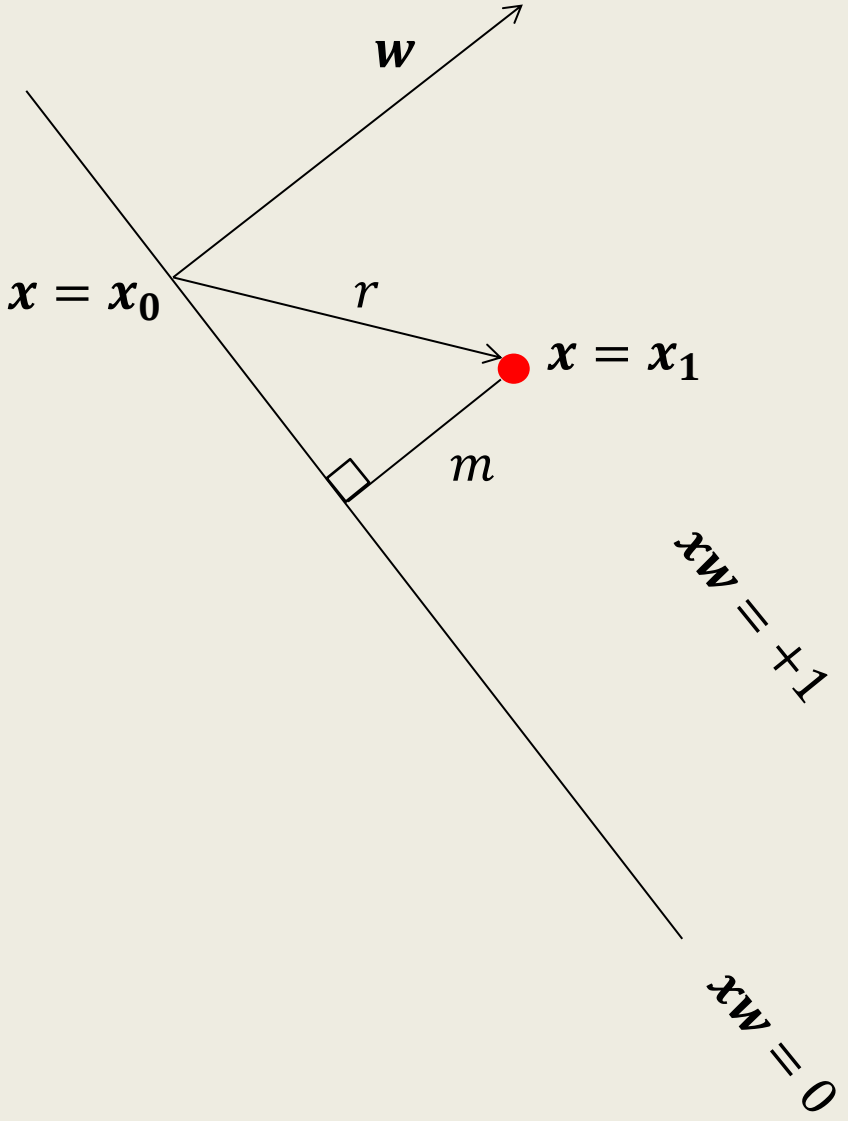
and

$$y = \boldsymbol{x_1 w} = +1 \qquad (2)$$

subtracting (1) from (2)

$$\boldsymbol{x_1 w} - \boldsymbol{x_0 w} = +1 - 0$$

$$\Longrightarrow (\boldsymbol{x_1} - \boldsymbol{x_0})\boldsymbol{w} = +1$$

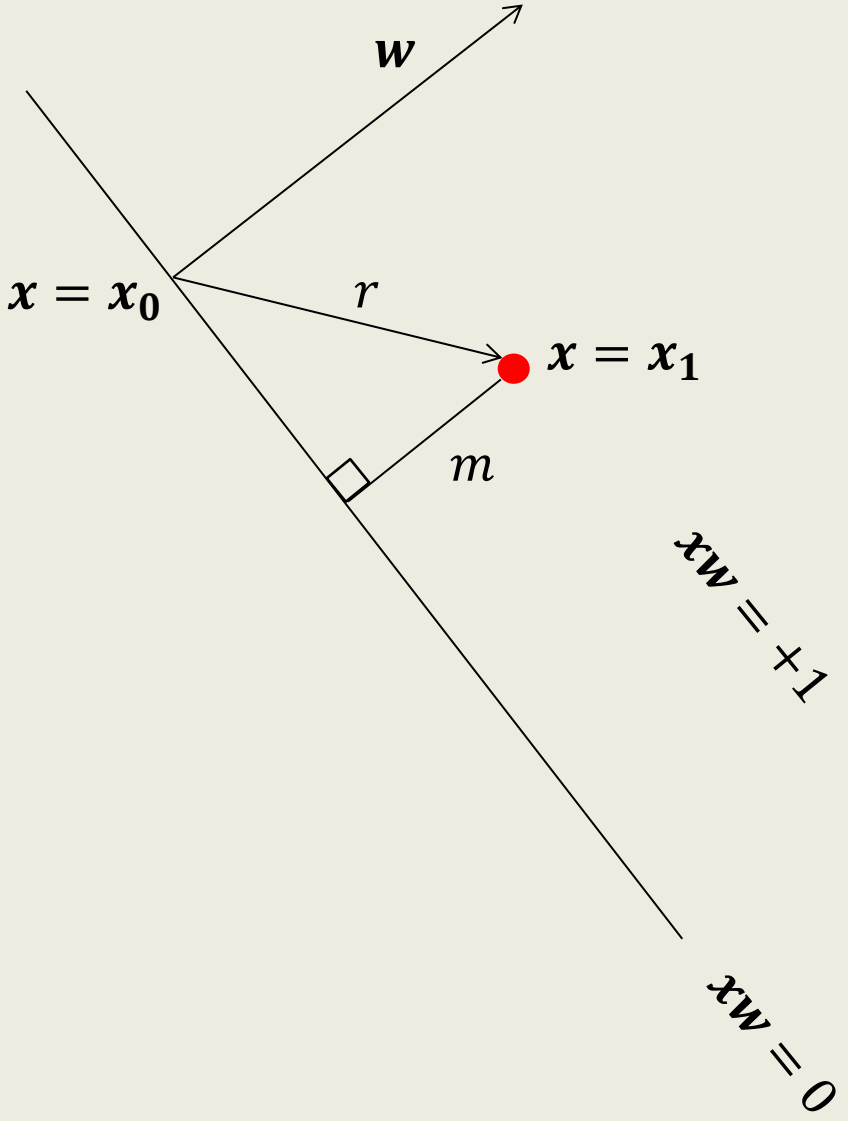


Deriving the SVM loss function (4) – MATH

$$m = \left| (\boldsymbol{x_1} - \boldsymbol{x_0}) \cdot \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \right|$$

substituting for $(\boldsymbol{x_1} - \boldsymbol{x_0})\boldsymbol{w} = 1$

$$\Longrightarrow m = {}^{1}/_{\|w\|}$$



Distance to hyperplane for class -1

m here is also $^{1}/_{\|w\|}$

Try deriving it following the approach used for class *+1*

Re: Deriving SVM loss function (0) – MATH

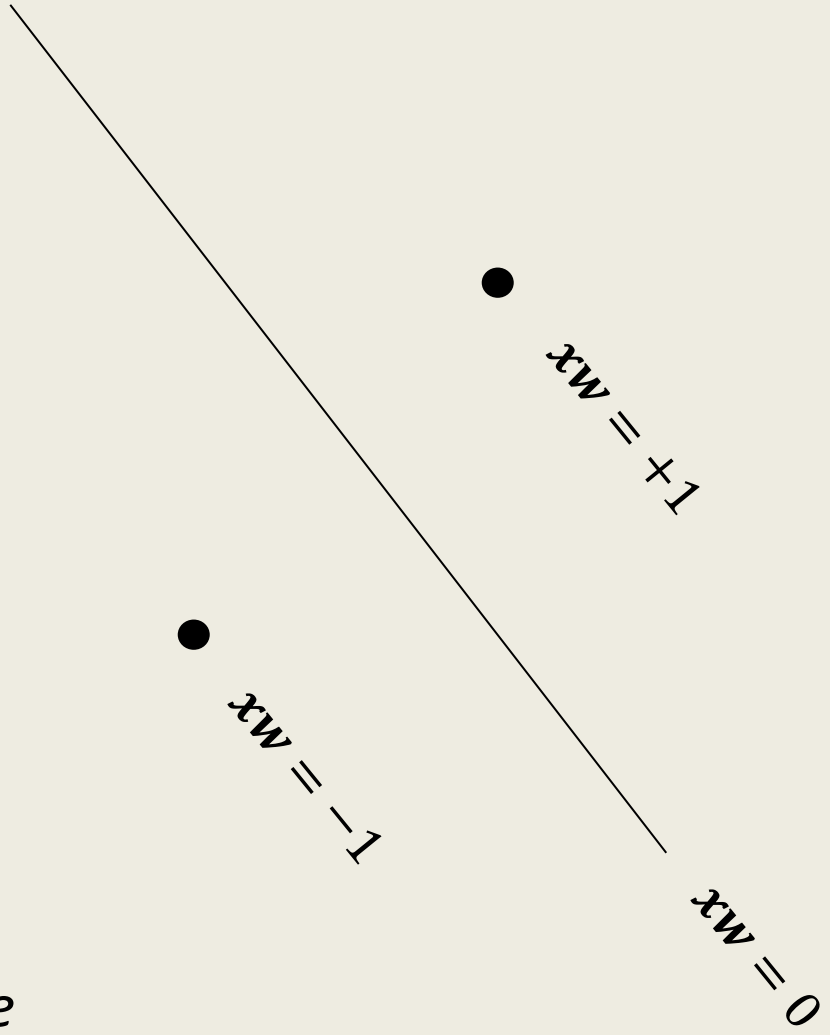
The optimal hyperplane is:

- **at the maximum margin from both classes**

  i.e. maximising $\frac{2}{\|\boldsymbol{w}\|}$

  equivalent to minimizing $\frac{\|\boldsymbol{w}\|^2}{2}$

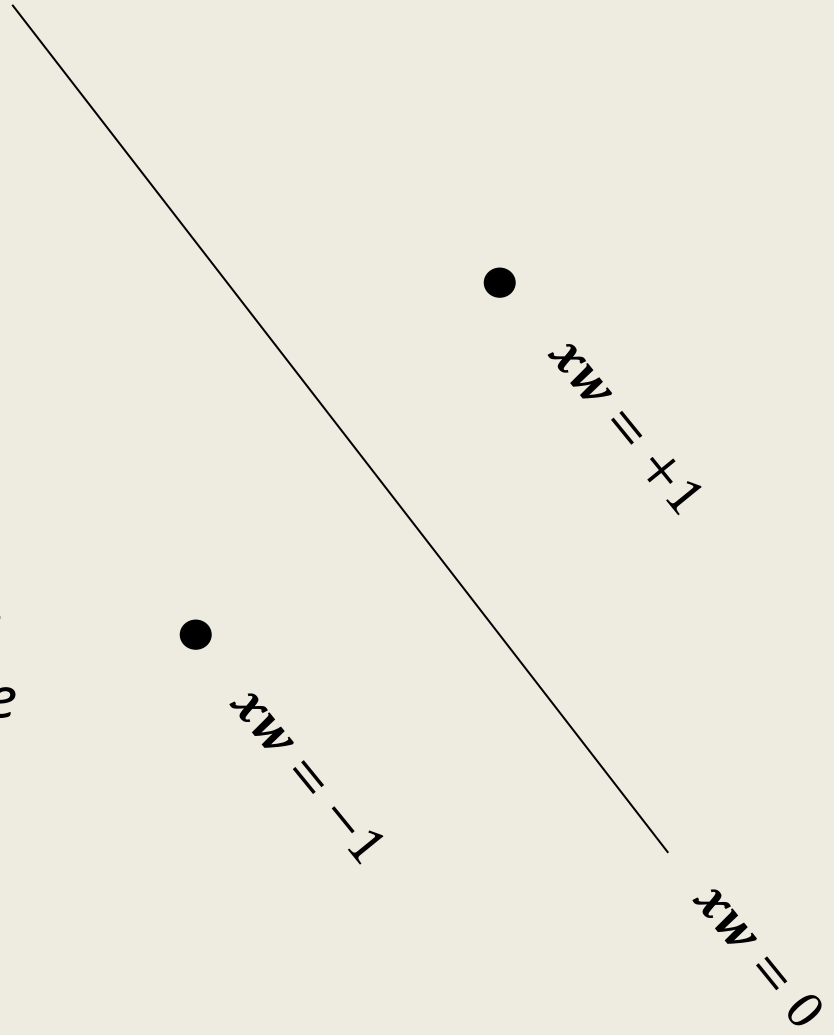
- while still correctly classifying
- *this means that the hyperplane has to be midpoint between the two classes*



Re: Deriving SVM loss function (0) – MATH

The optimal hyperplane is:

- at the maximum margin from both classes
- **while still correctly classifying**
- *this means that the hyperplane has to be midpoint between the two classes*



Deriving the SVM loss function (5) – MATH

Correct classification requires that predicted class ($\hat{y}_n = \boldsymbol{x_n w}^*$) and true class ($y_n$) are on the same side of the hyperplane

$$\Longrightarrow y_n(\boldsymbol{x_n w}^*) > 0$$

since this implies that $y_n(\boldsymbol{x_n w})$ is at least some value γ

$$y_n(\boldsymbol{x_n w}^*) \geq \gamma$$

dividing through by γ

$$y_n\left(\boldsymbol{x_n}\frac{\boldsymbol{w}^*}{\gamma}\right) \geq 1$$

let $\boldsymbol{w} = \frac{\boldsymbol{w}^*}{\gamma}$

$$\Longrightarrow y_n(\boldsymbol{x_n w}) \geq 1$$

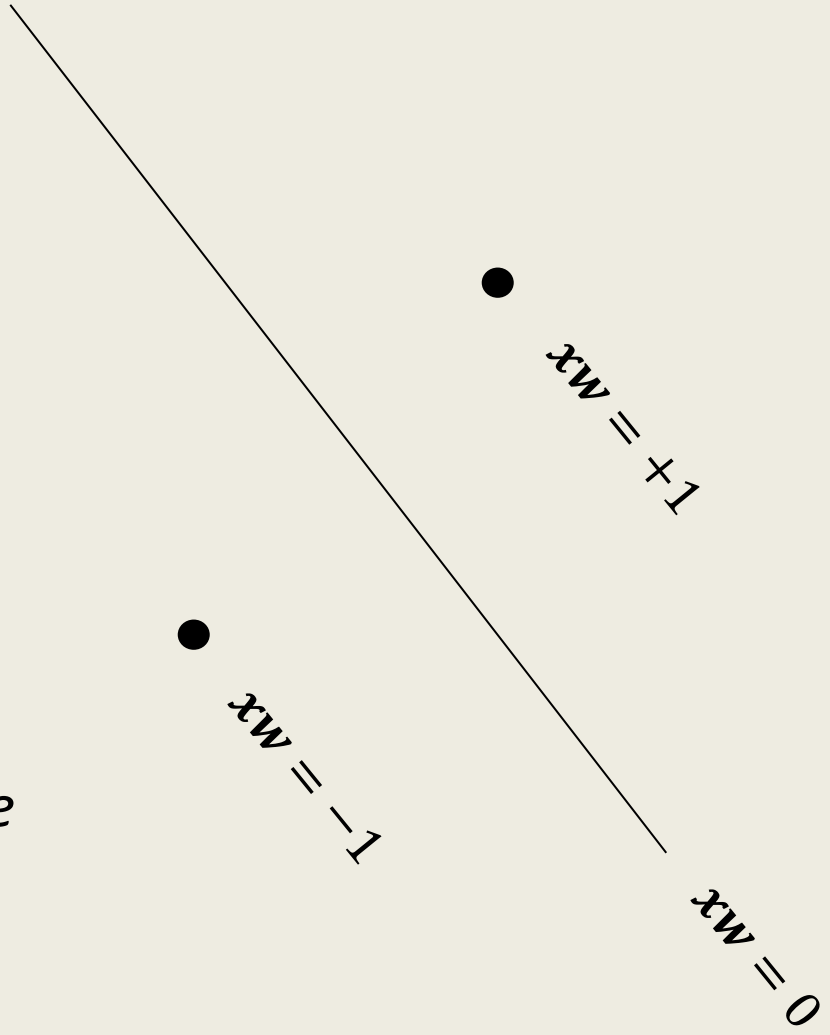
Re: Deriving SVM loss function (0) – MATH

The optimal hyperplane is:

- at the maximum margin from both classes
- **while still correctly classifying**

i.e. such that $y_n(\boldsymbol{x_n w}) \geq 1 \quad \forall n$

- *this means that the hyperplane has to be midpoint between the two classes*



Deriving the SVM loss function (6) – MATH

The optimal hyperplane is:

1. at maximum margin from both classes
   - i.e. maximising $\frac{2}{\|\boldsymbol{w}\|}$
   - equivalent to minimizing $\frac{\|\boldsymbol{w}\|^2}{2}$
2. while still correctly classifying
   - i.e. such that $y_n(\boldsymbol{x_n w}) \geq 1 \quad \forall n$

Deriving the dual formulation (1) – MATH

Recall that minimum of a loss function is at $\frac{dL(\boldsymbol{w})}{d\boldsymbol{w}} = 0$

substituting for $L(\boldsymbol{w}) = -\sum_{n=1}^{N} \beta_n (y_n(\boldsymbol{x}_n \boldsymbol{w}) - 1) + \frac{1}{2}\|\boldsymbol{w}\|^2$, *i.e. the primal formulation*

$$\frac{d\left(-\sum_{n=1}^{N} \beta_n (y_n(\boldsymbol{x}_n \boldsymbol{w}) - 1) + \frac{1}{2}\|\boldsymbol{w}\|^2\right)}{d\boldsymbol{w}} = 0$$

applying rules of differentiation

$$-\sum_{n=1}^{N} \beta_n\big(y_n(\boldsymbol{x}_n)\big) + \boldsymbol{w} = 0$$

making $\boldsymbol{w}$ *the subject of the formula*

$$\Longrightarrow \boldsymbol{w} = \sum_{n=1}^{N} \beta_n\big(y_n(\boldsymbol{x}_n)\big)$$

Deriving the dual formulation (2) – MATH

$$\boldsymbol{w} = \sum_{n=1}^{N} \beta_n \big(y_n(\boldsymbol{x}_n)\big)$$

and

$$L_{SVM}(\boldsymbol{w}) = -\sum_{n=1}^{N} \beta\left(y_n(\boldsymbol{x}_n \boldsymbol{w}) - 1\right) + \frac{1}{2}\|\boldsymbol{w}\|^2$$

substituting for $\boldsymbol{w}$ in $L_{SVM}(\boldsymbol{w})$

$$\Rightarrow -\sum_{n=1}^{N} \beta_n \left(y_n \left(\boldsymbol{x}_n \sum_{n=1}^{N} \beta_n \big(y_n(\boldsymbol{x}_n)\big) \right) - 1 \right) + \frac{1}{2} \left(\sum_{n=1}^{N} \beta_n \big(y_n(\boldsymbol{x}_n)\big) \right)^2$$

Deriving the dual formulation (3) – MATH

referring to this as $g(\beta)$ since it is no longer a function of $\boldsymbol{w}$

$$g(\beta) = -\sum_{n=1}^{N} \beta_n \left(y_n \left(\boldsymbol{x}_n \sum_{n=1}^{N} \beta_n \left(y_n(\boldsymbol{x}_n) \right) \right) - 1 \right) + \frac{1}{2} \left(\sum_{n=1}^{N} \beta_n \left(y_n(\boldsymbol{x}_n) \right) \right)^2$$

on expanding each term

$$g(\beta) = -\sum_{n,m=1}^{N} \beta_n \beta_m y_n y_m \boldsymbol{x}_n \boldsymbol{x}_m + \sum_{n=1}^{N} \beta_n + \frac{1}{2} \sum_{n,m=1}^{N} \beta_n \beta_m y_n y_m \boldsymbol{x}_n \boldsymbol{x}_m$$

collecting like terms

$$g(\beta) = \sum_{n=1}^{N} \beta_n - \frac{1}{2} \sum_{n,m=1}^{N} \beta_n \beta_m y_n y_m \boldsymbol{x}_n \boldsymbol{x}_m \,, \qquad \beta_n \geq 0$$