NIO

Exercise 13: SVMs, Exam Hints

Duc Duy Pham, M.Sc.

Raum: BC 410
Tel.: 0203-379-3734
Email: duc.duy.pham@uni-due.de

Revision of Lecture

- What is the general idea of a Support Vector Machine?

Revision of Lecture

- What is the general idea of a Support Vector Machine?
  1) Find optimal hyperplane to linearly separate two classes
  2) Exactly the same as the perceptron
  3) Feature reduction
  4) Find optimal hyperplane for regression problems

Revision of Lecture

- What is the general idea of a Support Vector Machine?
  1) Find optimal hyperplane to linearly separate two classes
  2) Exactly the same as the perceptron
  3) Feature reduction
  4) Find optimal hyperplane for regression problems

A: 1, 2, 3	B: all
C: 1	D: 1,4

Revision of Lecture

- What is the general idea of a Support Vector Machine?
    1) Find optimal hyperplane to linearly separate two classes
    2) Exactly the same as the perceptron
    3) Feature reduction
    4) Find optimal hyperplane for regression problems

A: 1, 2, 3	B: all
C: 1	D: 1,4

Revision of Lecture

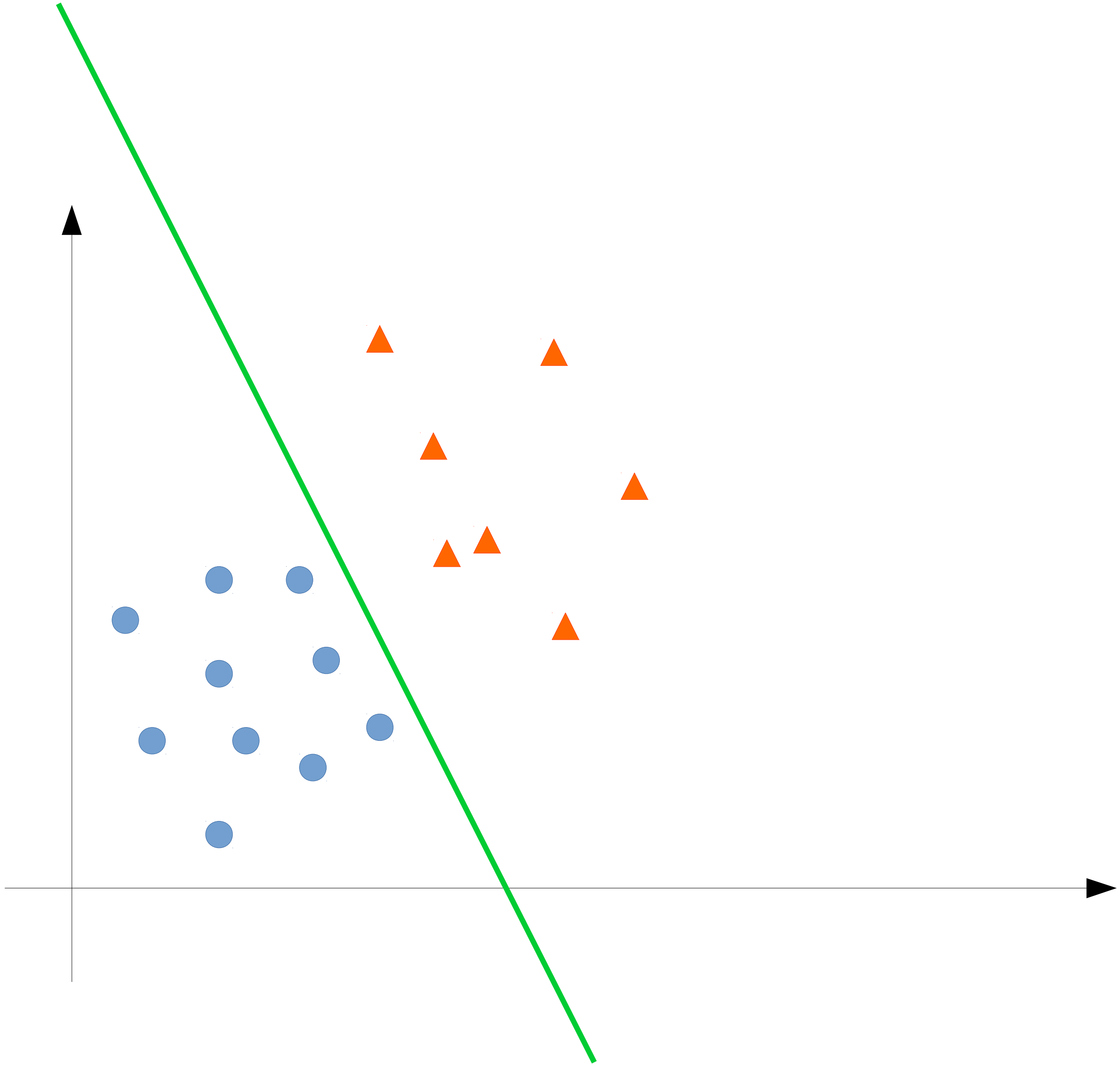
- What is the general idea of a Support Vector Machine?

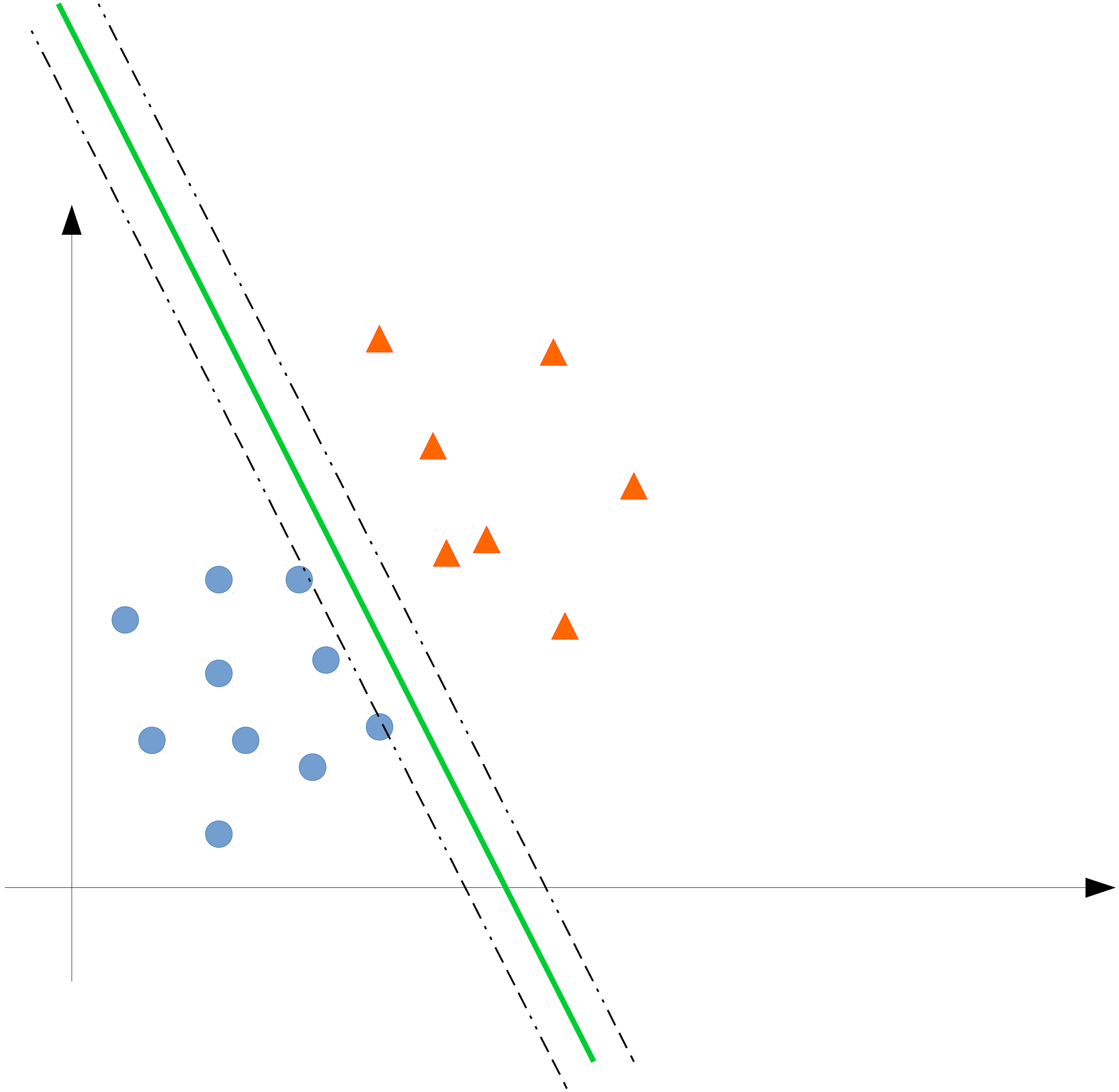
Revision of Lecture

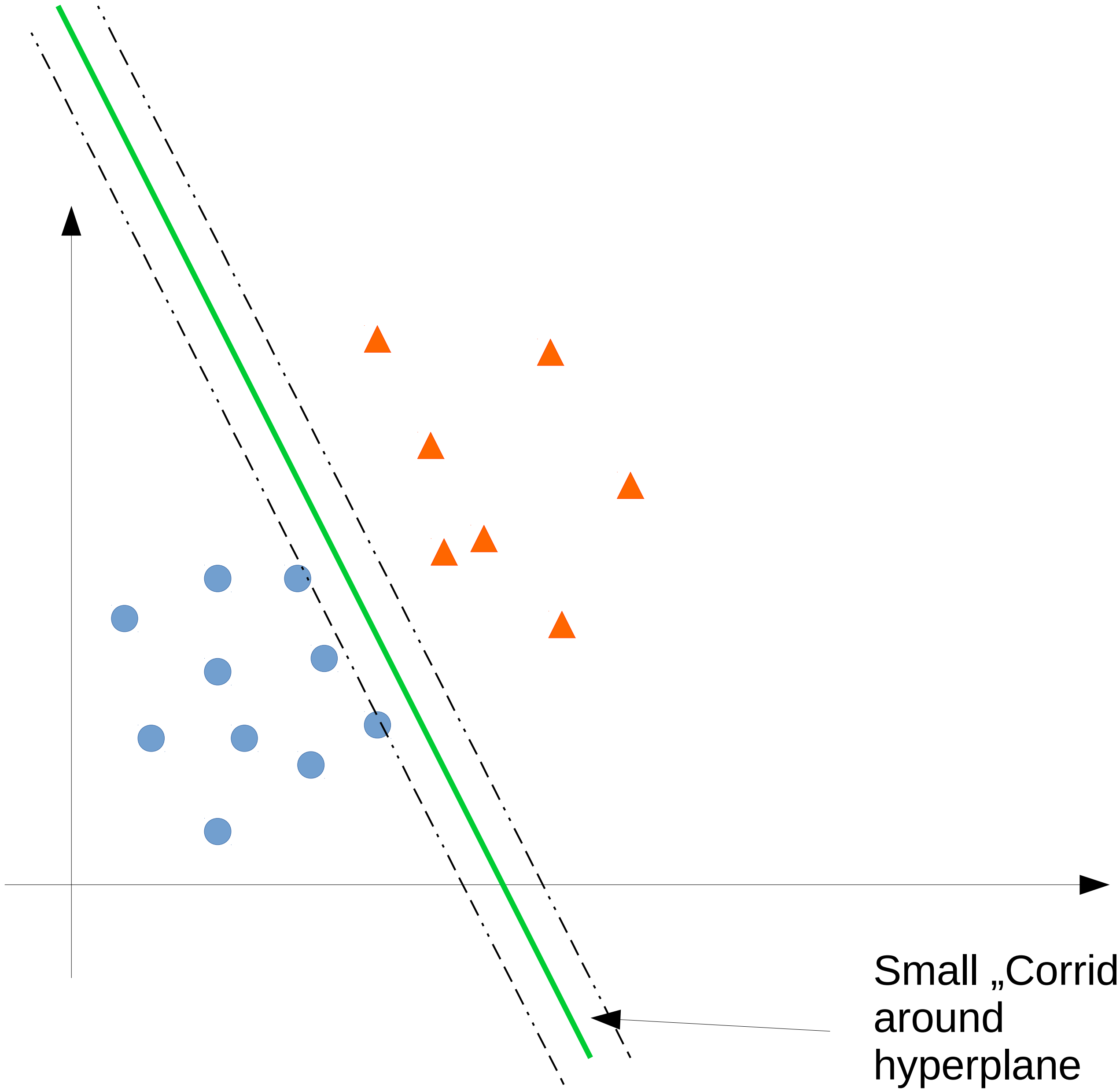
- What is the general idea of a Support Vector Machine?
  - Use „optimally“ positioned hyperplane to linearly separate two classes
    (similar general idea as perceptron!)

Revision of Lecture

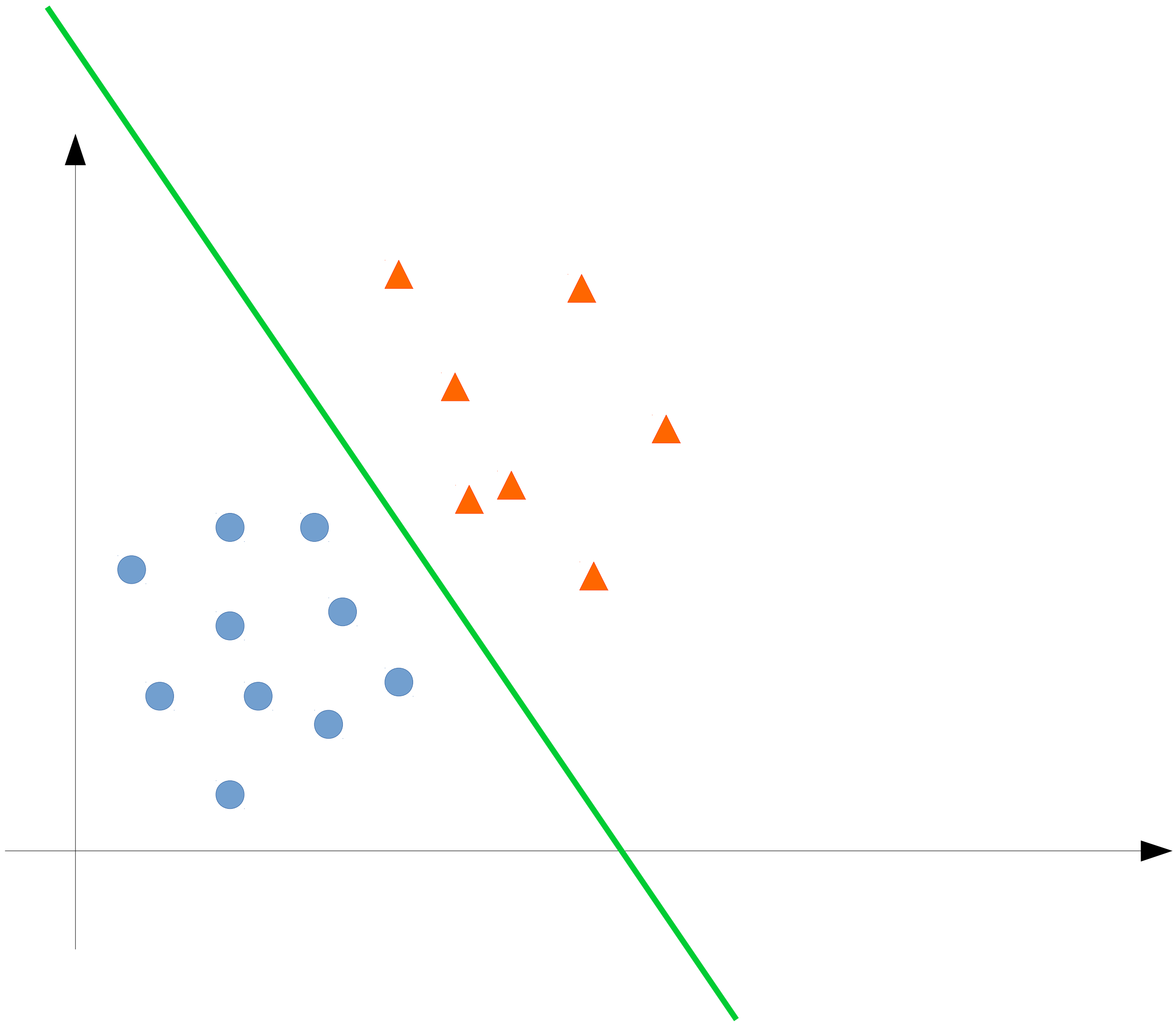
- What is the general idea of a Support Vector Machine?
  - Use „optimally“ positioned hyperplane to linearly separate two classes
    (similar general idea as perceptron!)
  - Maximize distance between hyperplane and both classes
    (maximize width of sample free „corridor“ [=functional margin] around hyperplane)

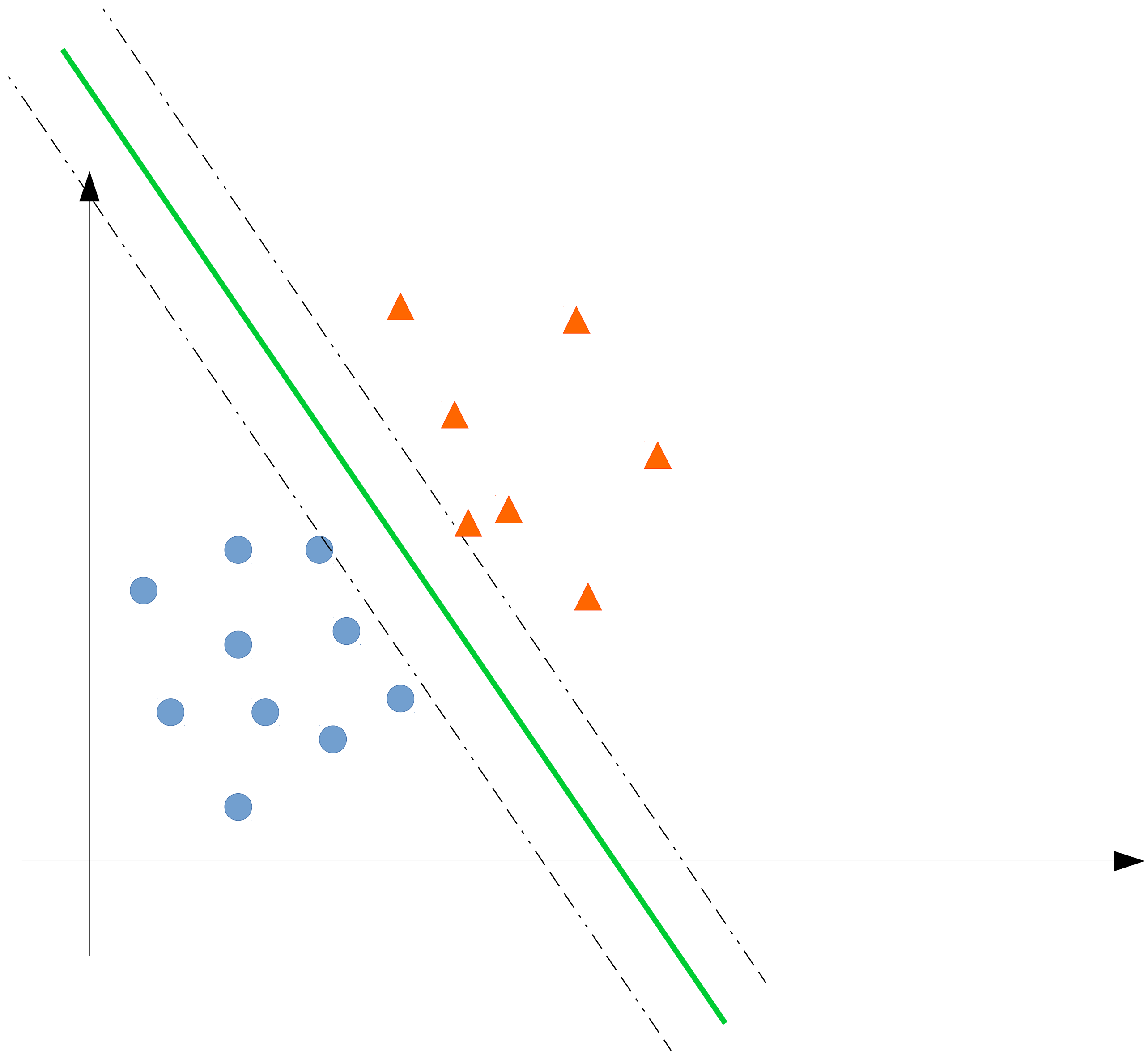


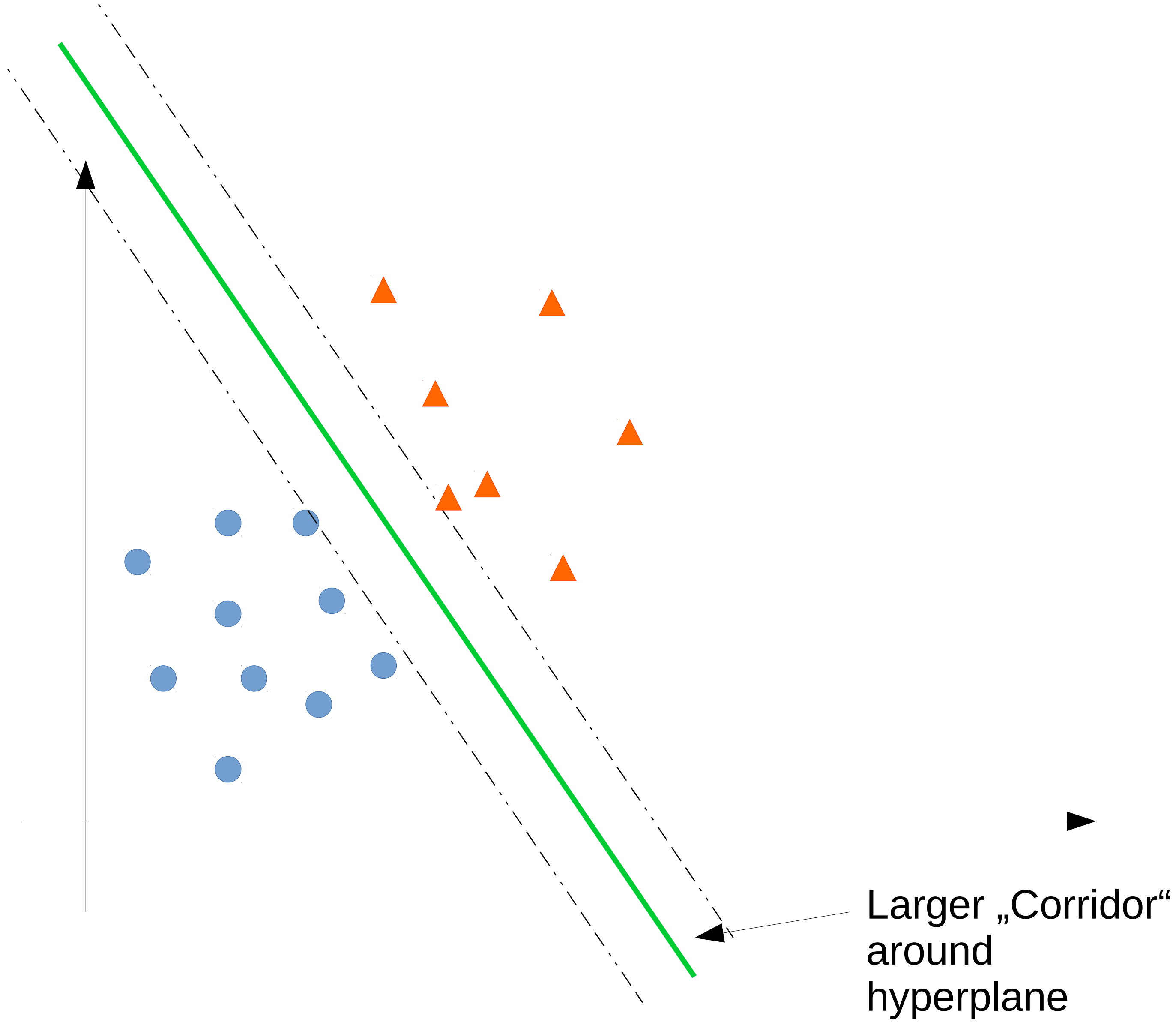




Small „Corridor“
around
hyperplane

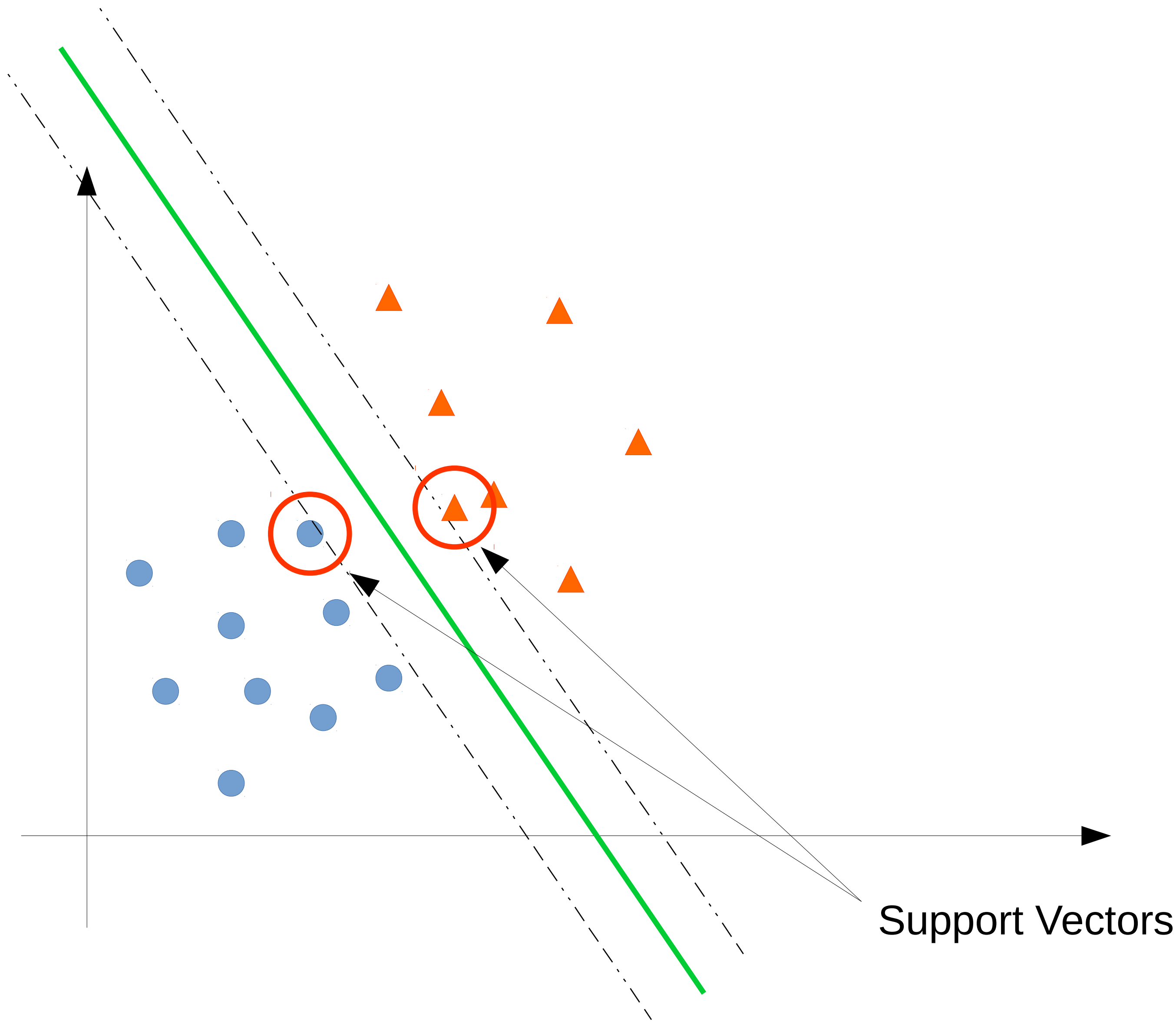


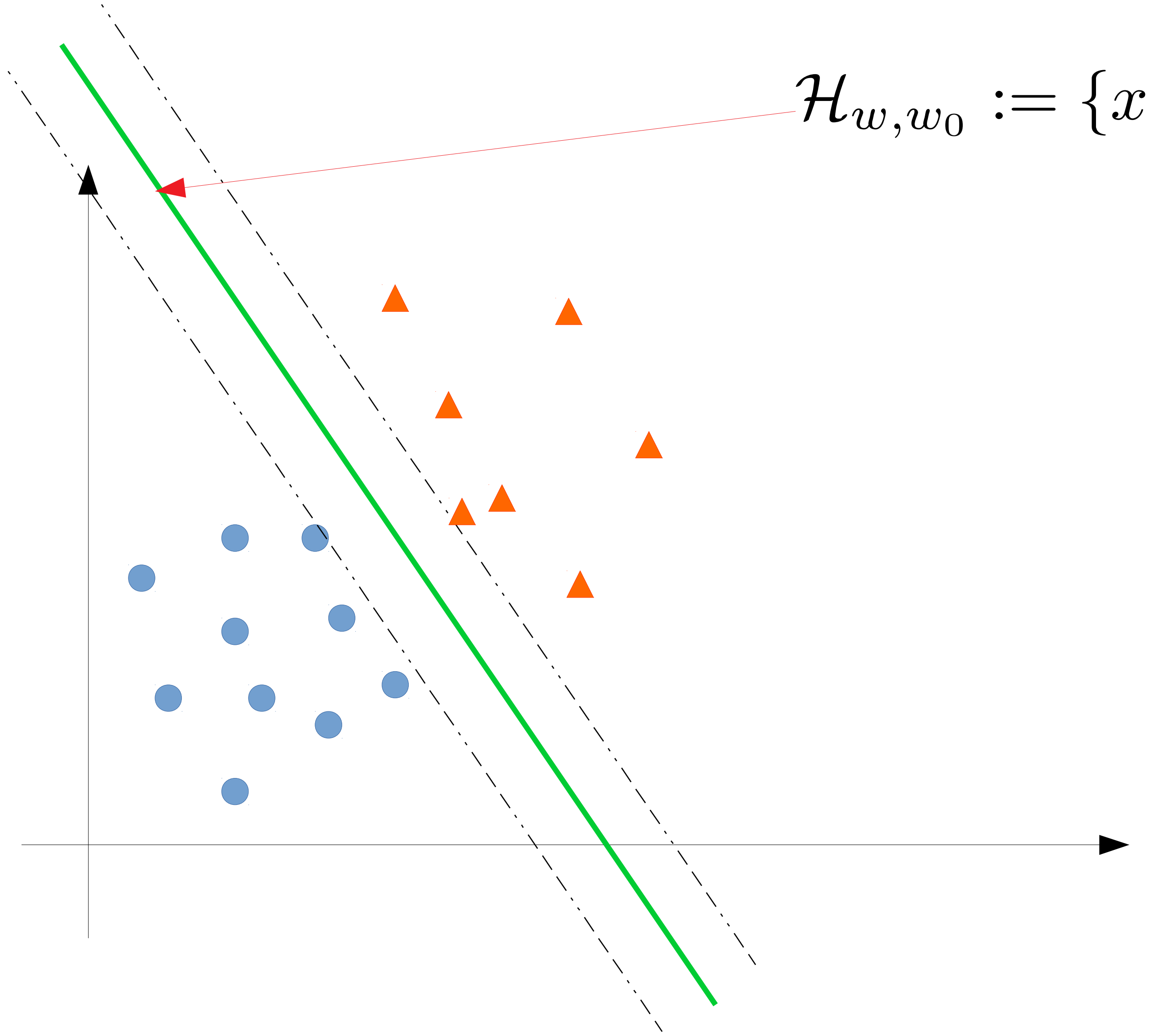



Larger „Corridor“
around
hyperplane

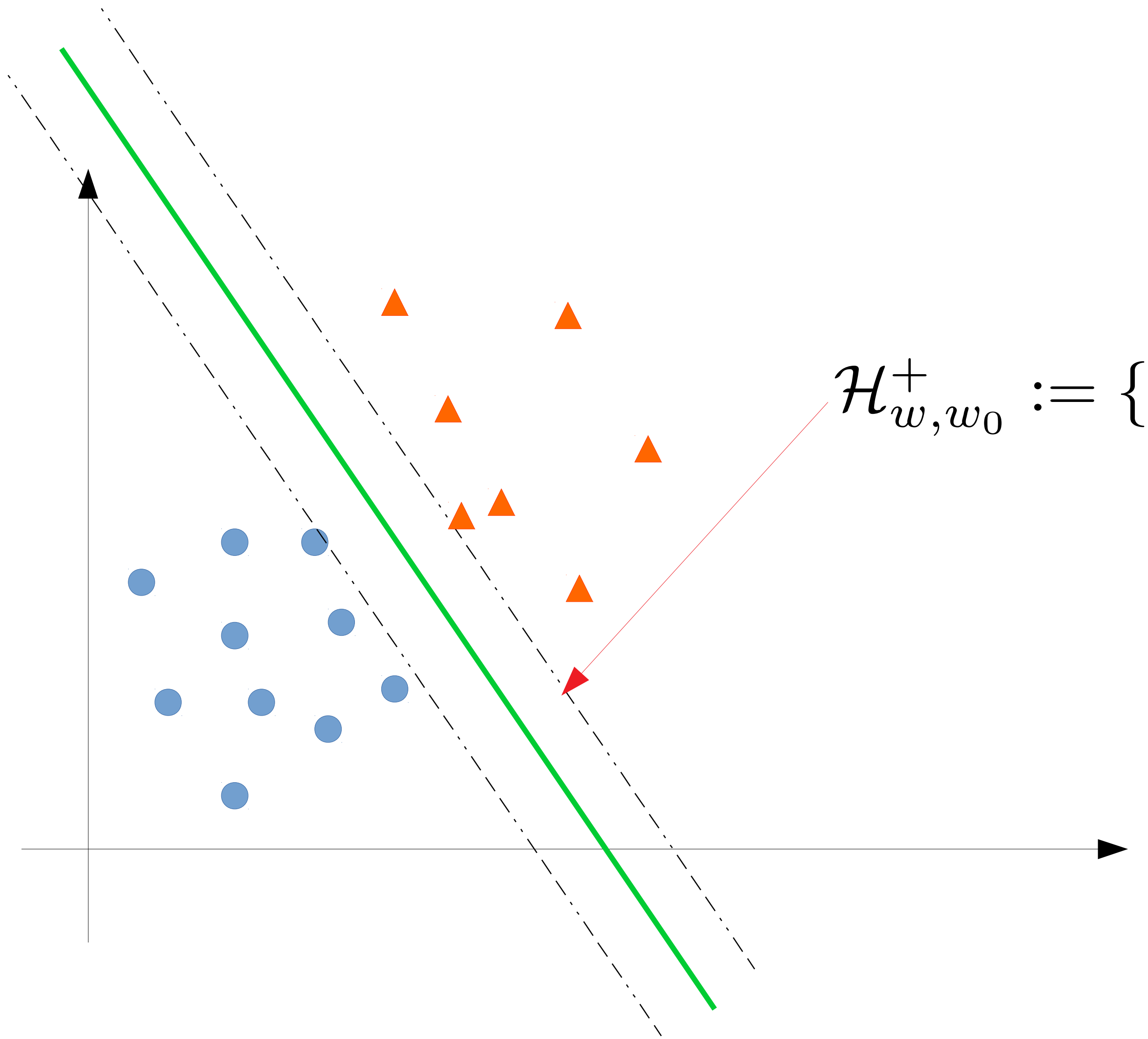
Revision of Lecture

- What are Support Vectors?
  - Support Vectors are the vectors that are on the margin of the corridor (functional margin)
  - They are the support/ the foundation of the margin hyperplanes
  - => functional margin width is restricted by Support Vectors


Support Vectors

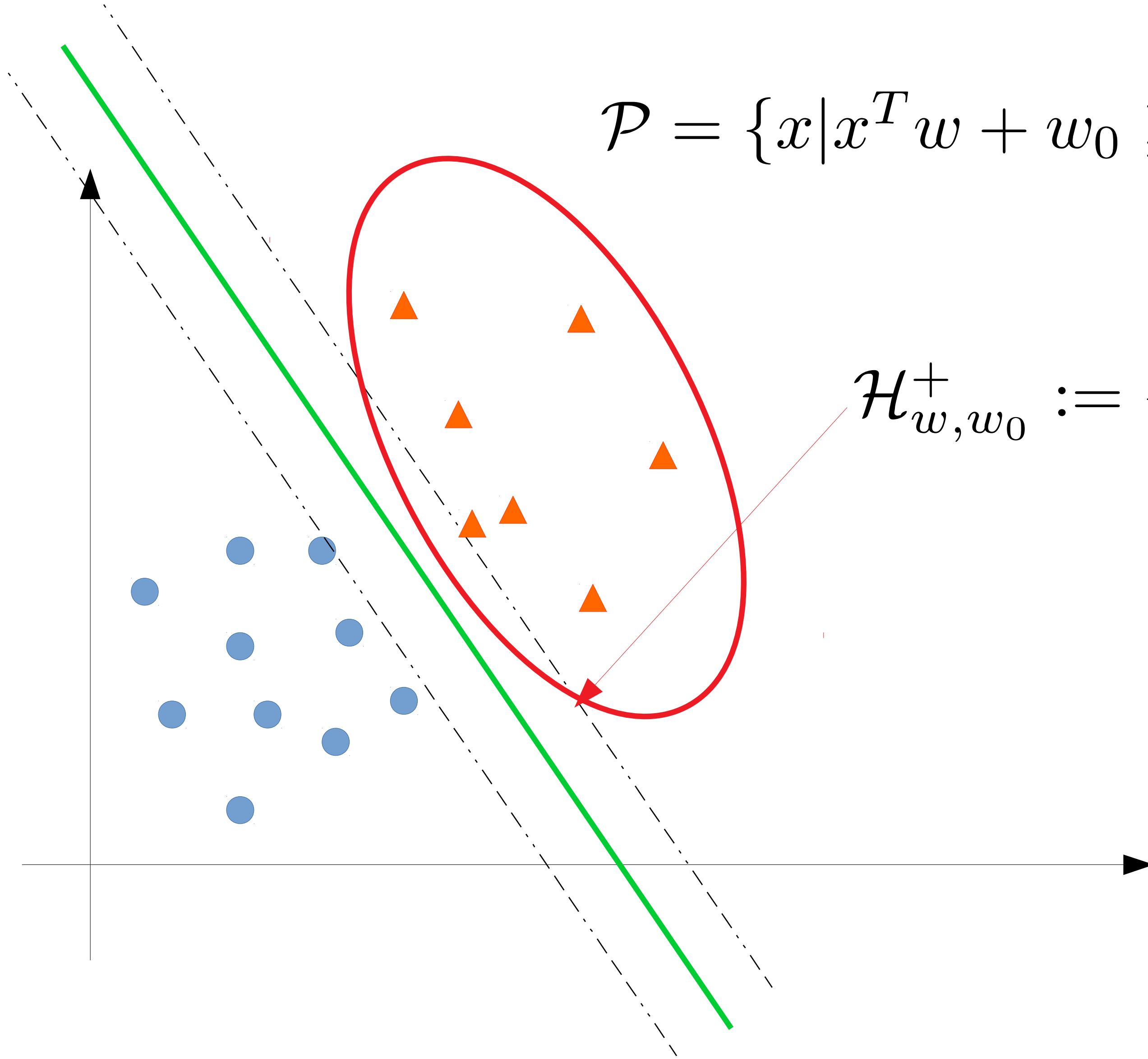


$$\mathcal{H}_{w,w_0} := \{x | x^T w + w_0 = 0\}$$

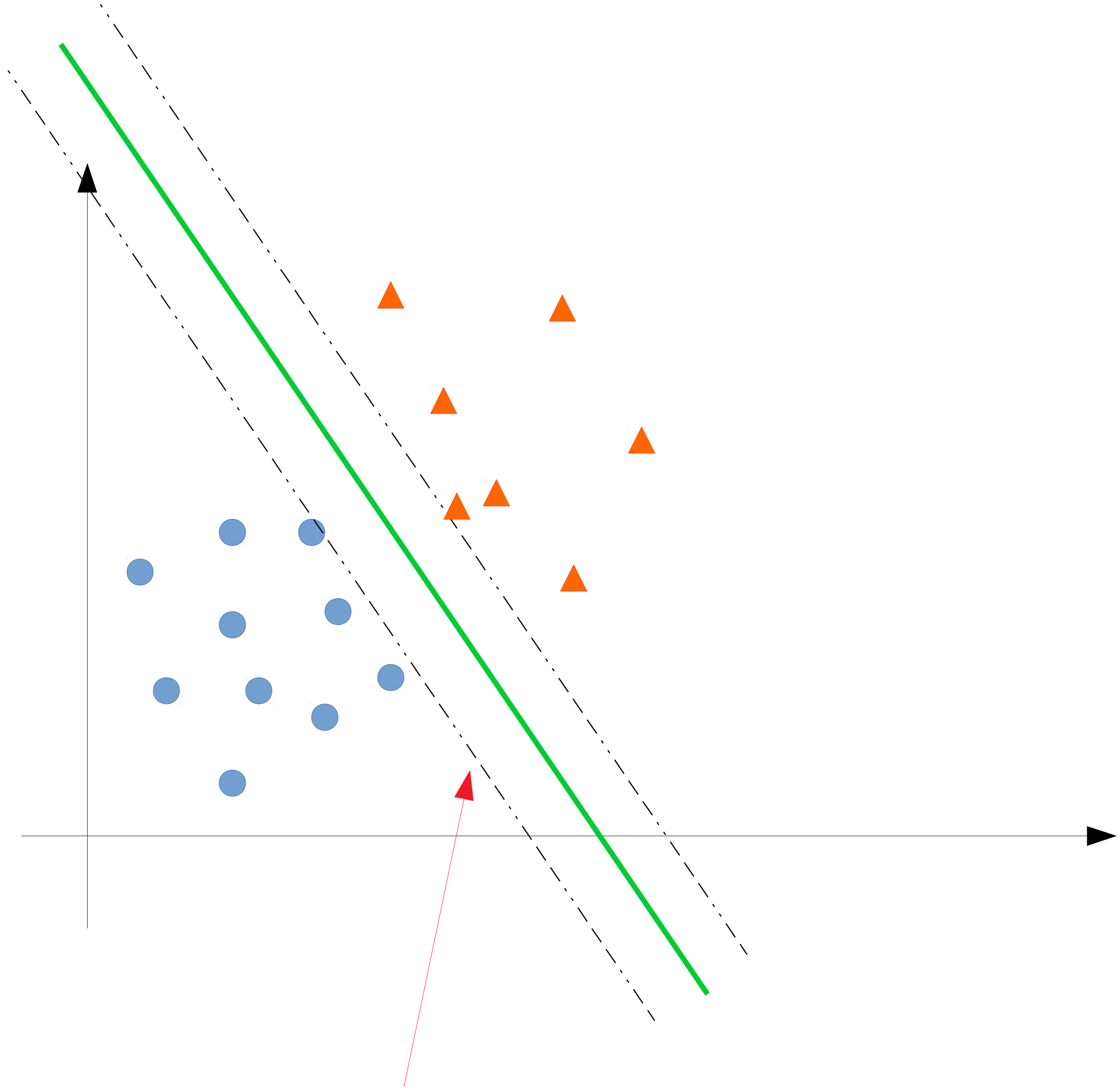


$$\mathcal{H}^{+}_{w,w_0} := \{x | x^T w + w_0 = 1\}$$

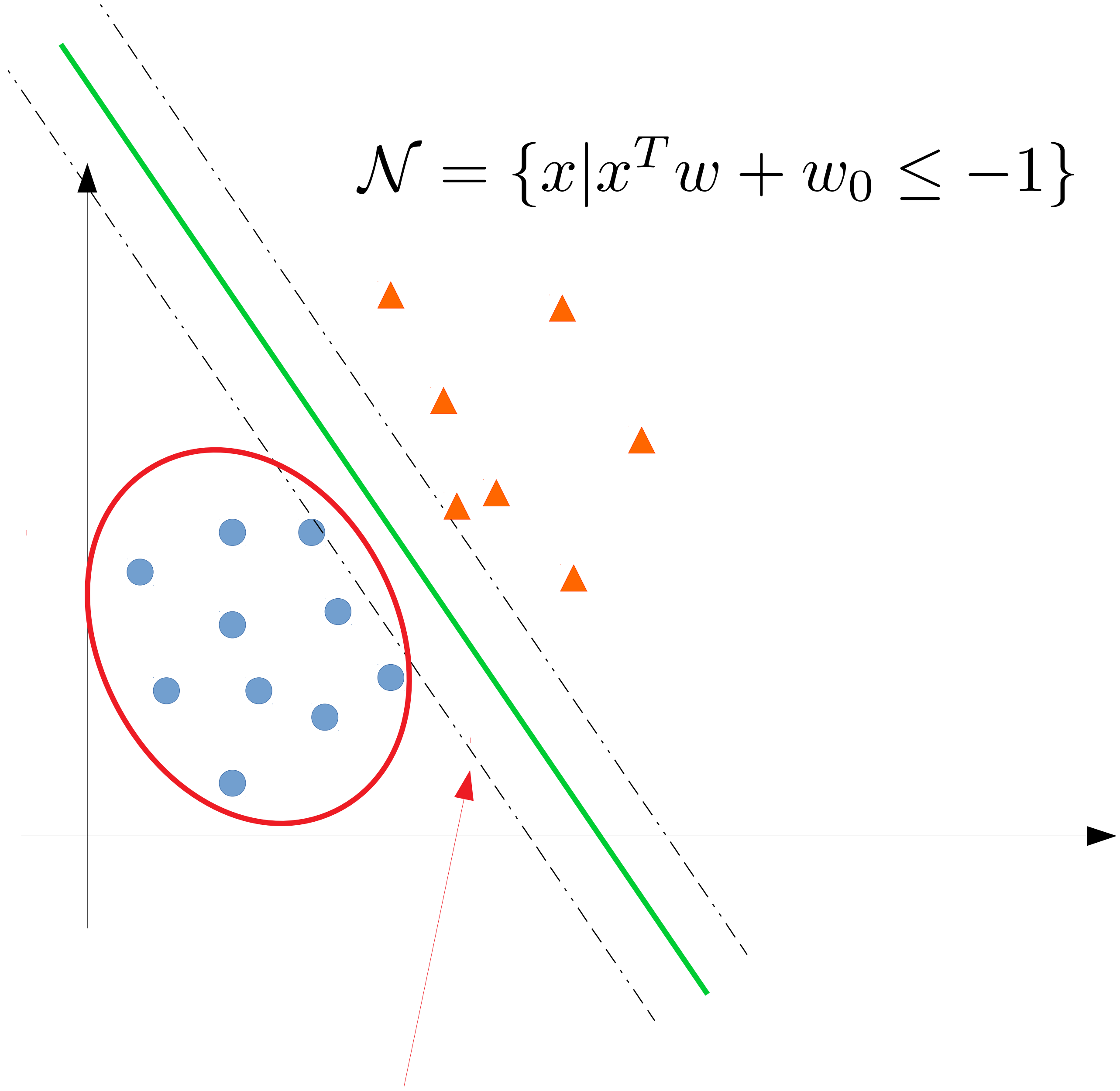
$$\mathcal{P} = \{x | x^T w + w_0 \geq 1\}$$



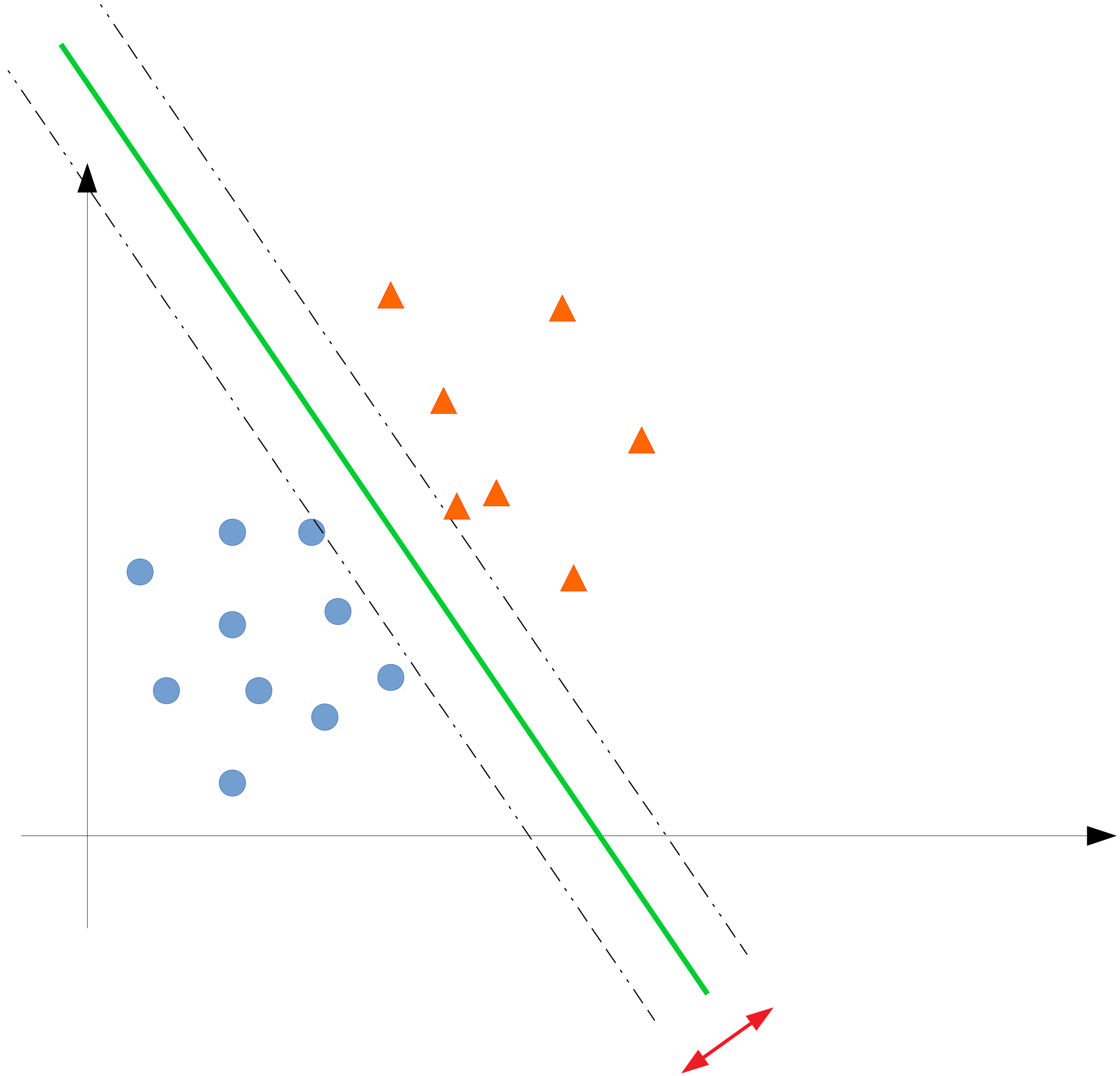
$$\mathcal{H}^{+}_{w,w_0} := \{x | x^T w + w_0 = 1\}$$



$$\mathcal{H}^{-}_{w,w_0} := \{x | x^T w + w_0 = -1\}$$



$$\mathcal{H}^{-}_{w,w_0} := \{x | x^T w + w_0 = -1\}$$



$$d_{w,w_0} = \left| \frac{1-w_0}{||w||} - \frac{-1-w_0}{||w||} \right| = \frac{2}{||w||}$$

Formalization of SVM goal

- Maximize $d_{w,w_0} = \left| \frac{1-w_0}{||w||} - \frac{-1-w_0}{||w||} \right| = \frac{2}{||w||}$
- Subject to:
  - $x^T w + w_0 \geq 1 \qquad \forall x \in \mathcal{P}$
  - $x^T w + w_0 \leq -1 \quad \forall x \in \mathcal{N}$

Formalization of SVM goal

- Minimize $\frac{1}{d_{w,w_0}} = \frac{||w||}{2} = \frac{1}{2} w^T w$
- Subject to:
  - $x^T w + w_0 \geq 1 \qquad \forall x \in \mathcal{P}$
  - $x^T w + w_0 \leq -1 \quad \forall x \in \mathcal{N}$

Formalization of SVM goal

- Minimize $\frac{1}{d_{w,w_0}} = \frac{||w||}{2} = \frac{1}{2} w^T w$
- Subject to:
  - $r_x \cdot (x^T w + w_0) \geq 1 \quad \forall x \in \mathcal{P} \cup \mathcal{N}$

Offen im Denken

Formalization of SVM goal

- Minimize $$\frac{1}{d_{w,w_0}} = \frac{||w||}{2} = \frac{1}{2}w^T w$$
- Subject to:
  - $$r_x \cdot (x^T w + w_0) \geq 1 \quad \forall x \in \mathcal{P} \cup \mathcal{N}$$

$$(r_x = 1 \text{ if } x \in \mathcal{P})$$

$$(r_x = -1 \text{ if } x \in \mathcal{N})$$

Now use Lagrange Multipliers!

- Objective Function: $\frac{1}{2}w^T w$
- Conditions:

$$r_x \cdot (x^T w + w_0) - 1 \geq 0 \qquad \forall x \in \mathcal{P} \cup \mathcal{N}$$

UNIVERSITÄT DUISBURG ESSEN

IS

Offen im Denken

$$f_{Lagr}(w, w_0, B) :=$$

$$\frac{1}{2}w^T w - \sum_{m=1}^{M} \beta^m (r^m (w^T x^m + w_0) - 1)$$

$$f_{Lagr}(w, w_0, B) :=$$

$$\frac{1}{2} w^T w - \sum_{m=1}^{M} \beta^m (r^m (w^T x^m + w_0) - 1)$$

Mimimize wrt w and w_0

Maximize wrt beta

$$
\begin{aligned}
f_{Lagr}(w, w_0, B) \;:=\; & \frac{1}{2} w^T w + \sum_{m=1}^{M} \beta^m - \\
& \sum_{m=1}^{M} \beta^m r^m w_0 - \\
& \sum_{m=1}^{M} \beta^m r^m w^T x^m
\end{aligned}
$$

$$
\begin{aligned}
f_{Lagr}(w, w_0, B) \; := \; & \frac{1}{2} w^T w + \sum_{m=1}^{M} \beta^m - \\
& \sum_{m=1}^{M} \beta^m r^m w_0 - \\
& \sum_{m=1}^{M} \beta^m r^m w^T x^m
\end{aligned}
$$

$$
\nabla_w f_{Lagr}(w, w_0, B) = w - \sum_{m=1}^{M} \beta^m r^m x^m \overset{!}{=} \vec{0}
$$

$$
\begin{aligned}
\boldsymbol{f}_{Lagr}(\boldsymbol{w}, \boldsymbol{w}_0, \boldsymbol{B}) \ := \ & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum_{m=1}^{M} \boldsymbol{\beta}^m - \\
& \sum_{m=1}^{M} \boldsymbol{\beta}^m \boldsymbol{r}^m \boldsymbol{w}_0 - \\
& \sum_{m=1}^{M} \boldsymbol{\beta}^m \boldsymbol{r}^m \boldsymbol{w}^T \boldsymbol{x}^m
\end{aligned}
$$

$$
\nabla_{\boldsymbol{w}} \boldsymbol{f}_{Lagr}(\boldsymbol{w}, \boldsymbol{w}_0, \boldsymbol{B}) = \boldsymbol{w} - \sum_{m=1}^{M} \boldsymbol{\beta}^m \boldsymbol{r}^m \boldsymbol{x}^m \stackrel{!}{=} \vec{\mathbf{0}}
$$

$$
\Rightarrow \qquad \boldsymbol{w} := \sum_{m=1}^{M} \boldsymbol{\beta}^m \boldsymbol{r}^m \boldsymbol{x}^m
$$

$$
\begin{aligned}
f_{Lagr}(w, w_0, B) \;:=\; & \frac{1}{2} w^T w + \sum_{m=1}^{M} \beta^m - \\
& \sum_{m=1}^{M} \beta^m r^m w_0 - \\
& \sum_{m=1}^{M} \beta^m r^m w^T x^m
\end{aligned}
$$

$$
\nabla_w f_{Lagr}(w, w_0, B) = w - \sum_{m=1}^{M} \beta^m r^m x^m \stackrel{!}{=} \vec{0}
$$

$$
\Rightarrow \quad \boxed{w := \sum_{m=1}^{M} \beta^m r^m x^m}
$$

Need Lagrange
Multipliers to get w!

$$
\begin{aligned}
f_{Lagr}(w, w_0, B) \; := \; & \frac{1}{2} w^T w + \sum_{m=1}^{M} \beta^m - \\
& \sum_{m=1}^{M} \beta^m r^m w_0 - \\
& \sum_{m=1}^{M} \beta^m r^m w^T x^m
\end{aligned}
$$

$$
\nabla_w f_{Lagr}(w, w_0, B) = w - \sum_{m=1}^{M} \beta^m r^m x^m \stackrel{!}{=} \vec{0}
$$

$$
\Rightarrow \qquad w := \sum_{m=1}^{M} \beta^m r^m x^m
$$

$$
\frac{\partial f_{Lagr}(w, w_0, B)}{\partial w_0} = \sum_{m=1}^{M} \beta^m r^m \stackrel{!}{=} 0
$$

$$
\begin{aligned}
f_{Lagr}(w, w_0, B) \;:=\; & \frac{1}{2} w^T w + \sum_{m=1}^{M} \beta^m - \\
& \sum_{m=1}^{M} \beta^m r^m w_0 - \\
& \sum_{m=1}^{M} \beta^m r^m w^T x^m
\end{aligned}
$$

$$
\begin{aligned}
f_{Lagr}(w, w_0, B) \ := \ & \frac{1}{2} w^T w + \sum_{m=1}^{M} \beta^m - \\
& \sum_{m=1}^{M} \beta^m r^m w_0 - \\
& \sum_{m=1}^{M} \beta^m r^m w^T x^m
\end{aligned}
$$

$$
\begin{aligned}
f_{Lagr}(w, w_0, B) \; := \; & \frac{1}{2} w^T w + \sum_{m=1}^{M} \beta^m - \\
& \sum_{m=1}^{M} \beta^m r^m w_0 - \\
& \sum_{m=1}^{M} \beta^m r^m w^T x^m
\end{aligned}
$$

$\boxed{w^T w}$

$$
\begin{aligned}
f_{Lagr}(w, w_0, B) \;:=\; & \frac{1}{2} w^T w + \sum_{m=1}^{M} \beta^m - \\
& \sum_{m=1}^{M} \beta^m r^m w_0 - \\
& \sum_{m=1}^{M} \beta^m r^m w^T x^m
\end{aligned}
$$

$$
\begin{aligned}
= \; & \sum_{m=1}^{M} \beta^m - \\
& \frac{1}{2} \underbrace{(\sum_{m=1}^{M} \beta^m r^m x^m)}_{w}{}^T \underbrace{(\sum_{l=1}^{M} \beta^l r^l x^l)}_{w}
\end{aligned}
$$

$$
\begin{aligned}
f_{Lagr}(w, w_0, B) \;:=\; & \frac{1}{2}w^T w + \sum_{m=1}^{M} \beta^m - \\
& \sum_{m=1}^{M} \beta^m r^m w_0 - \\
& \sum_{m=1}^{M} \beta^m r^m w^T x^m
\end{aligned}
$$

$$
\begin{aligned}
= \; & \sum_{m=1}^{M} \beta^m - \\
& \frac{1}{2}(\underbrace{\sum_{m=1}^{M} \beta^m r^m x^m}_{w})^T(\underbrace{\sum_{l=1}^{M} \beta^l r^l x^l}_{w})
\end{aligned}
$$

Maximize this **Dual form** to get Lagrange Multipliers!
(Condition: Betas >= 0)

$$
\begin{aligned}
f_{Lagr}(w, w_0, B) \; &:= \; \frac{1}{2} w^T w + \sum_{m=1}^{M} \beta^m - \\
&\sum_{m=1}^{M} \beta^m r^m w_0 - \\
&\sum_{m=1}^{M} \beta^m r^m w^T x^m
\end{aligned}
$$

$$
= \sum_{m=1}^{M} \beta^m - \frac{1}{2} \underbrace{\left(\sum_{m=1}^{M} \beta^m r^m x^m \right)^T}_{w} \underbrace{\left(\sum_{l=1}^{M} \beta^l r^l x^l \right)}_{w}
$$

Maximize this **Dual form** to get Lagrange Multipliers!
(Condition: Betas >= 0)

From lecture:
x^m is support vector iff beta^m > 0!!!

UNIVERSITÄT DUISBURG ESSEN IS

Offen im Denken

Berechnung von $\boldsymbol{w}_0^*$:

Sei $\boldsymbol{x}^m$ ein Support-Vektor, z.B. derjenigen Klasse mit $\boldsymbol{r}^m = \mathbf{1}$.

Dann gilt $\quad \boldsymbol{w}^{*,T}\boldsymbol{x}^m + \boldsymbol{w}_0 = \mathbf{1} \Rightarrow \boldsymbol{w}_0^* := \mathbf{1} - \boldsymbol{w}^{*,T}\boldsymbol{x}^m$

Summary

- How to get optimal weight vector?
  - Formulate Lagrange Approach
  - Calculate partial derivatives of Lagrange function wrt to weights
  - Optimal weight vector is dependant on Lagrange multipliers
  - Insert finding into formular to get Dual Form
  - (dual form is independent of weight vector)
  - Max Dual Form to get Lagrange Multipliers!
  - (Lagrange Multipliers determine optimal weight vector)
  - Positive Lagrange Multipliers indicate support vector
  - Use support vector to calculate offset w_0

Revision of Lecture

- What to do if classes are not linearly separable?

Revision of Lecture

- What to do if classes are not linearly separable?
  - Apply Transformation (embedding function) f_{Tran} on input space to establish linear separability

Revision of Lecture

- What to do if classes are not linearly separable?
  - Apply Transformation (embedding function) f_{Tran} on input space to establish linear separability
  - The discriminant function would then change to:

$$f_{Disk}(x) := \text{sign}(w^{*,T} f_{Tran}(x) + w_0^*)$$

$$w^* := \sum_{m=1}^{M} \beta^{*,m} r^m f_{Tran}(x^m)$$
$$\beta^{*,m} \neq 0$$

Revision of Lecture

- What to do if classes are not linearly separable?
  - Apply Transformation (embedding function) f_{Tran} on input space to establish linear separability
  - The discriminant function would then change to:

$$f_{Disk}(x) := \text{sign}(w^{*,T} f_{Tran}(x) + w_0^*)$$

Transformed input

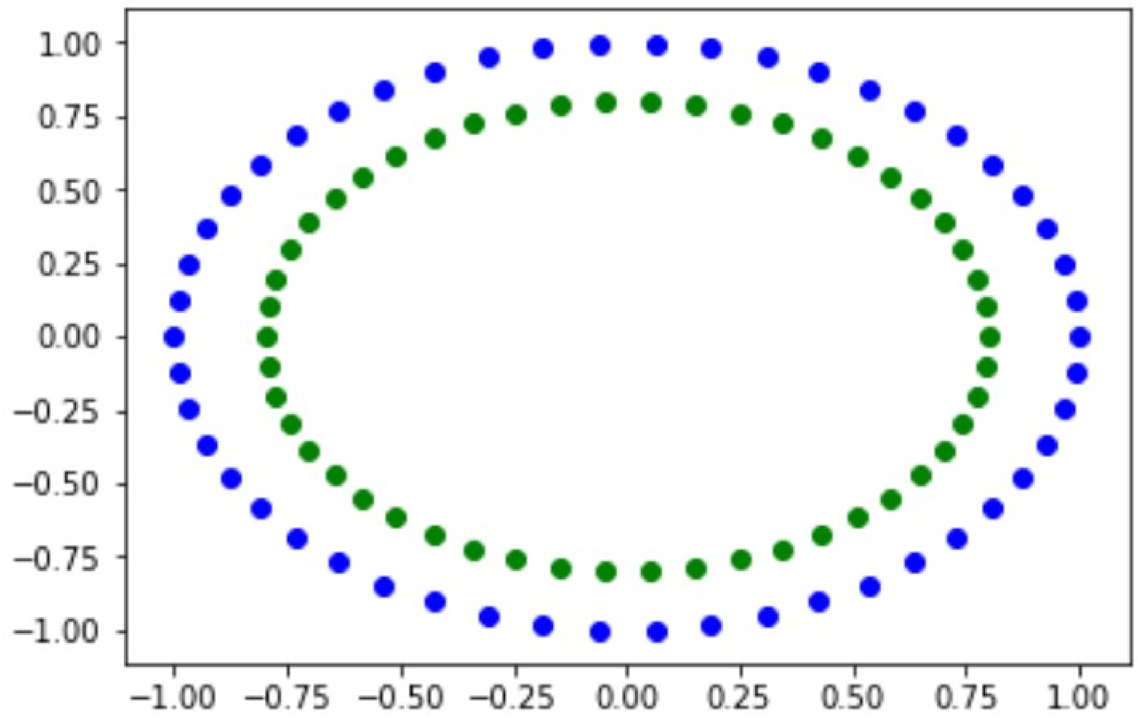
$$w^* := \sum_{m=1}^{M} \beta^{*,m} r^m f_{Tran}(x^m)$$

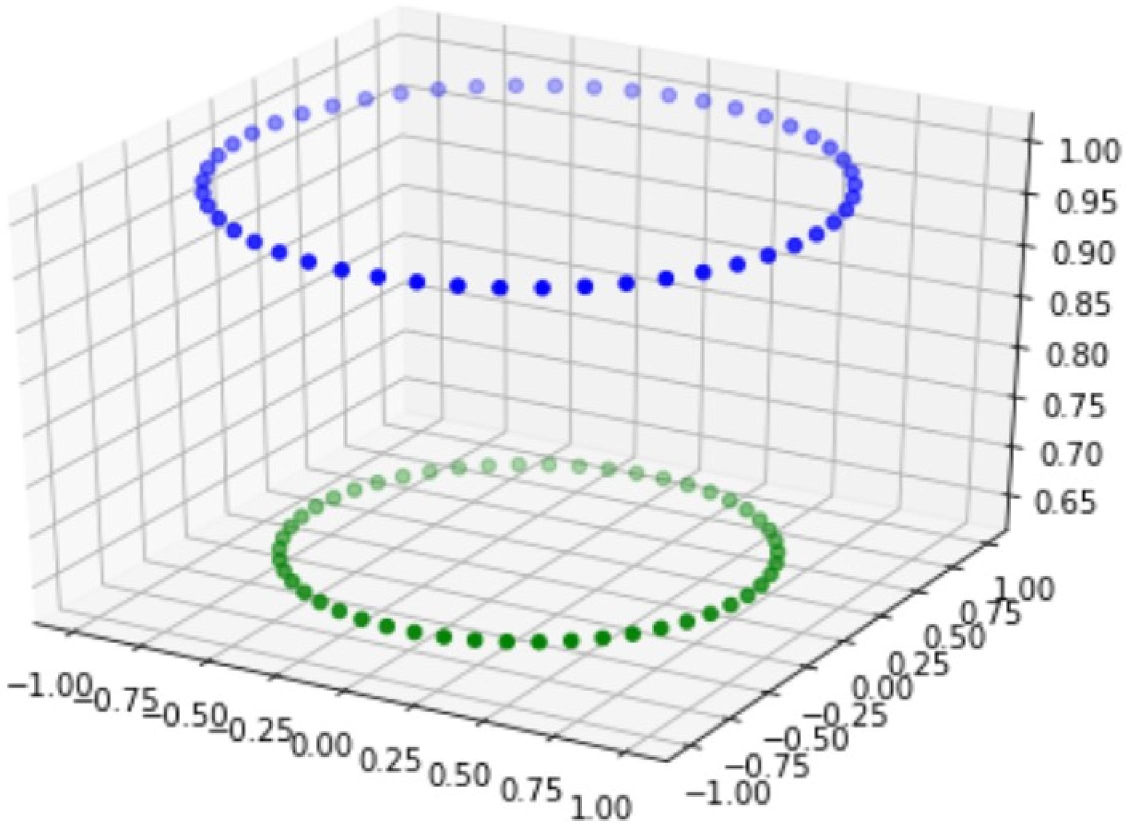
$$\beta^{*,m} \neq 0$$

Optimal weight vector for hyperplane
in transformed input space

(Lagrange coefficient)

(Desired ouput)


1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00 −0.75 −0.50 −0.25 0.00 0.25 0.50 0.75 1.00



$$f_{Tran}(x_1, x_2) := \begin{pmatrix} x_1 \\ x_2 \\ x_1^2 + x_2^2 \end{pmatrix}$$

Revision of Lecture

- What is the Kernel Trick?

Revision of Lecture

- What is the Kernel Trick?
  - Embeddings into very high dimensions demand more computational power and capacity for calculation of scalar product

$$\boldsymbol{w}^{*,T}\boldsymbol{f}_{Tran}(\boldsymbol{x}) = \sum \beta^{*,m} r^{m} \boldsymbol{f}_{Tran}(\boldsymbol{x}^{m})^{T} \boldsymbol{f}_{Tran}(\boldsymbol{x})$$

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Revision of Lecture

- What is the Kernel Trick?
  - Embeddings into very high dimensions demand more computational power and capacity for calculation of scalar product

$$w^{*,T} f_{Tran}(x) = \sum \beta^{*,m} r^m \boxed{f_{Tran}(x^m)^T f_{Tran}(x)}$$

Revision of Lecture

- What is the Kernel Trick?
  - Embeddings into very high dimensions demand more computational power and capacity for calculation of scalar product

$$\boldsymbol{w}^{*,T}\boldsymbol{f}_{Tran}(\boldsymbol{x}) = \sum \beta^{*,m} r^m \boxed{\boldsymbol{f}_{Tran}(\boldsymbol{x}^m)^T \boldsymbol{f}_{Tran}(\boldsymbol{x})}$$

  - There exist combination of embedding function f_{Tran} and kernel function f_{Kern} , such that Mercer condition holds:

$$\boldsymbol{f}_{Tran}(\boldsymbol{x}^m)^T \cdot \boldsymbol{f}_{Tran}(\boldsymbol{x}^\ell) = \boldsymbol{f}_{Kern}(\boldsymbol{x}^m, \boldsymbol{x}^\ell) \quad \forall \boldsymbol{x}^m, \boldsymbol{x}^\ell$$

Revision of Lecture

- What is the Kernel Trick?
  - Embeddings into very high dimensions demand more computational power and capacity for calculation of scalar product

$$\boldsymbol{w}^{*,T} \boldsymbol{f}_{Tran}(\boldsymbol{x}) = \sum \beta^{*,m} r^m \boxed{\boldsymbol{f}_{Tran}(\boldsymbol{x}^m)^T \boldsymbol{f}_{Tran}(\boldsymbol{x})}$$

  - There exist combination of embedding function f_{Tran} and kernel function f_{Kern} , such that Mercer condition holds:

$$\boldsymbol{f}_{Tran}(\boldsymbol{x}^m)^T \cdot \boldsymbol{f}_{Tran}(\boldsymbol{x}^\ell) = \boldsymbol{f}_{Kern}(\boldsymbol{x}^m, \boldsymbol{x}^\ell) \quad \forall \boldsymbol{x}^m, \boldsymbol{x}^\ell$$

  - This way, we do not even need to compute the transformation! Only the <u>result</u> of the scalar product!

UNIVERSITÄT DUISBURG ESSEN IS

Offen im Denken

Revision of Lecture

- Some Kernel functions:

$$f_{Kern}(x^m, x^\ell) := (x^{m,T} \cdot x^\ell)^2$$

$$f_{Kern}(x^m, x^\ell) := e^{-\frac{\|x^m - x^\ell\|^2}{2\sigma^2}}$$

$$f_{Kern}(x^m, x^\ell) := \tanh(x^{m,T} \cdot x^\ell)$$

$$f_{Kern}(x^m, x^\ell) := (x^{m,T} \cdot x^\ell + 1)^d$$

Topics covered in this course

- Neuroinformatics & Machine Learning Basics
- Statistical Decision Theory
  & Statistical Classifiers
- McCulloch Pitts Cell & Perceptron
  & Learning Algorithms
- Adaline
- Multi Layer Perceptron
- Convolutional Neural Networks
- Bias-Variance-Dilemma
  & Statistical Analyses (e.g. Precision/Recall)
- RBF- Nets
- SVMs

Exam hints

- No proofs!
- No programming tasks!
- Content from the tutorials is expected to be known (except for proofs and programming assignments!)
- It is more important to be able to accurately explain concepts and relationships than to learn complicated mathematical deriviations by heart.
- You should understand the derivations, though!
- Simple (usually short) formulas, introduced in the lecture, are expected to be known (e.g. Bayes etc...)

Exam hints

- These lecture topics are definitely not going to be in the exam:
  - Chapter 1 : slides 3-4
  - Chapter 3: slides 42-47
  - Chapter 5: slides 19-24, slides 29-32, slide 38
  - Chapter 6: slides 3-4, slide 10, slides 62-64
  - Chapter 7: slides 4-9
  - Chapter 8:
    - Do not learn formulas by heart, but rather understand them and know the meaning of the used symbols!

Types of Questions

- Definitions:
  - E.g.: When are two sets of classes said to be linearly separable?
  - Assume two sets: $\mathbf{P},\mathbf{N} \subset \mathbb{R}^N$
  - P, N linearly separable if there exists a weight vector $w \in \mathbb{R}^N$ and a threshold $\Theta \in \mathbb{R}$ such that

$$w^T x \geq \Theta \quad \forall x \in P$$

$$w^T x < \Theta \quad \forall x \in N$$

Types of Questions

- Calculations (only minor part of exam):
  - Some easy calculations
  - No calculator required
  - e.g. covariance matrix, bayes, conditional probabilities or something similarly easy
  - No calculations of inverse matrix expected

Offen im Denken

Types of Questions

- Algorithms:
  - Description of algorithm
    - How does it work?
    - Do not leave out important aspects!
    - Make sure the corrector sees, that you know what you are talking about
    - (Do not just write down some formulas without any explanation!)
    - E.g. Perceptron learning algorithm

Types of Questions

- Algorithms:
  - Idea of algorithm
    - What is the general idea of this algorithm?
    - Explain it in an understandable fashion
      (like you would explain it to a collegue)
    - Draw sketches to illustrate your explanations
    -

Types of Questions

- Derivations:
  - Do not learn derivations/proofs by heart!
  - Make sure you understand WHY we had to make these derivations!
  - To which conclusions did these poofs/derivations lead us?

Types of Questions

- Formulas:
  - Do not learn very complex formulas by heart!
    (see list in previous slide)
  - Make sure you understand which components these formulas consist of!
  - Easier formulas are expected to be known!
  - (E.g. Bias, Variance, Definition of a hyperplane, linear associator, bayes, etc...)

Types of Questions

- Transfer of knowledge:
  - Get an overall understanding of which topics this course covers
  - What are the differences?
  - What are common properties?
  - Is there any relation between chapter X and Y?

Good luck in your exam!