Machine Learning Review

Support Vector Machines

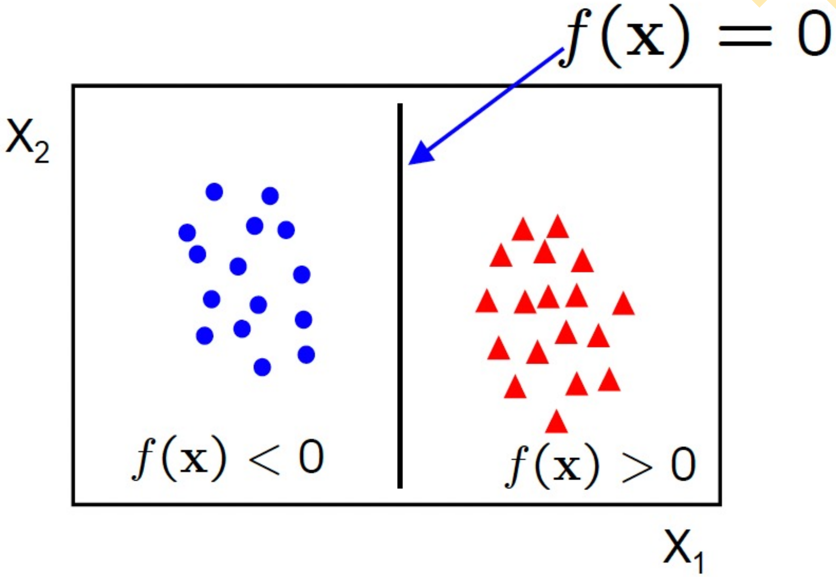
SVM

- We approach the classification problem in a direct way:
  - We try and find a plane that separates the classes in the feature space.

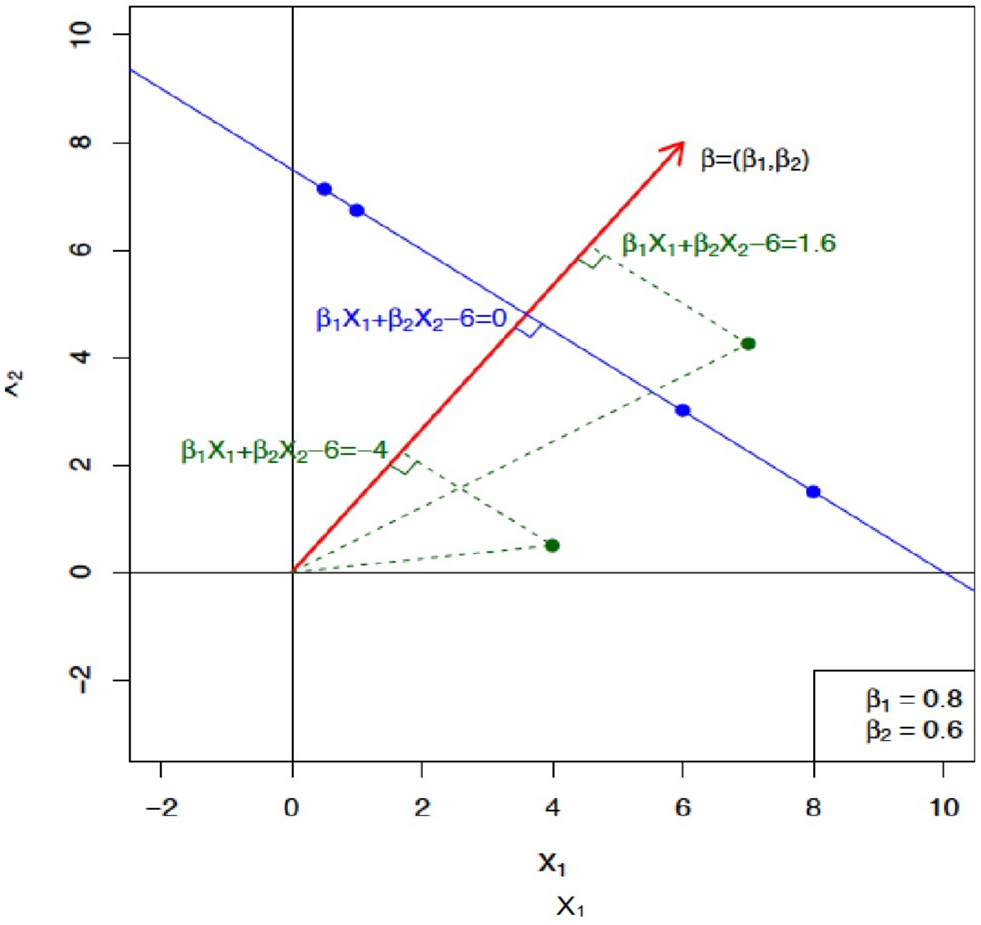
- If we cannot, we get creative:
  - We soften what we mean by “separates”

What is a hyperplane?

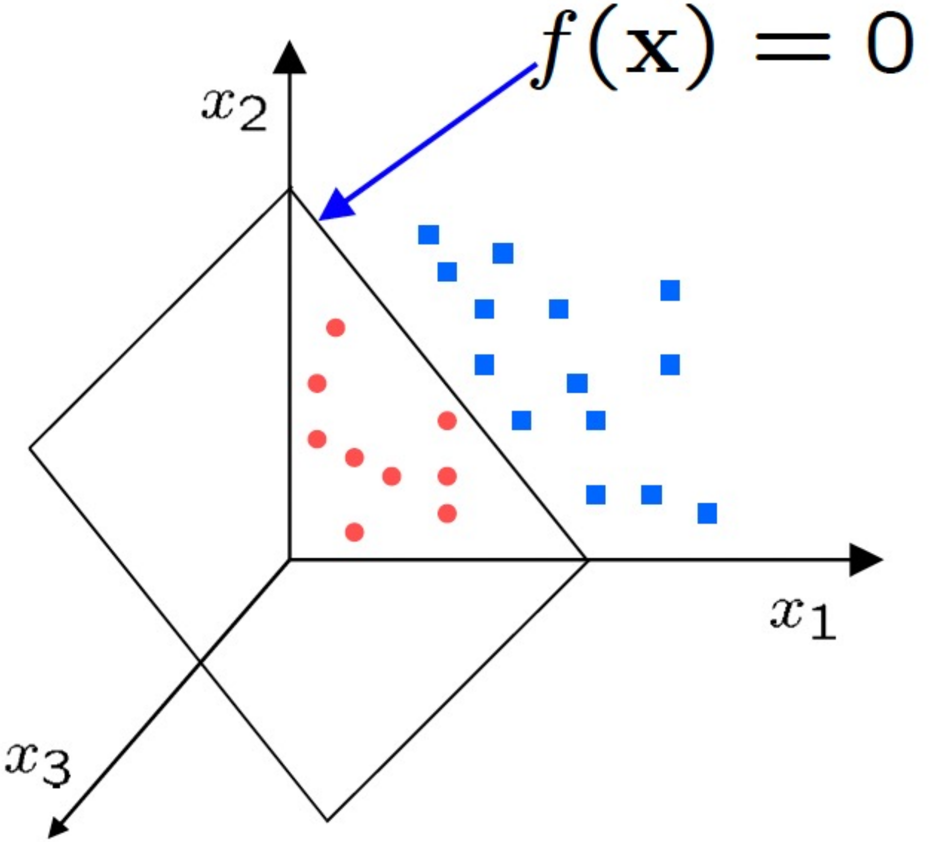
- $f\ (x) = W^T X + b$
  - A hyperplane in p dimensions is a subspace of dimension p − 1.
  - In 2-D, a hyperplane is a line.
  - General form: $\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_p X_p = 0$
  - The weight vector W is called the normal vector — it points in a direction orthogonal to the surface of a hyperplane.



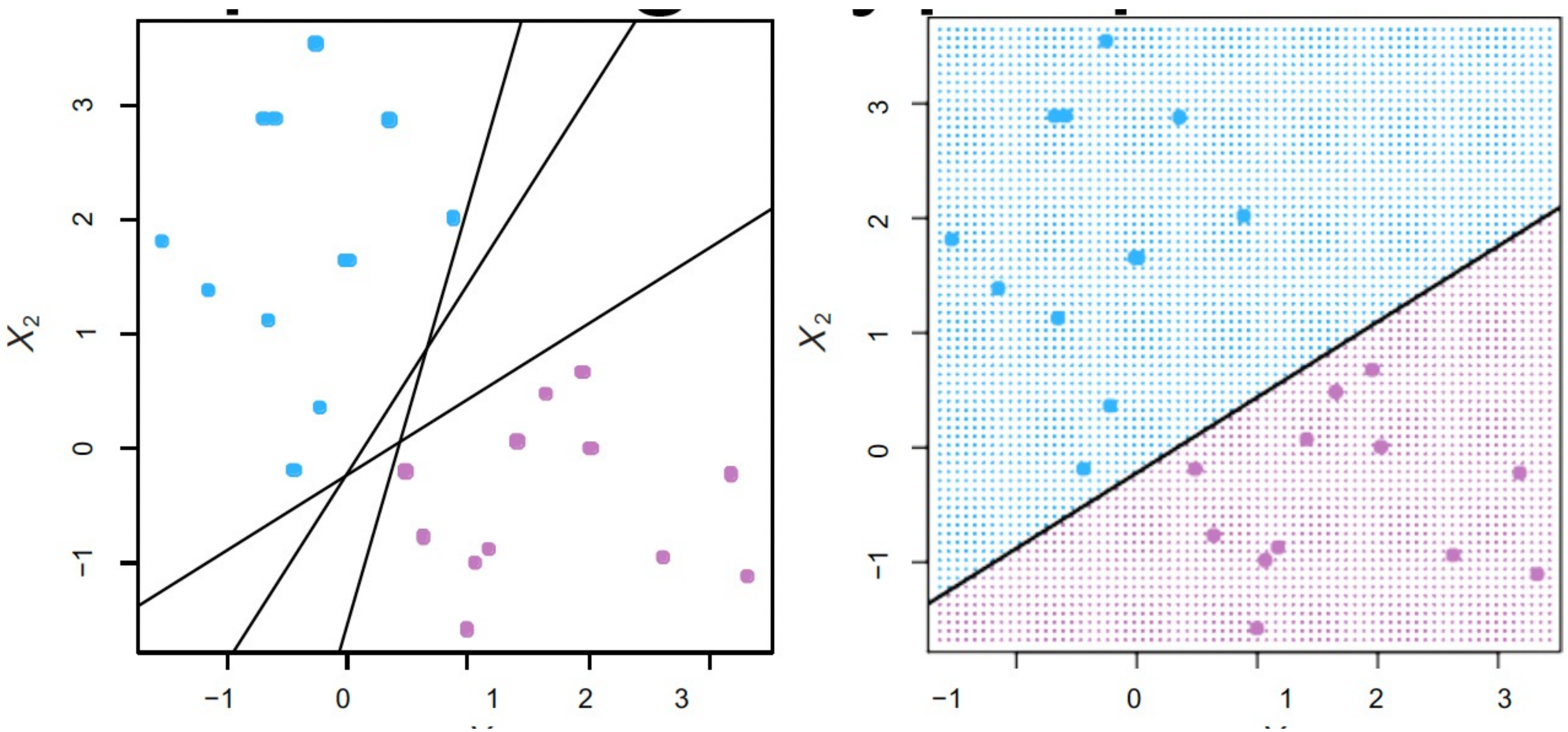
2-D


$\beta=(\beta_1,\beta_2)$
$\beta_1X_1+\beta_2X_2-6=1.6$
$\beta_1X_1+\beta_2X_2-6=0$
$\beta_1X_1+\beta_2X_2-6=-4$
$\beta_1 = 0.8$
$\beta_2 = 0.6$
10
8
6
4
2
0
−2
−2
0
2
4
6
8
10
X_2
X_1
X_1

3-D


$f(\mathbf{x}) = 0$
x_2
x_1
x_3

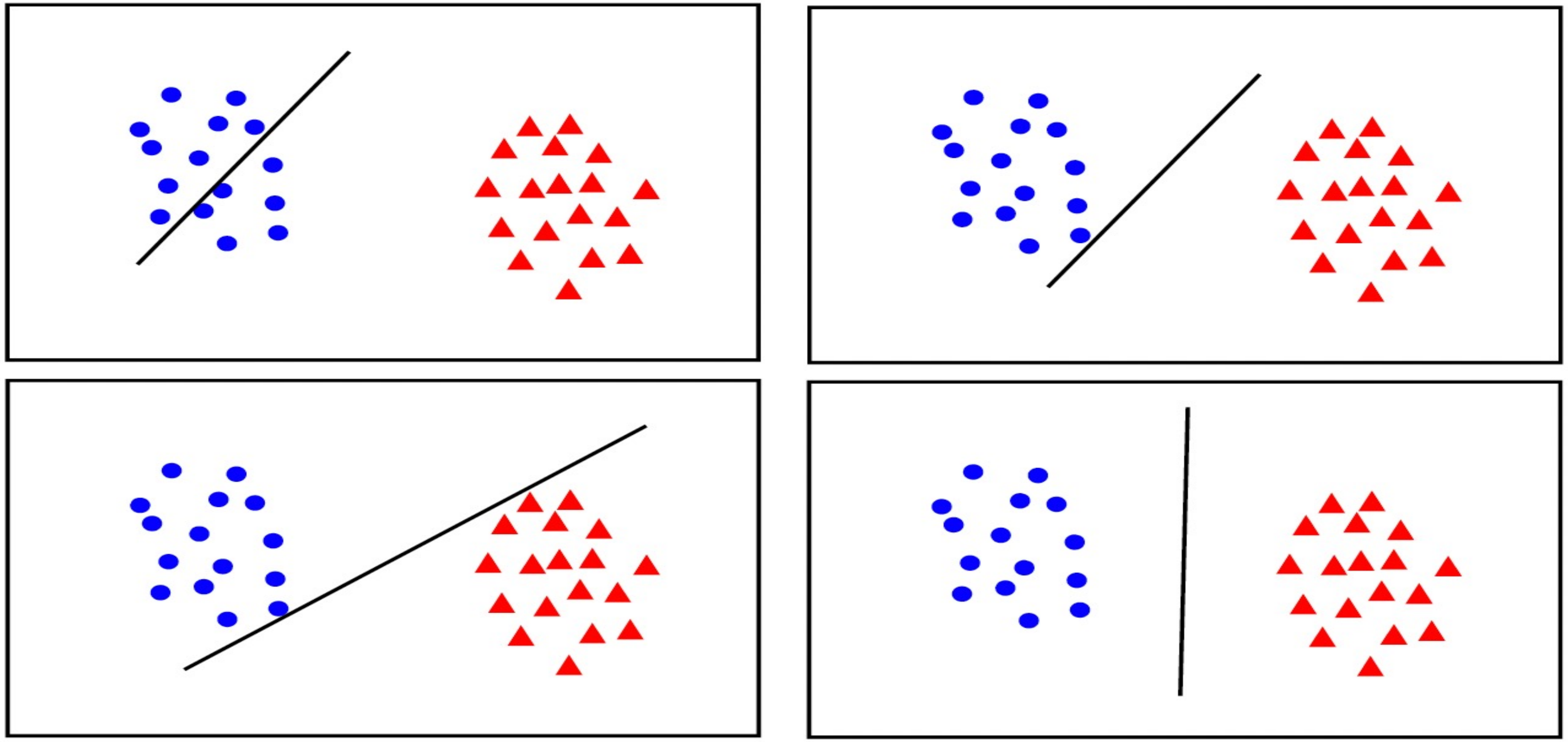
Separating hyperplanes



$f\ (x) > 0$ for data points on one side of the hyperplane;

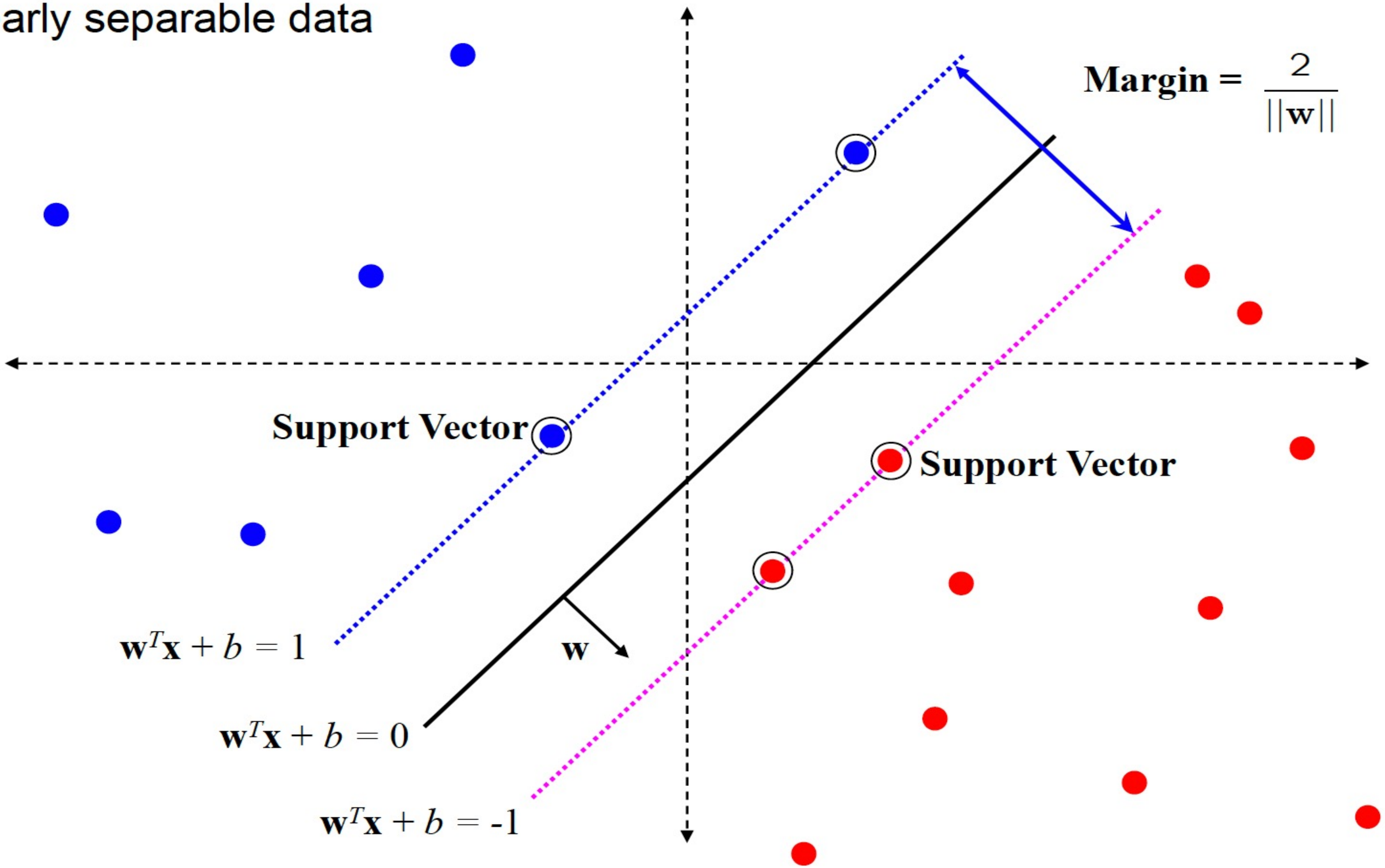
$f\ (x) < 0$ for data points on the other side of the hyperplane.

Which W is the best?



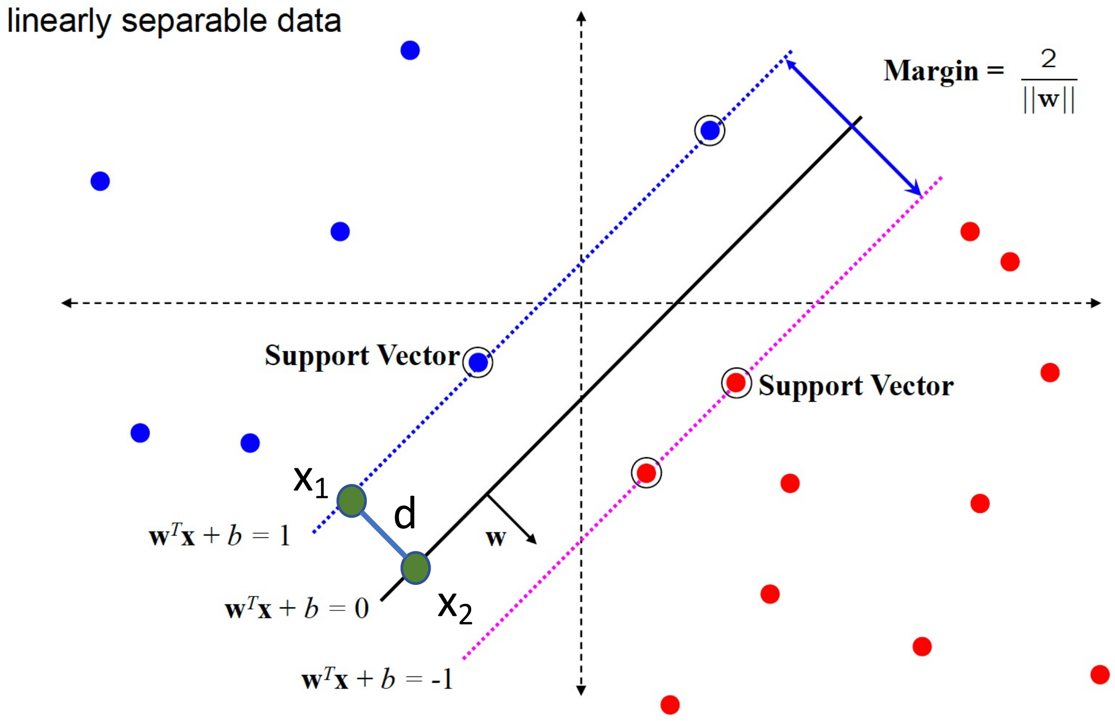
- Maximum margin solution: among all separating hyperplanes, find the one that makes the biggest gap or margin between the two classes.

linearly separable data


Margin = $\frac{2}{||\mathbf{w}||}$
Support Vector
Support Vector
$\mathbf{w}^T\mathbf{x} + b = 1$
$\mathbf{w}$
$\mathbf{w}^T\mathbf{x} + b = 0$
$\mathbf{w}^T\mathbf{x} + b = -1$

Margin representation

- $w^T x_1 + \mathrm{b} = 1$
- $w^T x_2 + \mathrm{b} = 0$
- $w^T (x_1 - x_2) = 1$
- $d = x_1 - x_2$
- $w^T \mathrm{d} = 1$
- because $\cos(\theta) = 1$ (w and d are parallel)
- and $\cos(\theta) = \dfrac{w^T \mathrm{d}}{\|w\| \|d\|}$
- so $\dfrac{w^T \mathrm{d}}{\|w\| \|d\|} = 1$
- $\dfrac{1}{\|w\| \|d\|} = 1$
- Therefore $\| d \| = \dfrac{1}{\|w\|}$



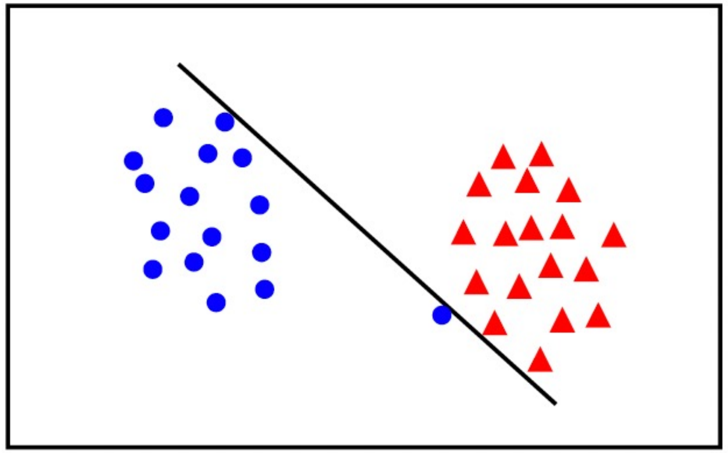
linearly separable data
Margin = $\frac{2}{||\mathbf{w}||}$
Support Vector
Support Vector
x_1
d
x_2
w
$\mathbf{w}^T\mathbf{x} + b = 1$
$\mathbf{w}^T\mathbf{x} + b = 0$
$\mathbf{w}^T\mathbf{x} + b = -1$

SVM optimization

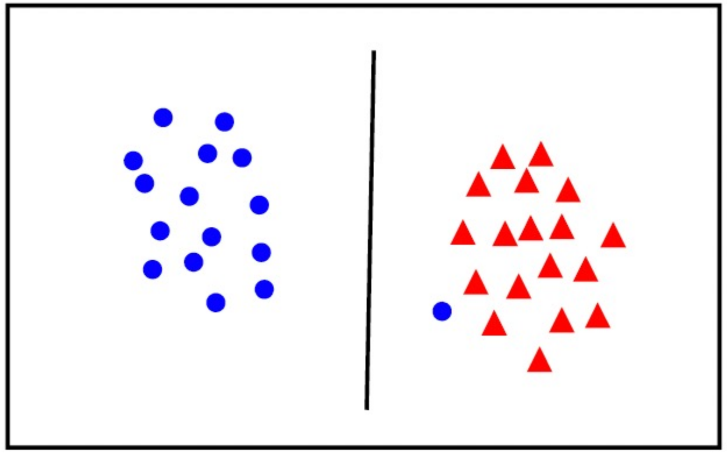
- max $\frac{2}{\|w\|}$ subject to $w^T x_i + b \ \geq +1 \ if \ y_i$ = +1 for i = 1…N
- max $\frac{2}{\|w\|}$ subject to $w^T x_i + b \ \leq -1 \ if \ y_i$ = -1 for i = 1…N

- Or min $\| \ w \ \|^2$ subject to $y_i \ (w^T x_i + b) \ \geq 1$ for i = 1…N

Soft margin



- the points can be linearly separated but there is a very narrow margin

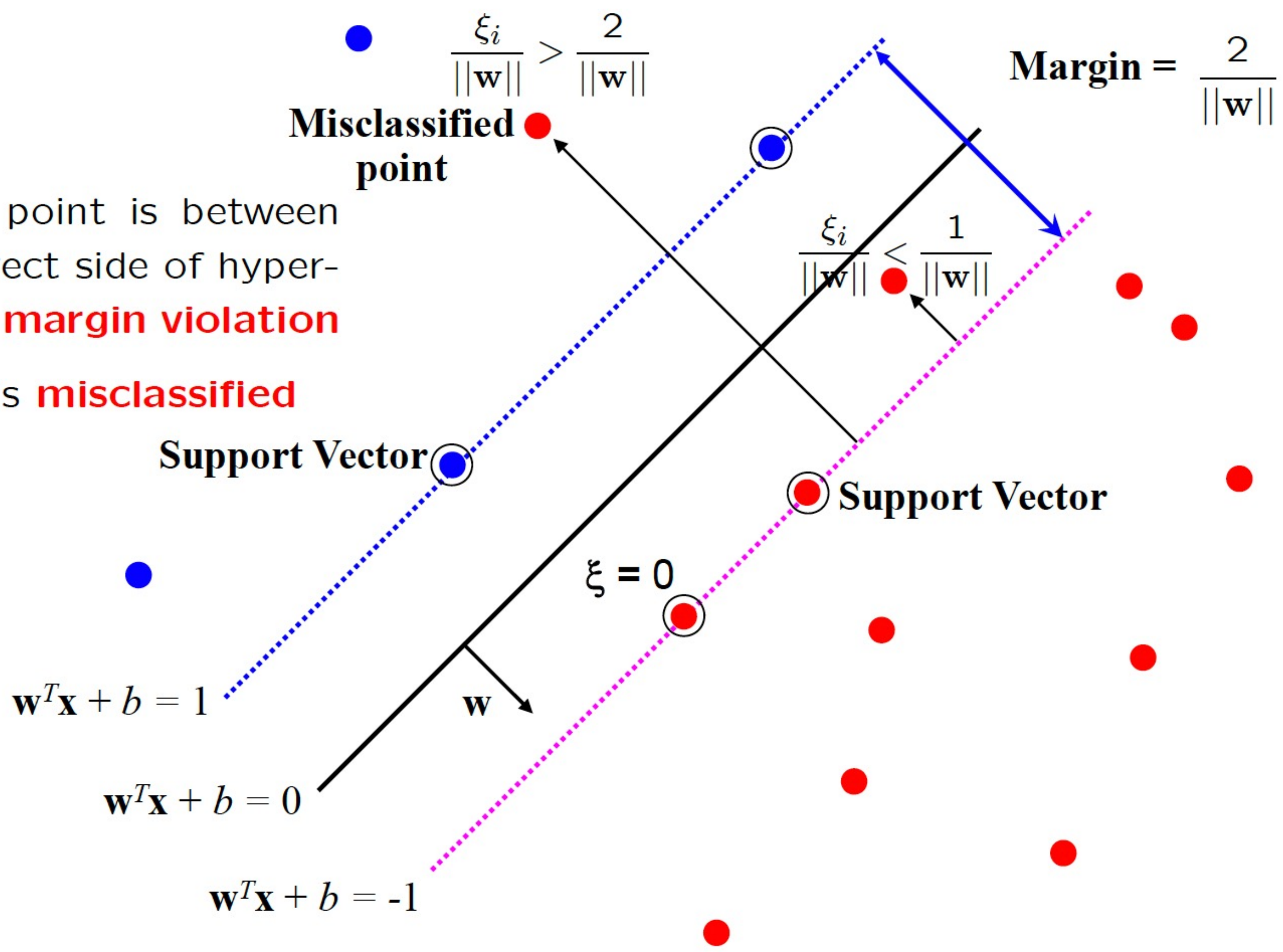


- but possibly the large margin solution is better, even though one constraint is violated

Slack variables

$$\xi_i \geq 0$$

- for $0 < \xi \leq 1$ point is between margin and correct side of hyper-plane. This is a **margin violation**
- for $\xi > 1$ point is **misclassified**



$\frac{\xi_i}{||\mathbf{w}||} > \frac{2}{||\mathbf{w}||}$
Misclassified point
Margin = $\frac{2}{||\mathbf{w}||}$
$\frac{\xi_i}{||\mathbf{w}||} < \frac{1}{||\mathbf{w}||}$
Support Vector
Support Vector
ξ = 0
$\mathbf{w}^T\mathbf{x} + b = 1$
$\mathbf{w}$
$\mathbf{w}^T\mathbf{x} + b = 0$
$\mathbf{w}^T\mathbf{x} + b = -1$

Support vectors

- Only observations that either lie on the margin or that violate the margin will affect the hyperplane, and hence the classifier obtained.
- In other words, an observation that lies strictly on the correct side of the margin does not affect the support vector classifier.
- Observations that lie directly on the margin, or on the wrong side of the margin for their class, are known as support vectors. These observations do affect the support vector classifier.

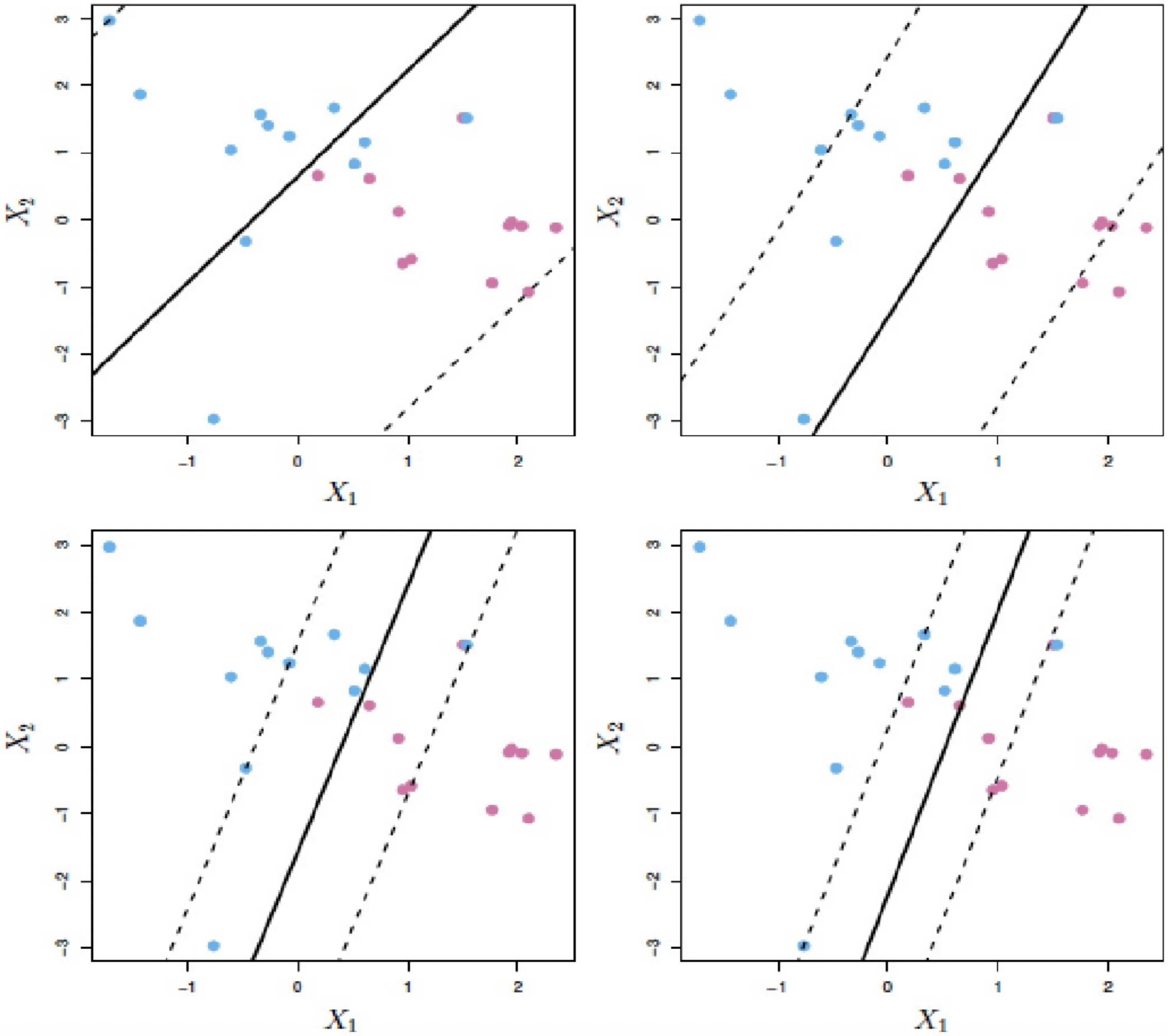
SVM optimization becomes…

$$\min \parallel w \parallel^2 + C \sum_i^N \xi_i$$

subject to

$$y_i \ (w^T x_i + b) \geq 1 - \xi_i \text{ for i = 1…N}$$

- C is a penalty parameter that determines the severity of the violations to the margin (and to the hyperplane) that we will tolerate.
- small C allows constraints to be easily ignored → large margin
- large C makes constraints hard to ignore → narrow margin
- C = ∞ enforces all constraints: hard margin



3
2
1
0
−1
−2
−3
X_2
−1
0
1
2
X_1
3
2
1
0
−1
−2
−3
X_2
−1
0
1
2
X_1
3
2
1
0
−1
−2
−3
X_2
−1
0
1
2
X_1
3
2
1
0
−1
−2
−3
X_2
−1
0
1
2
X_1

C is a regularization parameter

- C controls the bias-variance trade-off of the statistical learning technique.
- When C is large, we seek narrow margins that are rarely violated; this amounts to a classifier that is highly fit to the data, which may have low bias but high variance.
- On the other hand, when C is small, the margin is wider and we allow more violations to it; this amounts to fitting the data less hard and obtaining a classifier that is potentially more biased but may have lower variance.

Optimization

Learning an SVM has been formulated as a constrained optimization problem over w and ξ

$\min \parallel w \parallel^2 + C\, \sum_i^N \xi_i$ subject to $y_i\ (w^T x_i + b) \geq 1 - \xi_i$ for i = 1…N

The constraint $y_i\ (w^T x_i + b) \geq 1 - \xi_i$ can be written concisely as

$$y_i\ f(x_i) \geq 1 - \xi_i$$

because with $\xi_i \geq 0$

$$\xi_i = \max(0, 1 - y_i\ f(x_i))$$

Then optimization is

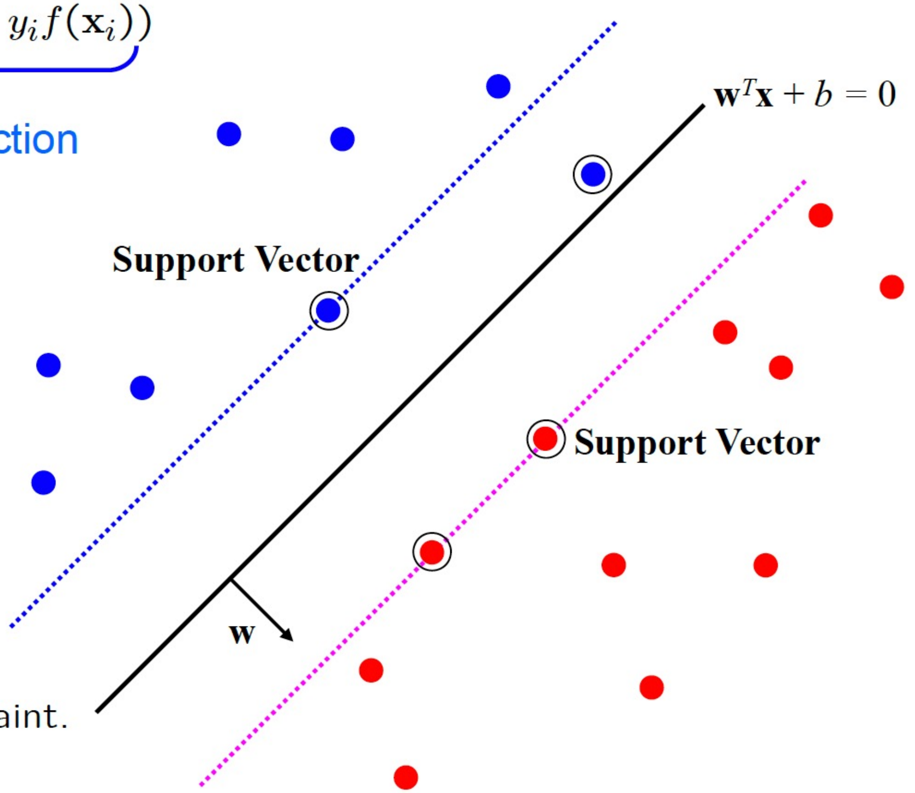
$$\min \underbrace{\parallel w \parallel^2}_{\text{regularization}} + C \sum_i^N \underbrace{\max(0, 1 - \ y_i\ f(x_i))}_{\text{loss function}}$$

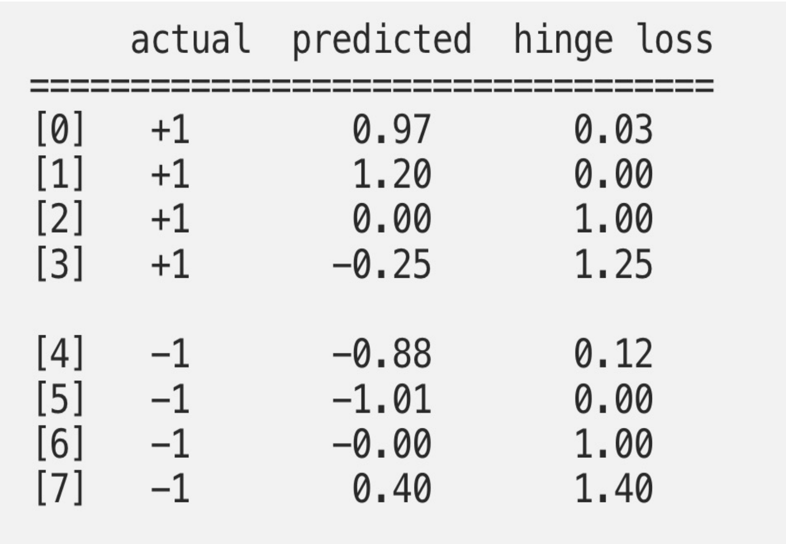
Hinge loss

$$\min_{\mathbf{w}\in\mathbb{R}^d} ||\mathbf{w}||^2 + C\sum_{i}^{N} \underbrace{\max\left(0, 1 - y_i f(\mathbf{x}_i)\right)}_{\text{loss function}}$$

Points are in three categories:

1. $y_i f(x_i) > 1$
   Point is outside margin.
   No contribution to loss
2. $y_i f(x_i) = 1$
   Point is on margin.
   No contribution to loss.
   As in hard margin case.
3. $y_i f(x_i) < 1$
   Point violates margin constraint.
   Contributes to loss





```
       actual  predicted  hinge loss
====================================
[0]    +1         0.97        0.03
[1]    +1         1.20        0.00
[2]    +1         0.00        1.00
[3]    +1        -0.25        1.25

[4]    -1        -0.88        0.12
[5]    -1        -1.01        0.00
[6]    -1        -0.00        1.00
[7]    -1         0.40        1.40
```