Machine Learning I

MICRO-455

Classification with SVM

The Teaching Team

Fall 2025

EPFL LASA

Q1

- Computational Aspects of SVM.
  - Number of datapoints: M
  - Data dimension: N
  - Number of support vectors: S

Q1 > A

- What is the number of parameters to store?

Q1 > A

- What is the number of parameters to store?

$$f(\boldsymbol{x}) = \text{sign}\left(\left(\sum_i \alpha_i y_i \langle \boldsymbol{x}, \boldsymbol{x}^i \rangle\right) + b\right)$$

Q1 > A

- What is the number of parameters to store?

$$f(\boldsymbol{x}) = \text{sign}\left(\left(\sum_i \alpha_i y_i \langle \boldsymbol{x}, \boldsymbol{x}^i \rangle\right) + b\right)$$

$$S \times N + S + 1 = S(N+1) + 1$$

Q1 > A

- What is the number of parameters to store?

$$f(\boldsymbol{x}) = \text{sign}\left(\left(\sum_i \alpha_i y_i \langle \boldsymbol{x}, \boldsymbol{x}^i \rangle\right) + b\right)$$

$$S \times N + S + 1 = S(N+1) + 1$$

Q1 > A

- What is the number of parameters to store?

$$f(\boldsymbol{x}) = \text{sign}\left(\left(\sum_i \alpha_i y_i \langle \boldsymbol{x}, \boldsymbol{x}^i \rangle\right) + b\right)$$

$$S \times N + S + 1 = S(N+1) + 1$$

Q1 > A

- What is the number of parameters to store?

$$f(\boldsymbol{x}) = \text{sign}\left(\left(\sum_i \alpha_i y_i \langle \boldsymbol{x}, \boldsymbol{x}^i \rangle\right) + b\right)$$

$$S \times N + S + 1 = S(N+1) + 1$$

Q1 > A

- What is the number of parameters to store?

$$f(\boldsymbol{x}) = \text{sign}\left(\left(\sum_i \alpha_i y_i \langle \boldsymbol{x}, \boldsymbol{x}^i \rangle\right) + b\right)$$

$$S \times N + S + 1 = S(N+1) + 1$$

- It is possible to store one less parameter; but it is not practical.

$$\sum_i^M \alpha_i y_i = \sum_i^S \alpha_i y_i = 0$$

Q1 > B

- What is the required memory to store the the trained model?
- S=10,000; N=100; each float takes 8B.

Q1 > B

- What is the required memory to store the the trained model?
- S=10,000; N=100; each float takes 8B.

$$S(N+1) = 10,000 \times (100+1) = 1,010,000$$

Q1 > B

- What is the required memory to store the the trained model?
- S=10,000; N=100; each float takes 8B.

$$S(N+1) = 10,000 \times (100+1) = 1,010,000$$

$$\frac{1,010,000 \times 8\text{B}}{1024 \times 1024} \approx 7.71\text{MB}$$

Q1 > C

- Training on the dataset with M=1,000 and N=10 takes 0.1s.
- What is training time for M=1,000,000 and N=100?
- The training time complexity is assumed to be: $\mathcal{O}(MN^2)$.

Q1 > C

- Training on the dataset with M=1,000 and N=10 takes 0.1s.
- What is training time for M=1,000,000 and N=100?
- The training time complexity is assumed to be: $\mathcal{O}(MN^2)$.

$$100,000 \times 0.1\text{s} = 10,000\text{s} = 166\frac{2}{3}\text{min} \approx 2.78\text{h} \approx 2\text{h}47\text{min}$$

Q1 > D

- How much energy (in Wh) is required for the training phase?
- CPU consumption: 50W; Kettle: 1500W for 5min

Q1 > D

- How much energy (in Wh) is required for the training phase?
- CPU consumption: 50W; Kettle: 1500W for 5min

$$\boxed{2.78\text{h} \times 50\text{W} = 139\text{Wh}}$$

Q1 > D

- How much energy (in Wh) is required for the training phase?
- CPU consumption: 50W; Kettle: 1500W for 5min

$$2.78\text{h} \times 50\text{W} = 139\text{Wh}$$

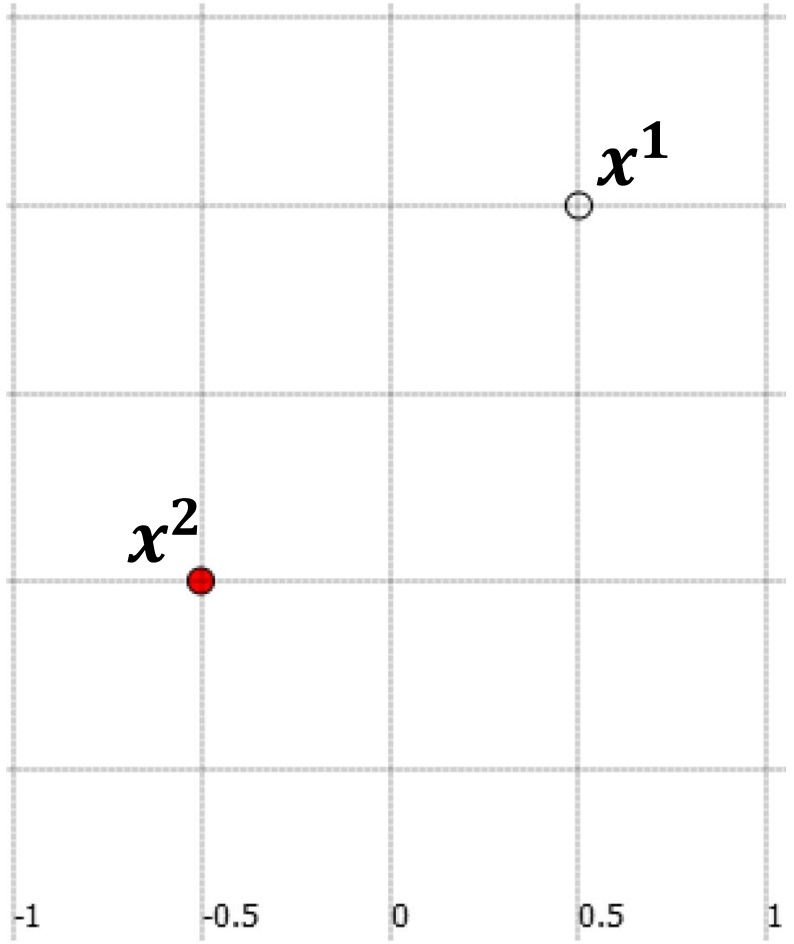
$$\frac{5}{60}\text{h} \times 1500\text{W} = 125\text{Wh}$$

Q2 > A

- Compute the coefficients and the bias term of the SVM classifier for two datapoints with RBF Kernel.
- 1: $\boldsymbol{x^1} = [\mathbf{0.5}, \mathbf{0.5}]^\top; y_1 = 1$
- 2: $\boldsymbol{x^2} = [\mathbf{-0.5}, \mathbf{-0.5}]^\top; y_2 = -1$

$$k(\boldsymbol{x}, \boldsymbol{x}_i) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_i\|^2}{2\sigma^2}\right)$$

- $k(\boldsymbol{x^1}, \boldsymbol{x^2}) = \mathbf{0.5} = k(\boldsymbol{x^2}, \boldsymbol{x^1})$
- $k(\boldsymbol{x^1}, \boldsymbol{x^1}) = \mathbf{1.0} = k(\boldsymbol{x^2}, \boldsymbol{x^2})$



Q2 > A

- Both datapoints must be support vectors because $\sum_i \alpha_i y_i = 0$.

Q2 > A

- Both datapoints must be support vectors because $\sum_i \alpha_i y_i = 0$.

$$\sum_i \alpha_i y_i = 0 \rightarrow \alpha_1 = \alpha_2$$

Q2 > A

- Both datapoints must be support vectors because $\sum_i \alpha_i y_i = 0$.

$$\sum_i \alpha_i y_i = 0 \rightarrow \alpha_1 = \alpha_2$$

$$f(\boldsymbol{x}) = \text{sign}\left(\left(\sum_i \alpha_i y_i k(\boldsymbol{x}, \boldsymbol{x}^i)\right) + b\right) \rightarrow \begin{cases} \sum_{i=1}^{M} \alpha_i y_i k(\boldsymbol{x}^1, \boldsymbol{x}^i) + b = 1 \\ \sum_{i=1}^{M} \alpha_i y_i k(\boldsymbol{x}^2, \boldsymbol{x}^i) + b = -1 \end{cases}$$

Q2 > A

- Both datapoints must be support vectors because $\sum_i \alpha_i y_i = 0$.

$$\sum_i \alpha_i y_i = 0 \rightarrow \alpha_1 = \alpha_2$$

$$f(\boldsymbol{x}) = \text{sign}\left(\left(\sum_i \alpha_i y_i k(\boldsymbol{x}, \boldsymbol{x}^i)\right) + b\right) \rightarrow \begin{cases} \sum_{i=1}^{M} \alpha_i y_i k(\boldsymbol{x}^1, \boldsymbol{x}^i) + b = 1 \\ \sum_{i=1}^{M} \alpha_i y_i k(\boldsymbol{x}^2, \boldsymbol{x}^i) + b = -1 \end{cases}$$

$$\begin{cases} \alpha_1 - 0.5\alpha_1 + b = 1 \\ 0.5\alpha_1 - \alpha_1 + b = -1 \end{cases} \rightarrow \begin{cases} 0.5\alpha_1 + b = 1 \\ -0.5\alpha_1 + b = -1 \end{cases}$$

Q2 > A

- Both datapoints must be support vectors because $\sum_i \alpha_i y_i = 0$.

$$\sum_i \alpha_i y_i = 0 \rightarrow \alpha_1 = \alpha_2$$

$$f(\boldsymbol{x}) = \text{sign}\left(\left(\sum_i \alpha_i y_i k(\boldsymbol{x}, \boldsymbol{x}^i)\right) + b\right) \rightarrow \begin{cases} \sum_{i=1}^{M} \alpha_i y_i k(\boldsymbol{x}^1, \boldsymbol{x}^i) + b = 1 \\ \sum_{i=1}^{M} \alpha_i y_i k(\boldsymbol{x}^2, \boldsymbol{x}^i) + b = -1 \end{cases}$$

$$\begin{cases} \alpha_1 - 0.5\alpha_1 + b = 1 \\ 0.5\alpha_1 - \alpha_1 + b = -1 \end{cases} \rightarrow \begin{cases} 0.5\alpha_1 + b = 1 \\ -0.5\alpha_1 + b = -1 \end{cases}$$

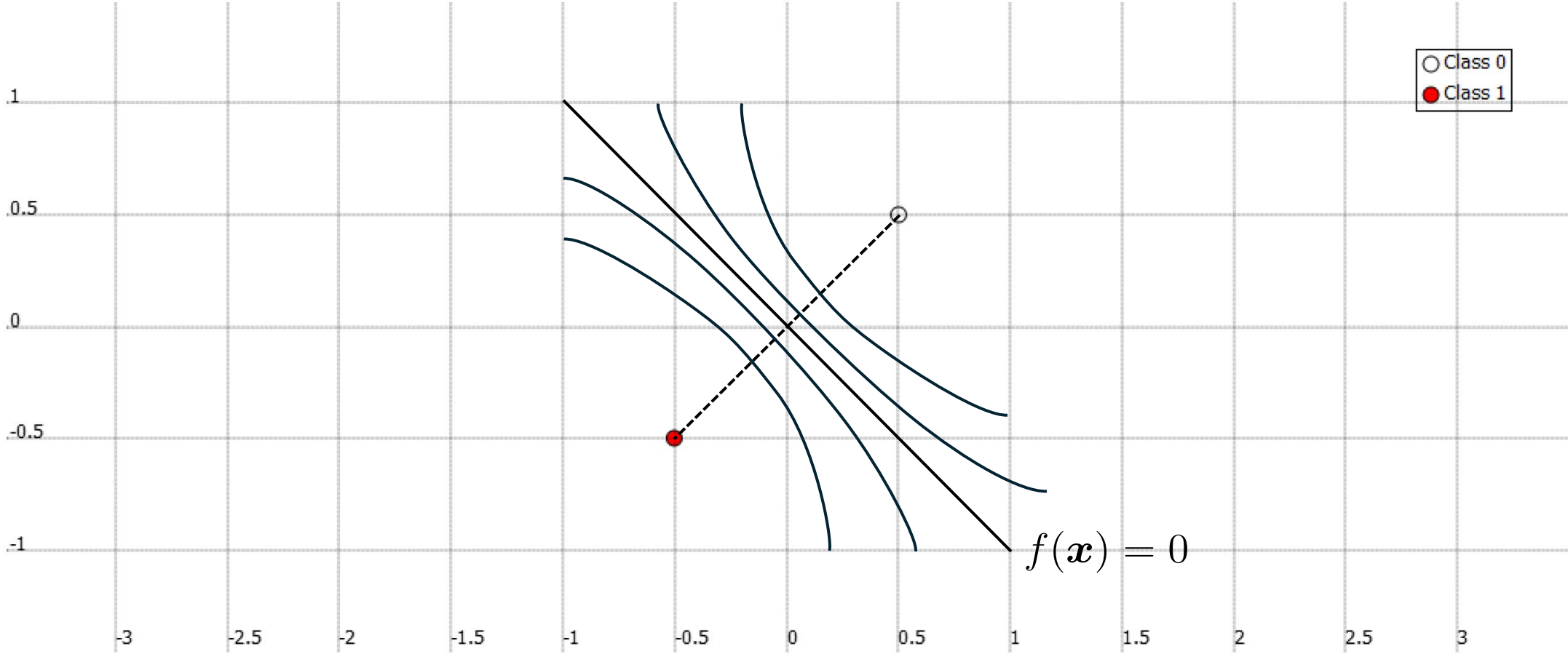
$$\alpha_1 = \alpha_2 = 2; \quad b = 0$$

Q2 > A

- Draw the isolines and the separating hyperplane.

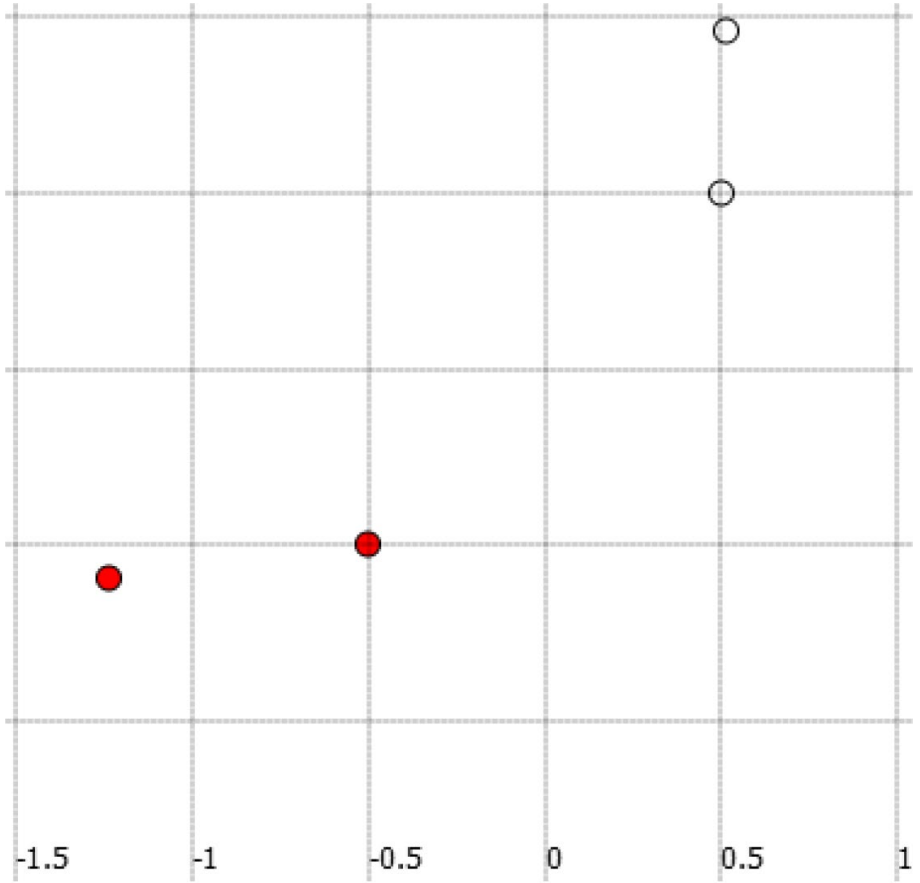
Q2 > A

- Draw the isolines and the separating hyperplane.



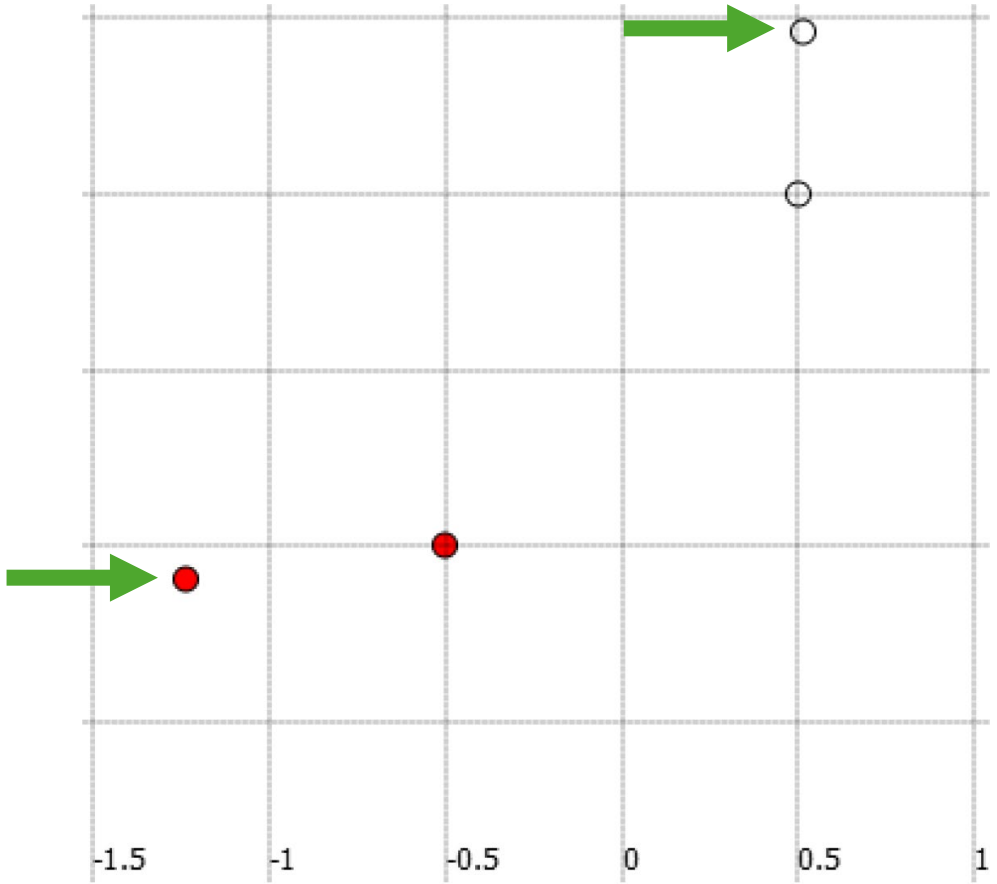
Q2 > B – Case 1

- Draw the new separating hyperplane.



Q2 > B – Case 1

- Draw the new separating hyperplane.

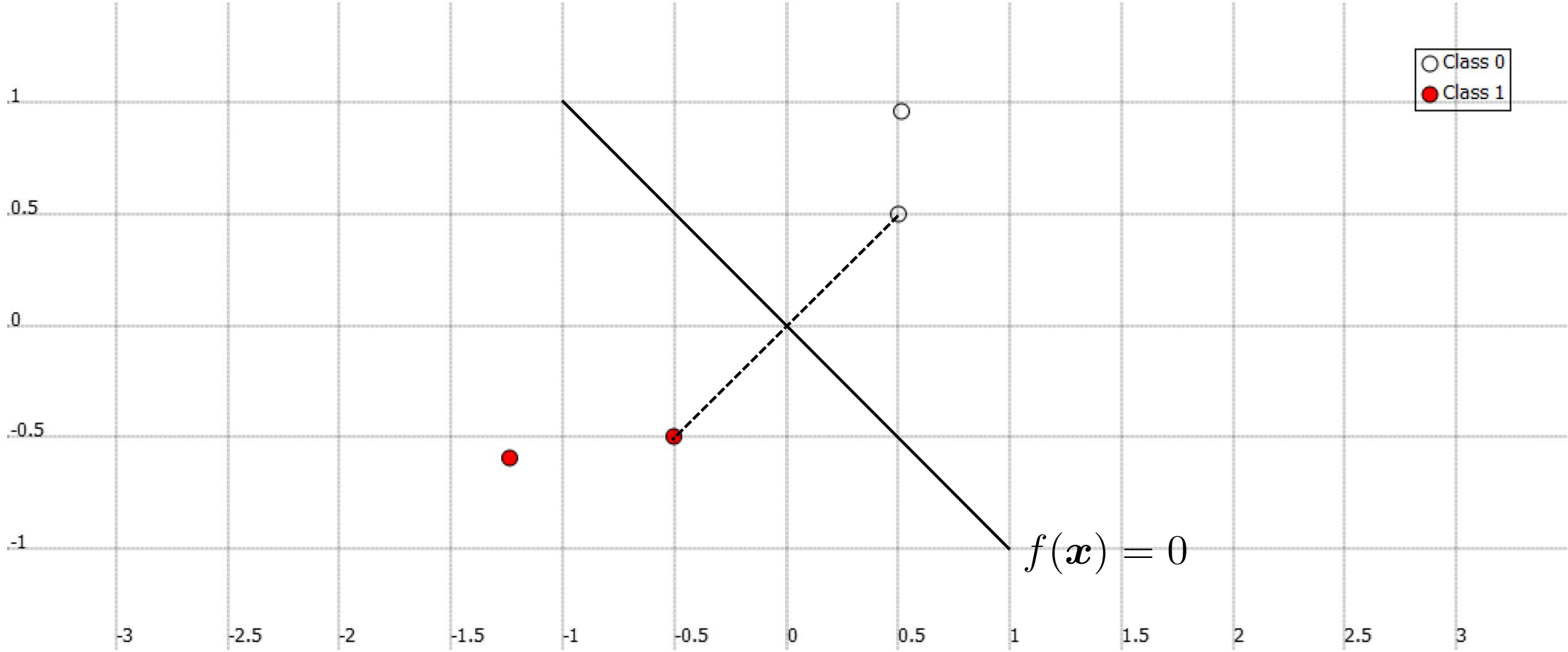


Q2 > B – Case 1

- The separating hyperplane and the support vectors do not change.

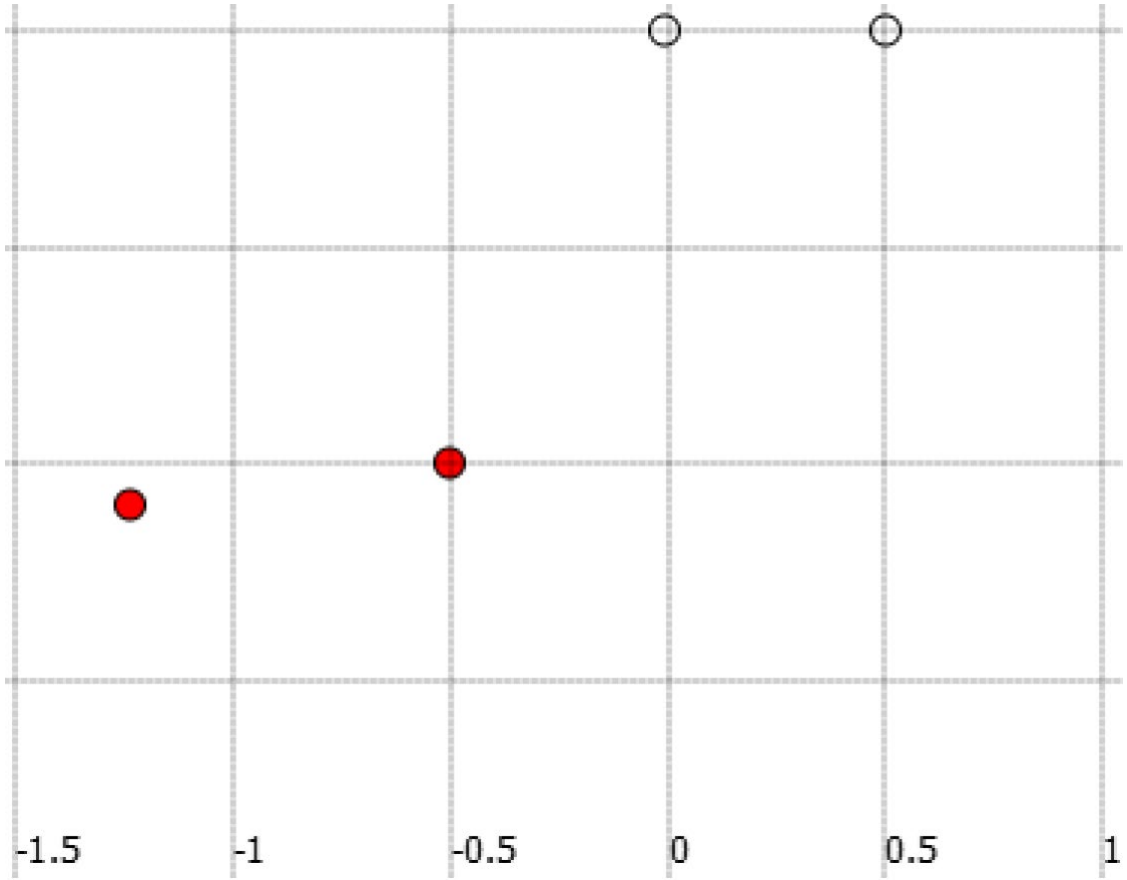
Q2 > B – Case 1

- Draw the isolines and the separating hyperplane.



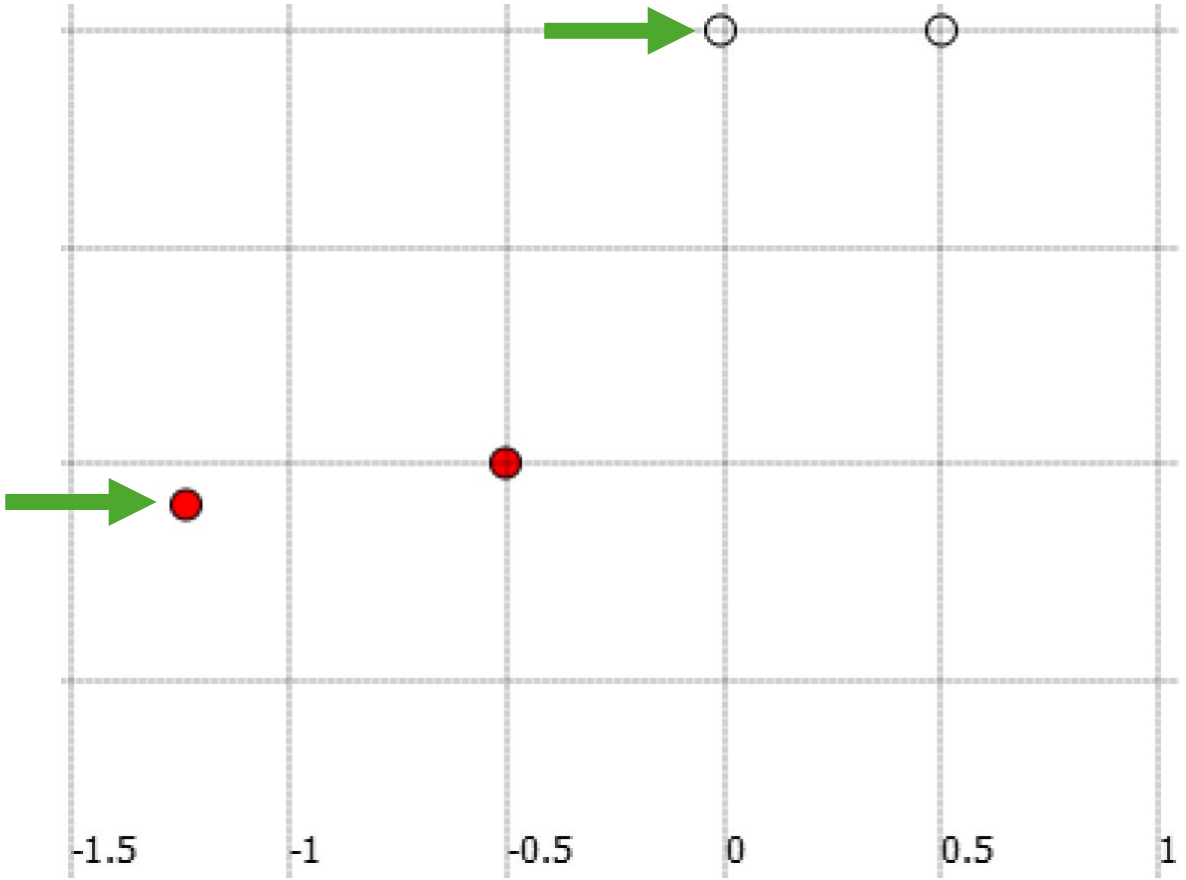
Q2 > B – Case 2

- Draw the new separating hyperplane.



Q2 > B – Case 2

- Draw the new separating hyperplane.

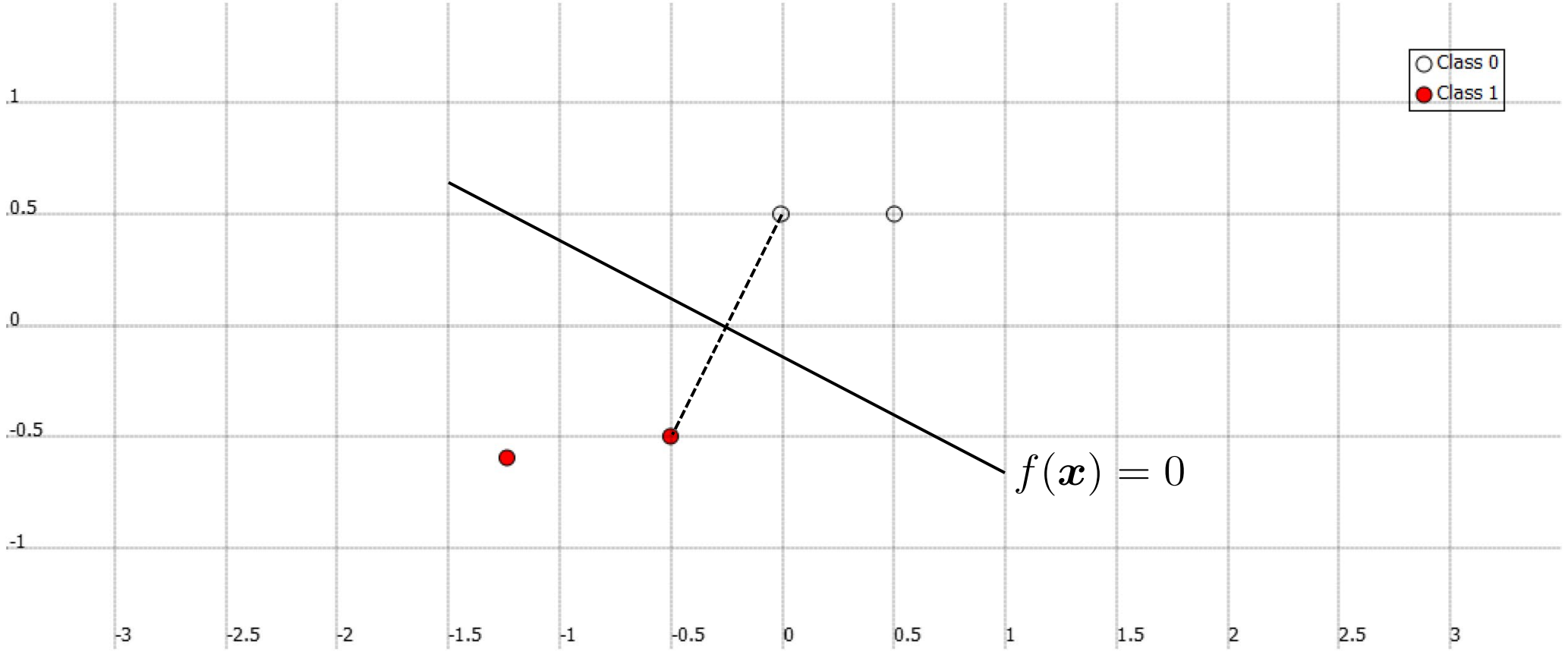


Q2 > B – Case 2

- Since the point added to the white class is inside the original margin, it now becomes a support vector instead of the original point in the white class.

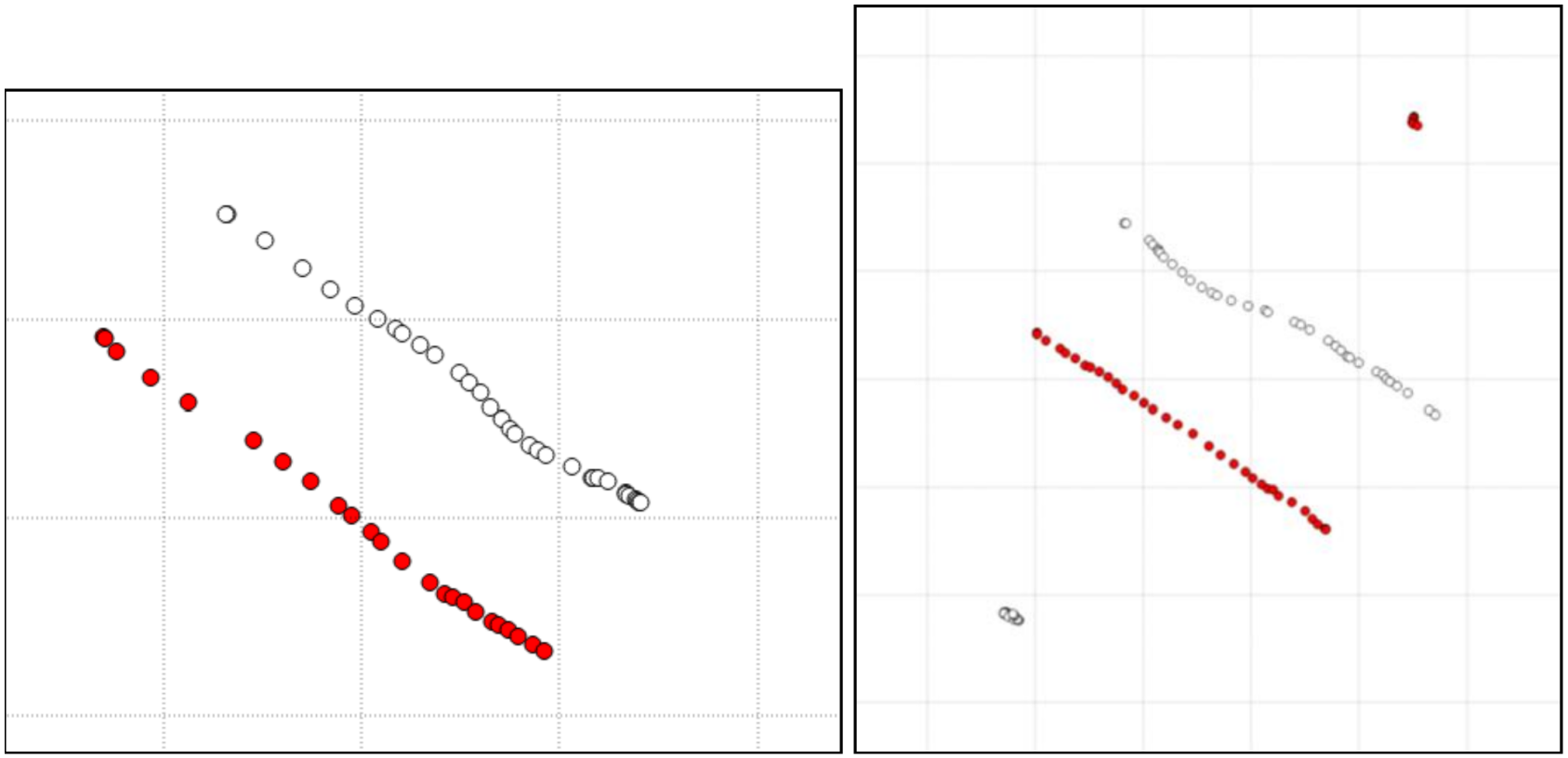
Q2 > B – Case 2

- Since the point added to the white class is inside the original margin, it now becomes a support vector instead of the original point in the white class.



Q2 > C

- Draw the separating hyperplane and discuss the associated effects of the values of the hyperparameters (penalty factor and kernel width).

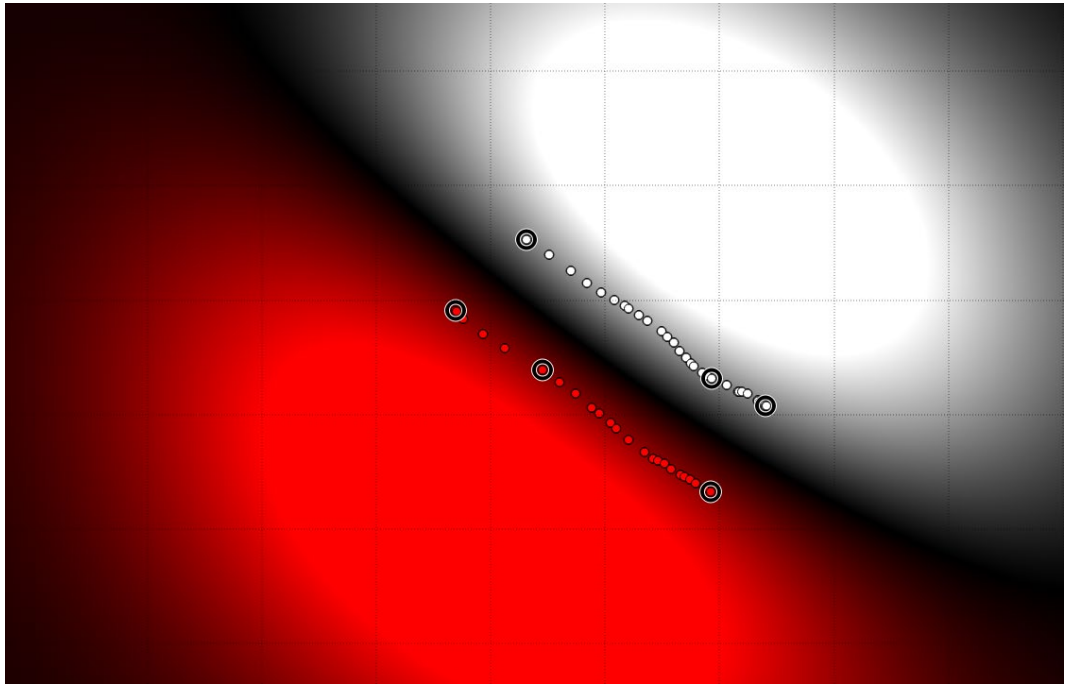


(i) (ii)

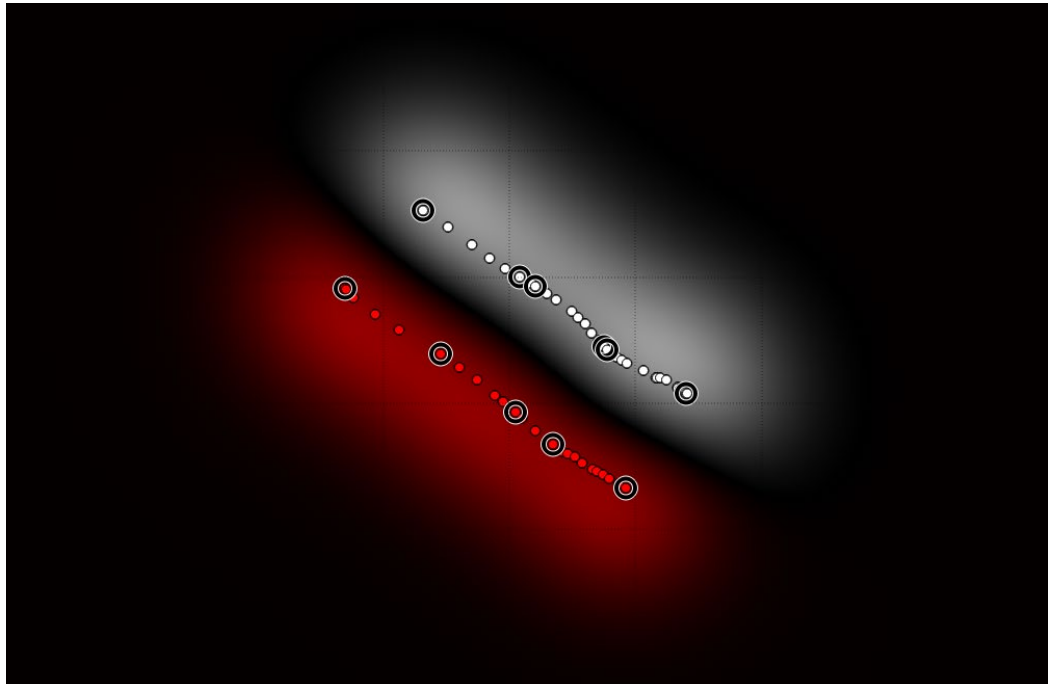
Q2 > C – Case i

- The separating line is unaffected by the value of the penalty C.
- The kernel width affects the number of support vectors.

$\sigma = 0.1$



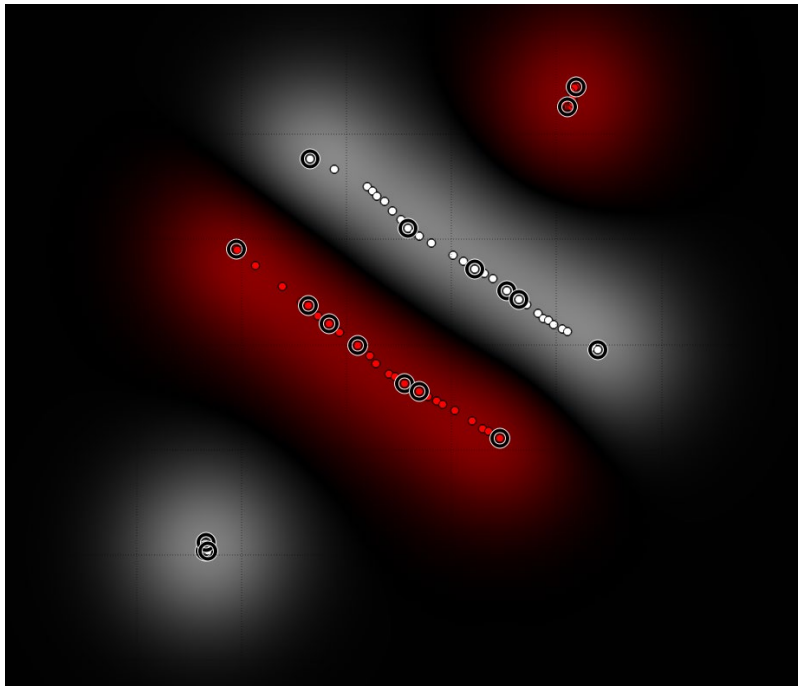
$\sigma = 0.01$



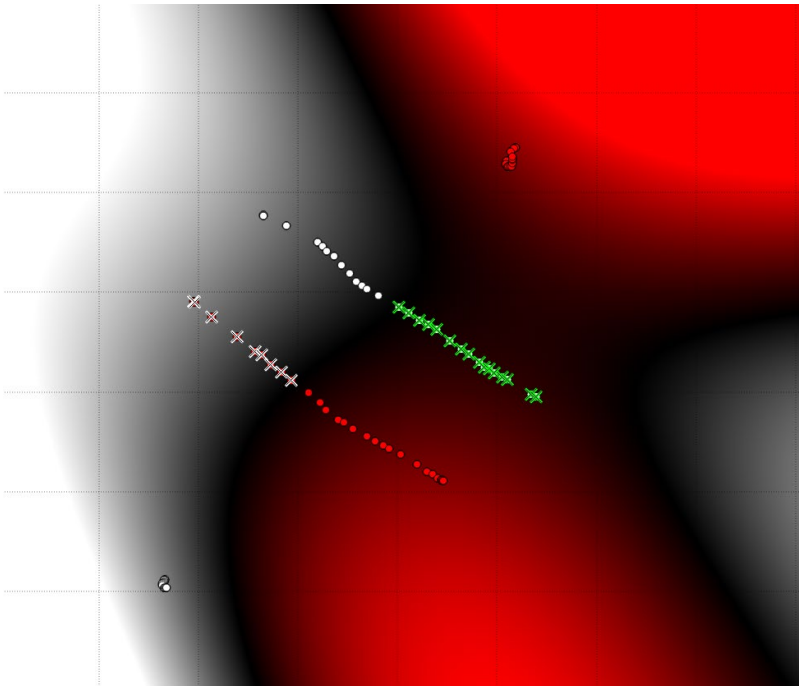
Q2 > C – Case ii

- The separating line will be influenced by both penalty C and the kernel width.

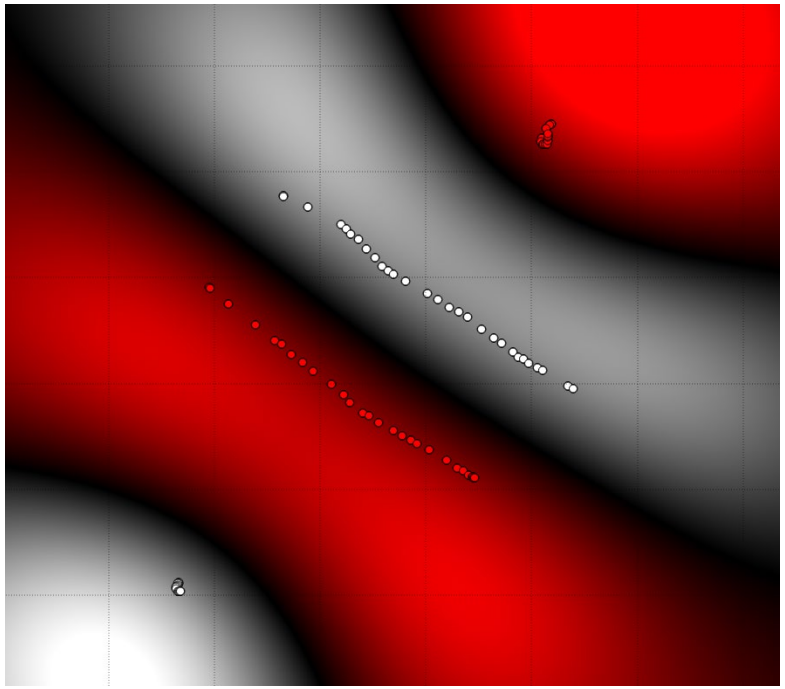
$\sigma = 0.01; C = 5000$



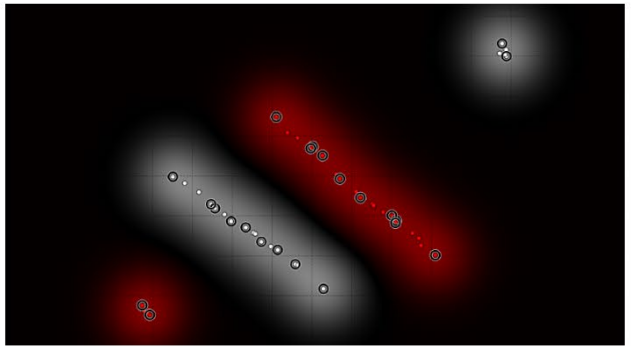
$\sigma = 0.5; C = 10$



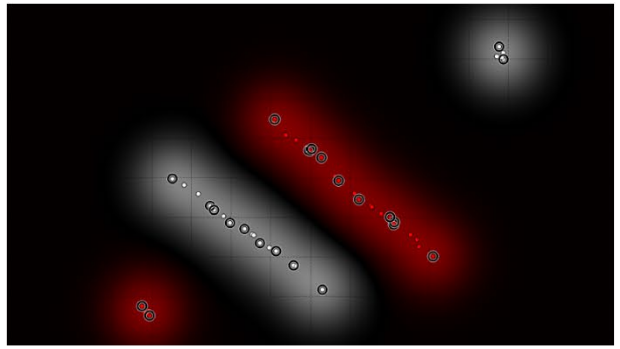
$\sigma = 0.1; C = 1000$



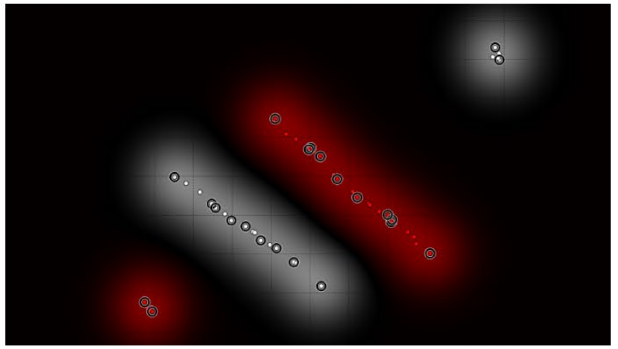
Q2 > C – Case ii



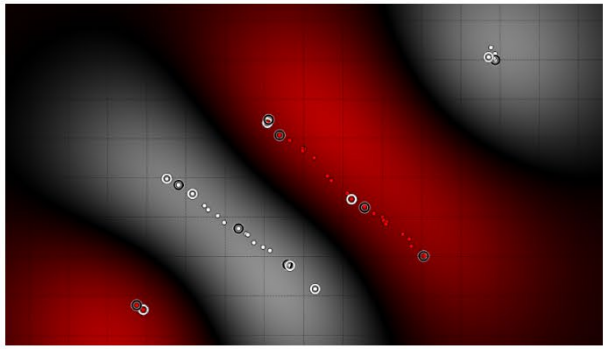
(a) $\sigma = 0.01; C = 1$



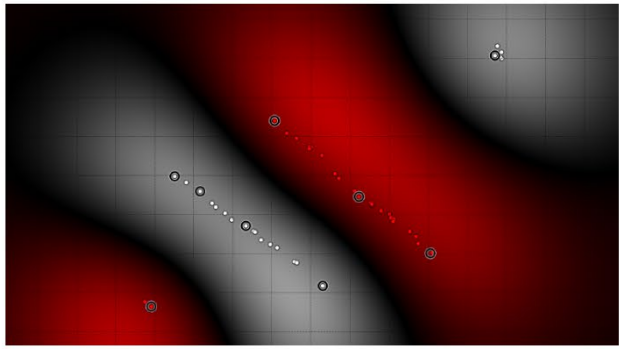
(b) $\sigma = 0.01; C = 10$



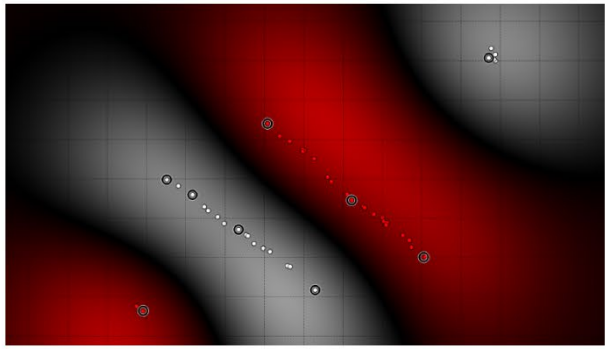
(c) $\sigma = 0.01; C = 5000$



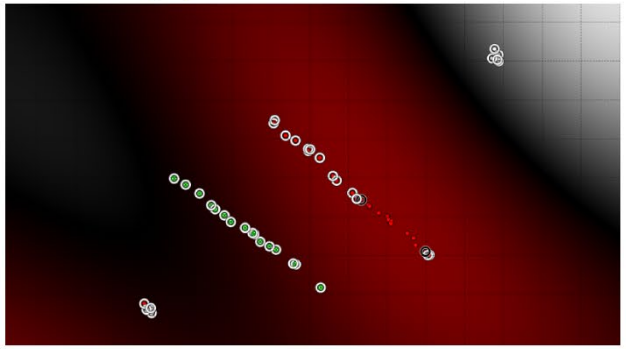
(d) $\sigma = 0.1; C = 1$



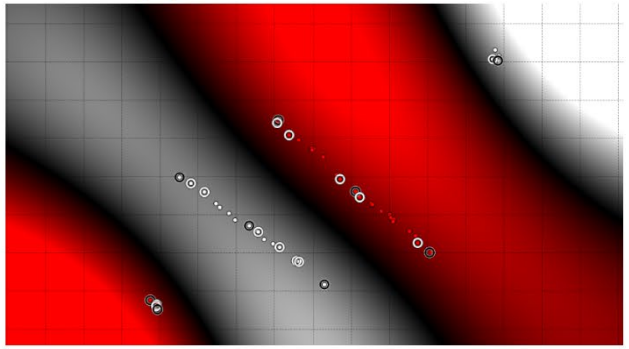
(e) $\sigma = 0.1; C = 10$



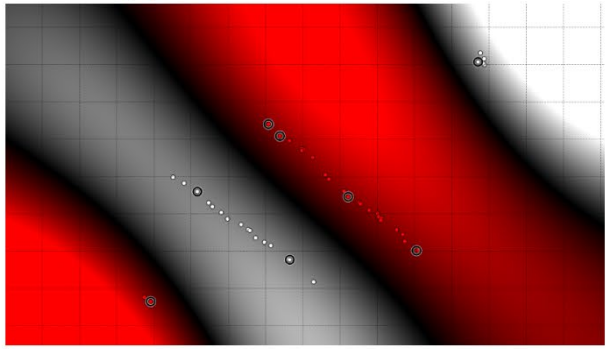
(f) $\sigma = 0.1; C = 5000$



(g) $\sigma = 0.5; C = 1$



(h) $\sigma = 0.5; C = 10$



(i) $\sigma = 0.5; C = 5000$

Q3 > A

- Uniqueness of $\boldsymbol{w} = \sum_{i=1}^{K} \alpha_i y_i \boldsymbol{x}^i$ with multiple ways of construction.
- N=2 and three non-zero non-colinear points $\boldsymbol{x^i}$ on the margin

Q3 > A

- Uniqueness of $\boldsymbol{w} = \sum_{i=1}^{K} \alpha_i y_i \boldsymbol{x}^i$ with multiple ways of construction.
- N=2 and three non-zero non-colinear points $\boldsymbol{x^i}$ on the margin
- Let optimal $\boldsymbol{w}$ be $\boldsymbol{w} = \alpha_1 y_1 \boldsymbol{x}^1 + \alpha_2 y_2 \boldsymbol{x}^2$.

Q3 > A

- Uniqueness of $\boldsymbol{w} = \sum_{i=1}^{K} \alpha_i y_i \boldsymbol{x}^i$ with multiple ways of construction.
- N=2 and three non-zero non-colinear points $\boldsymbol{x^i}$ on the margin
- Let optimal $\boldsymbol{w}$ be $\boldsymbol{w} = \alpha_1 y_1 \boldsymbol{x}^1 + \alpha_2 y_2 \boldsymbol{x}^2$.
- Linear independence of any pair of two $\boldsymbol{x}^i$
- Consider $\boldsymbol{x}^2 = \beta_1 \boldsymbol{x}^1 + \beta_3 \boldsymbol{x}^3$.

Q3 > A

- Uniqueness of $\boldsymbol{w} = \sum_{i=1}^{K} \alpha_i y_i \boldsymbol{x}^i$ with multiple ways of construction.
- N=2 and three non-zero non-colinear points $\boldsymbol{x}^i$ on the margin
- Let optimal $\boldsymbol{w}$ be $\boldsymbol{w} = \alpha_1 y_1 \boldsymbol{x}^1 + \alpha_2 y_2 \boldsymbol{x}^2$.
- Linear independence of any pair of two $\boldsymbol{x}^i$
- Consider $\boldsymbol{x}^2 = \beta_1 \boldsymbol{x}^1 + \beta_3 \boldsymbol{x}^3$.

$$\boldsymbol{w} = \underbrace{(\alpha_1 y_1 + \alpha_2 y_2 \beta_1)}_{\alpha'_1 y_1} \boldsymbol{x}^1 + \underbrace{\alpha_2 y_2 \beta_3}_{\alpha'_3 y_3} \boldsymbol{x}^3 = \alpha'_1 y_1 \boldsymbol{x}^1 + \alpha'_3 y_3 \boldsymbol{x}^3$$

Q3 > B

- The constant 1 is arbitrary in $\boldsymbol{w}^\top \boldsymbol{x} + b = \pm 1$.

Q3 > B

- The constant 1 is arbitrary in $\boldsymbol{w}^\top \boldsymbol{x} + b = \pm 1$.
- Given $\sum_i \alpha_i y_i = 0$, we have at least 1 SV per class with $y_i(\boldsymbol{w}^\top \boldsymbol{x}^i + b) = 1$.

Q3 > B

- The constant 1 is arbitrary in $\boldsymbol{w}^\top \boldsymbol{x} + b = \pm 1$.
- Given $\sum_i \alpha_i y_i = 0$, we have at least 1 SV per class with $y_i(\boldsymbol{w}^\top \boldsymbol{x}^i + b) = 1$.
- For $a > 0$, we can have $y_i(\boldsymbol{w}^\top \boldsymbol{x}^i + b) = a$.

$$\begin{cases} \boldsymbol{w}^\top \boldsymbol{x}^1 + b = a \\ \boldsymbol{w}^\top \boldsymbol{x}^2 + b = -a \end{cases} \longrightarrow \boldsymbol{w}^\top (\boldsymbol{x}^1 - \boldsymbol{x}^2) = 2a$$

Q3 > B

- The constant 1 is arbitrary in $\boldsymbol{w}^\top \boldsymbol{x} + b = \pm 1$.
- Given $\sum_i \alpha_i y_i = 0$, we have at least 1 SV per class with $y_i(\boldsymbol{w}^\top \boldsymbol{x}^i + b) = 1$.
- For $a > 0$, we can have $y_i(\boldsymbol{w}^\top \boldsymbol{x}^i + b) = a$.

$$\begin{cases} \boldsymbol{w}^\top \boldsymbol{x}^1 + b = a \\ \boldsymbol{w}^\top \boldsymbol{x}^2 + b = -a \end{cases} \longrightarrow \boldsymbol{w}^\top (\boldsymbol{x}^1 - \boldsymbol{x}^2) = 2a$$

$$\|\boldsymbol{w}\| = \frac{2a}{\|(\boldsymbol{x}^1 - \boldsymbol{x}^2)\| \cos(\theta)}$$

- Factor a only scales the norm of $\boldsymbol{w}$, not its direction and not the choices of SVs.

Q3 > C

- Prove the convexity of the relaxed linear SVM optimization problem.

$$
\begin{aligned}
\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad & \tfrac{1}{2}\|\boldsymbol{w}\|^2 + \tfrac{C}{M}\sum_{i=1}^{M}\xi_i \\
\text{s.t.} \quad & y_i(\boldsymbol{w}^\top \boldsymbol{x}^i + b) \geq 1 - \xi_i \quad \forall i = 1, \ldots, M \\
& \xi_i \geq 0 \quad \forall i = 1, \ldots, M
\end{aligned}
$$

Q3 > C

- Prove the convexity of the relaxed linear SVM optimization problem.

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{M}\sum_{i=1}^{M}\xi_i \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^\top \boldsymbol{x}^i + b) \geq 1 - \xi_i \quad \forall i = 1, \ldots, M \\ & \xi_i \geq 0 \quad \forall i = 1, \ldots, M \end{aligned}$$

- Convex: $f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq f(\lambda\boldsymbol{x}) + f((1-\lambda)\boldsymbol{y})$ for $\boldsymbol{\lambda \in [0,1]}$
- Strictly convex: $f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) < f(\lambda\boldsymbol{x}) + f((1-\lambda)\boldsymbol{y})$ for $\boldsymbol{\lambda \in [0,1]}$ and $\boldsymbol{x \neq y}$

Q3 > C

For two arbitrary vectors $\boldsymbol{x}, \boldsymbol{y}$ and $\lambda \in [0,1]$,

$$\begin{aligned}\|\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}\|^2 &= (\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y})^\top (\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \\ &= \lambda^2\|\boldsymbol{x}\|^2 + (1-\lambda)^2\|\boldsymbol{y}\|^2 + 2\lambda(1-\lambda)\boldsymbol{x}^\top\boldsymbol{y}.\end{aligned}$$

We now compare this with $\lambda\|\boldsymbol{x}\|^2 + (1-\lambda)\|\boldsymbol{y}\|^2$:

$$\begin{aligned}\lambda\|\boldsymbol{x}\|^2 + (1-\lambda)\|\boldsymbol{y}\|^2 - \|\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}\|^2 &= \lambda(1-\lambda)\left(\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - 2\boldsymbol{x}^\top\boldsymbol{y}\right) \\ &= \lambda(1-\lambda)\|\boldsymbol{x}-\boldsymbol{y}\|^2 \geq 0.\end{aligned}$$

Hence,

$$\boxed{\|\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}\|^2 \leq \lambda\|\boldsymbol{x}\|^2 + (1-\lambda)\|\boldsymbol{y}\|^2, \quad \forall\, \lambda \in [0,1],}$$

which proves that $\boxed{f(\boldsymbol{x}) = \|\boldsymbol{x}\|^2 \text{ is convex.}}$

Moreover, if $\boldsymbol{x} \neq \boldsymbol{y}$ and $0 < \lambda < 1$, then $\|\boldsymbol{x}-\boldsymbol{y}\|^2 > 0$, and the inequality becomes strict:

$$\boxed{\|\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}\|^2 < \lambda\|\boldsymbol{x}\|^2 + (1-\lambda)\|\boldsymbol{y}\|^2.}$$

Therefore, $\boxed{f(\boldsymbol{x}) = \|\boldsymbol{x}\|^2 \text{ is strictly convex.}}$

Q3 > C

- Prove the convexity of the relaxed linear SVM optimization problem.

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{M}\sum_{i=1}^{M}\xi_i \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^\top \boldsymbol{x}^i + b) \geq 1 - \xi_i \quad \forall i = 1, \ldots, M \\ & \xi_i \geq 0 \quad \forall i = 1, \ldots, M \end{aligned}$$

- The objective is convex for all decision variables and strictly convex in $\boldsymbol{w}$.

Q3 > C

- Prove the convexity of the relaxed linear SVM optimization problem.

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{M}\sum_{i=1}^{M}\xi_i$$

$$\text{s.t.} \quad y_i(\boldsymbol{w}^\top \boldsymbol{x}^i + b) \geq 1 - \xi_i \quad \forall i = 1, \ldots, M$$

$$\xi_i \geq 0 \quad \forall i = 1, \ldots, M$$

- The objective is convex for all decision variables and strictly convex in $\boldsymbol{w}$.
- Inequality constraints are affine in decision variables.

Q3 > C

- Prove the convexity of the relaxed linear SVM optimization problem.

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{M}\sum_{i=1}^{M}\xi_i$$

$$\text{s.t.} \quad y_i(\boldsymbol{w}^\top \boldsymbol{x}^i + b) \geq 1 - \xi_i \quad \forall i = 1, \ldots, M$$

$$\xi_i \geq 0 \quad \forall i = 1, \ldots, M$$

- The objective is convex for all decision variables and strictly convex in $\boldsymbol{w}$.
- Inequality constraints are affine in decision variables.
- The relaxed problem is convex.

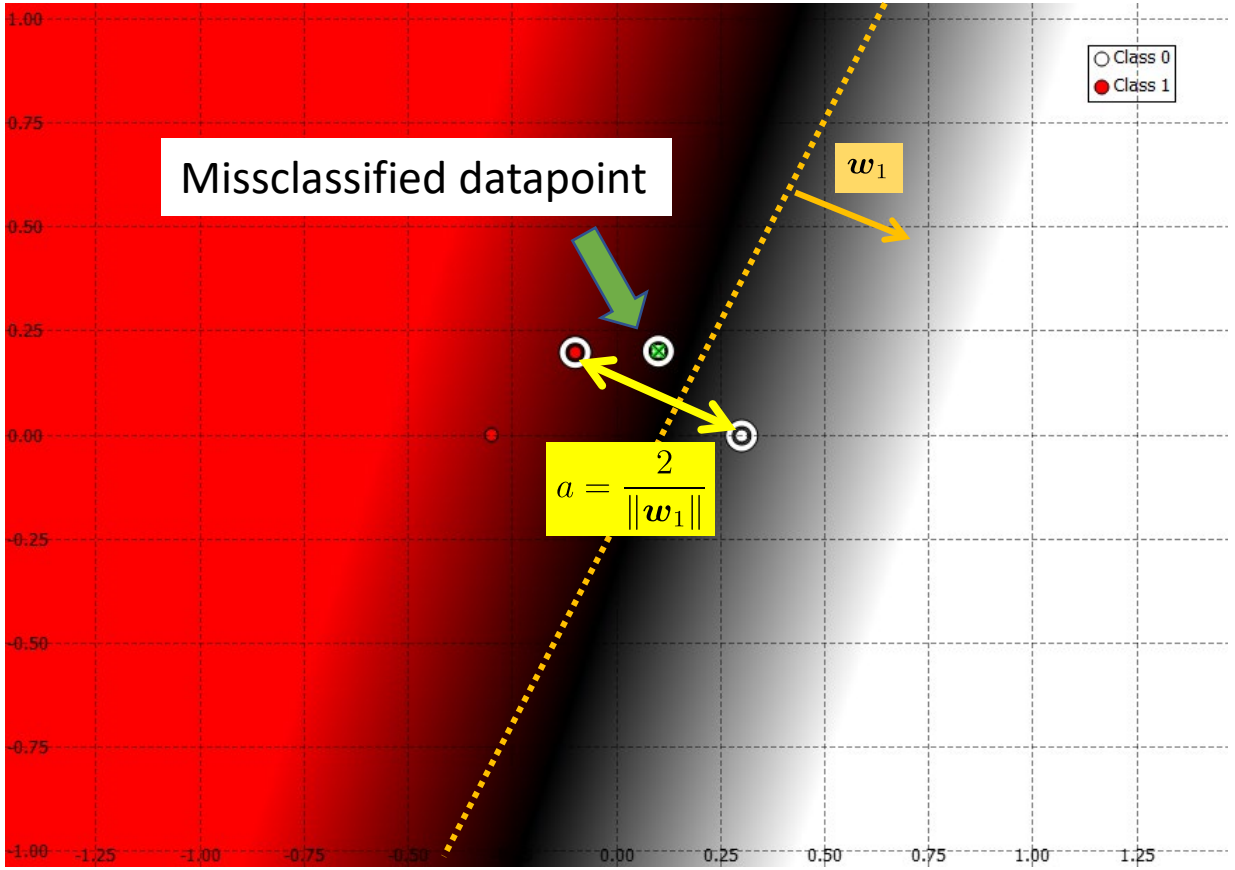
Q3 > C

- Investing the optimum in unrelaxed and relaxed linear SVM.

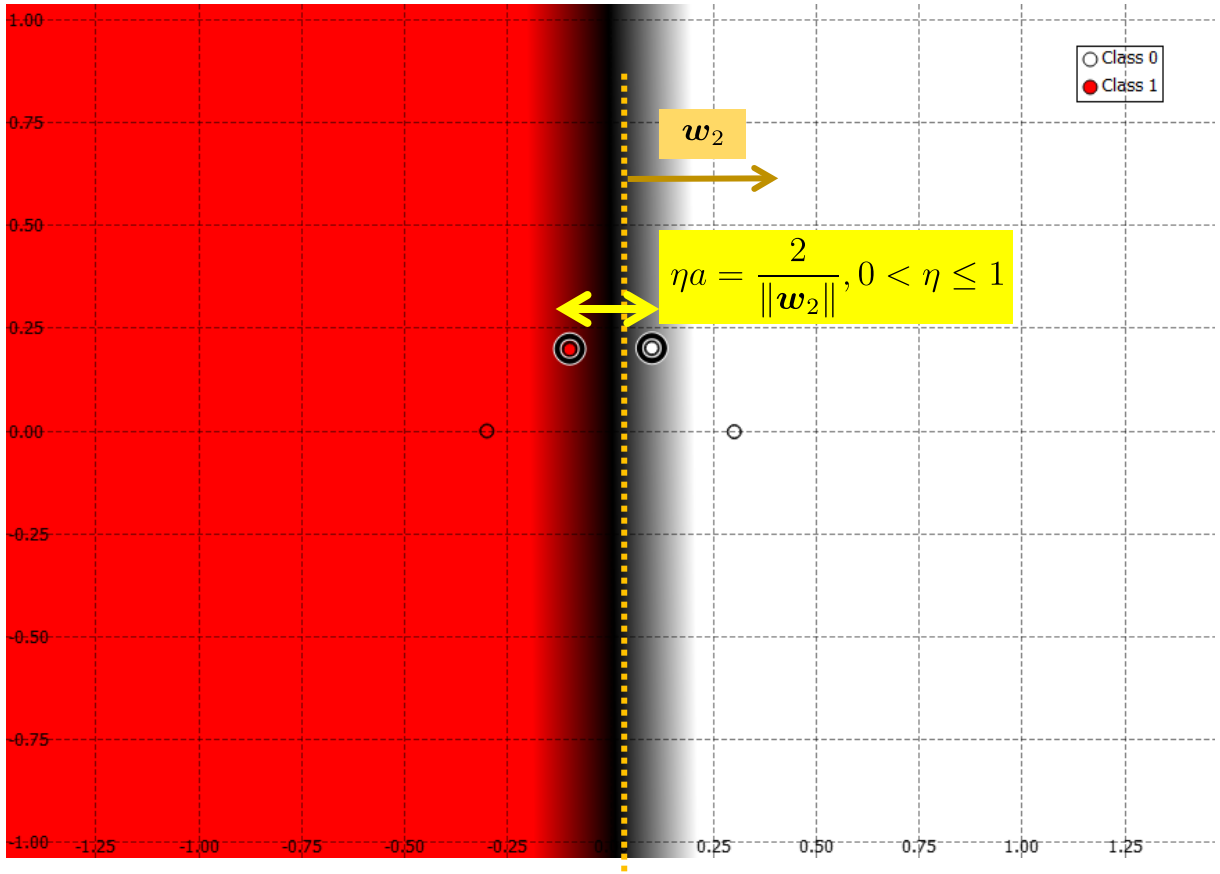
$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{M}\sum_{i=1}^{M}\xi_i$$
$$\text{s.t.} \quad y_i(\boldsymbol{w}^\top \boldsymbol{x}^i + b) \geq 1 - \xi_i \quad \forall i = 1, \ldots, M$$
$$\xi_i \geq 0 \quad \forall i = 1, \ldots, M$$

- Tradeoff between enlarging the margin and reducing the constraint violation cost

Q3 > C



Small penalty C=5, larger margin and 1 misclassified datapoint



C=100

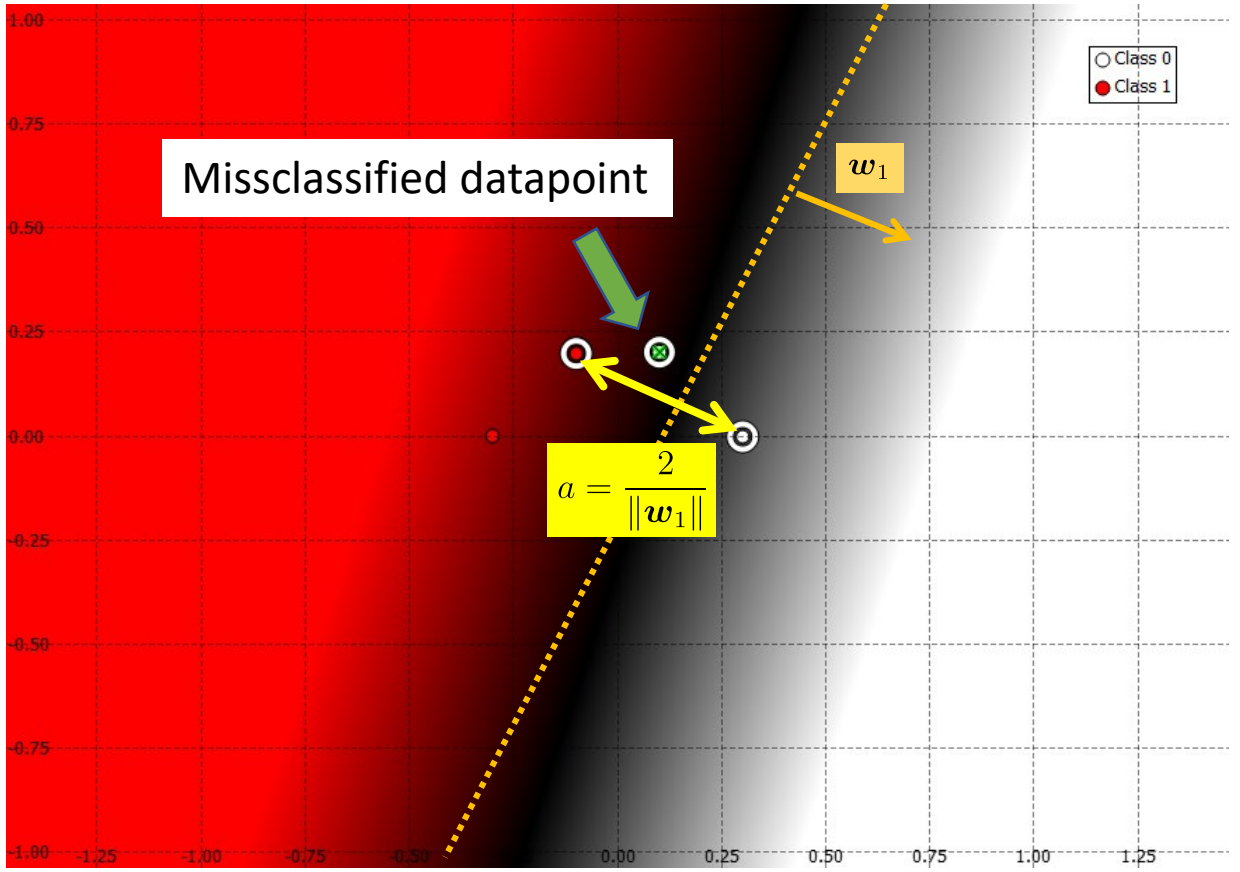
Q3 > C

- Assume b=0.
- For the misclassified datapoint:

$$\xi = 1 - y_i \boldsymbol{w}_1^\top \boldsymbol{x}^i = 1 - y_i \|\boldsymbol{w_1}\| \|\boldsymbol{x}^i\| \cos(\theta)$$

$$\xi = 1 - y_i \boldsymbol{w}_1^\top \boldsymbol{x}^i = 1 - y_i \|\boldsymbol{w}_1\| \underbrace{\frac{\boldsymbol{w}_1^\top \boldsymbol{x}^i}{\|\boldsymbol{w}_1\|}}_{d_i} = 1 - \|\boldsymbol{w}_1\| y_i d_i$$

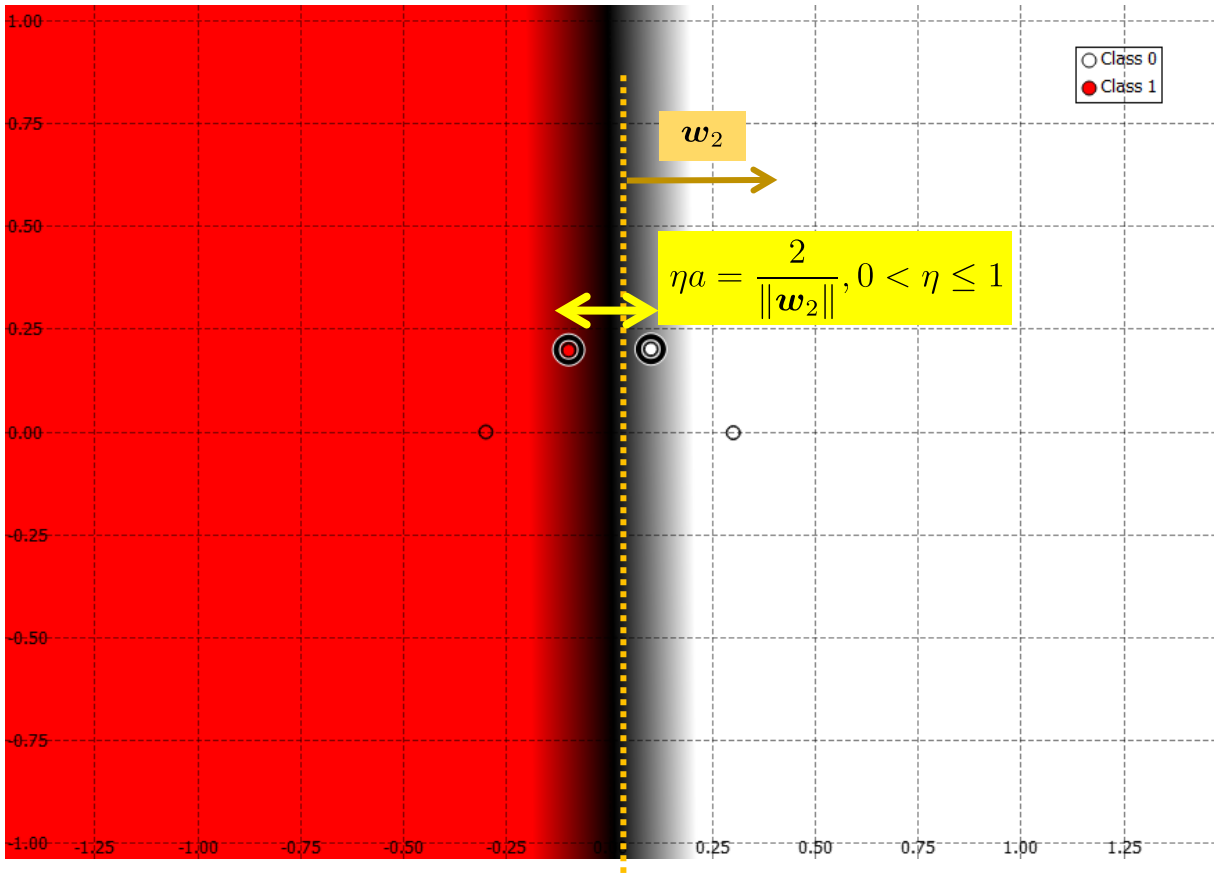
- Slack $\boldsymbol{\xi}$ varies linearly with $\|\boldsymbol{w}_1\|$ and distance to the hyperplane.



Small penalty C=5, larger margin and 1 misclassified datapoint

Q3 > C

- $\xi_i = 0 \ \ \forall i$



C=100

Q3 > C

- Dependece on C and η

$$\frac{1}{2}\|\boldsymbol{w}_1\|^2 + \frac{C}{M}\sum_i \xi_i \overset{?}{\leq} \frac{1}{2}\|\boldsymbol{w}_2\|^2 \xrightarrow{\sum_i \xi_i = \xi} \frac{2}{a^2} + \frac{C}{M}\xi \overset{?}{\leq} \frac{2}{(\eta a)^2}$$

Q3 > C

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad & \tfrac{1}{2}\|\boldsymbol{w}\|^2 + \tfrac{C}{M}\sum_{i=1}^{M}\xi_i \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^\top \boldsymbol{x}^i + b) \geq 1 - \xi_i \quad \forall i = 1, \ldots, M \\ & \xi_i \geq 0 \quad \forall i = 1, \ldots, M \end{aligned}$$