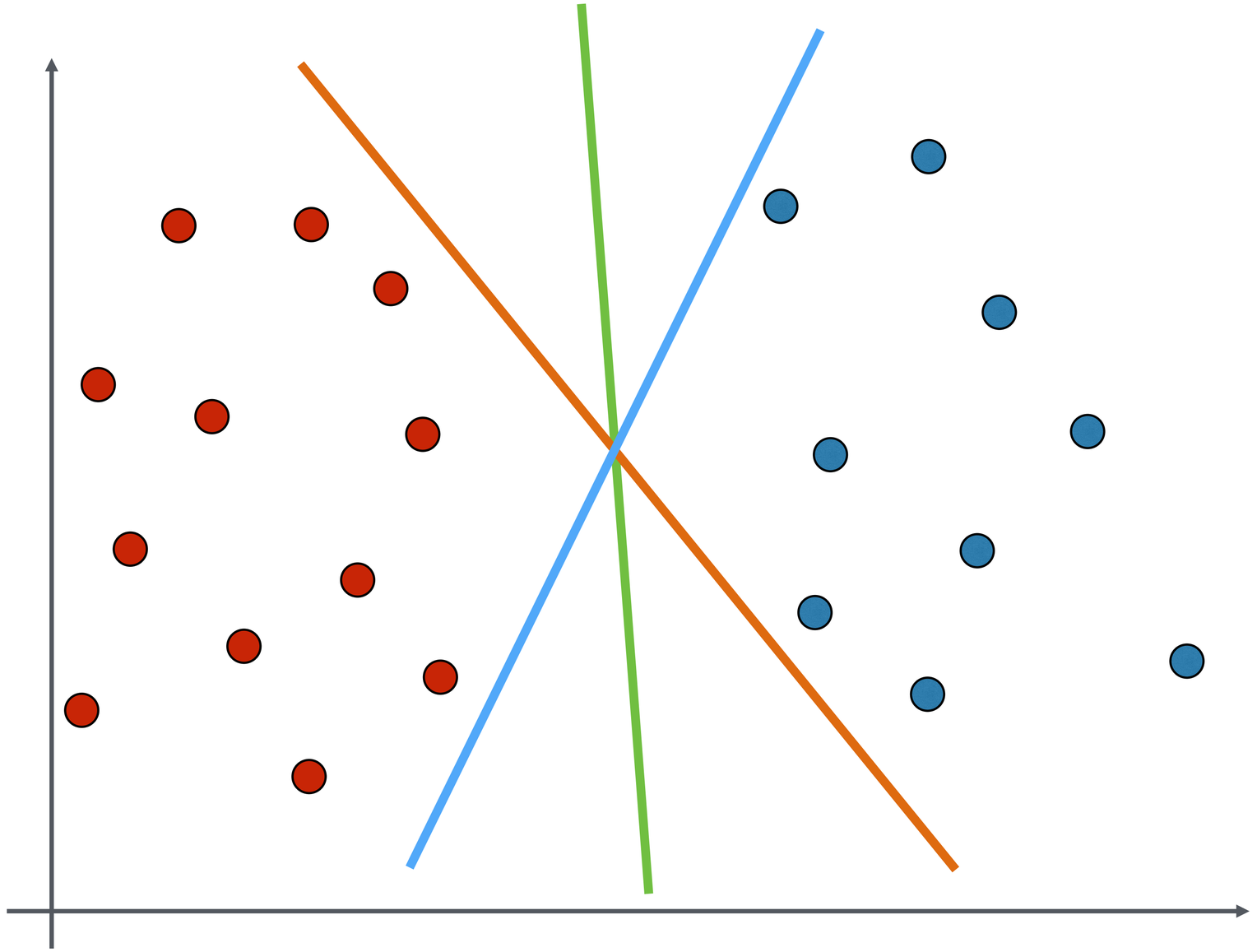
Statistical Methods in AI (CSE 471)
SVM

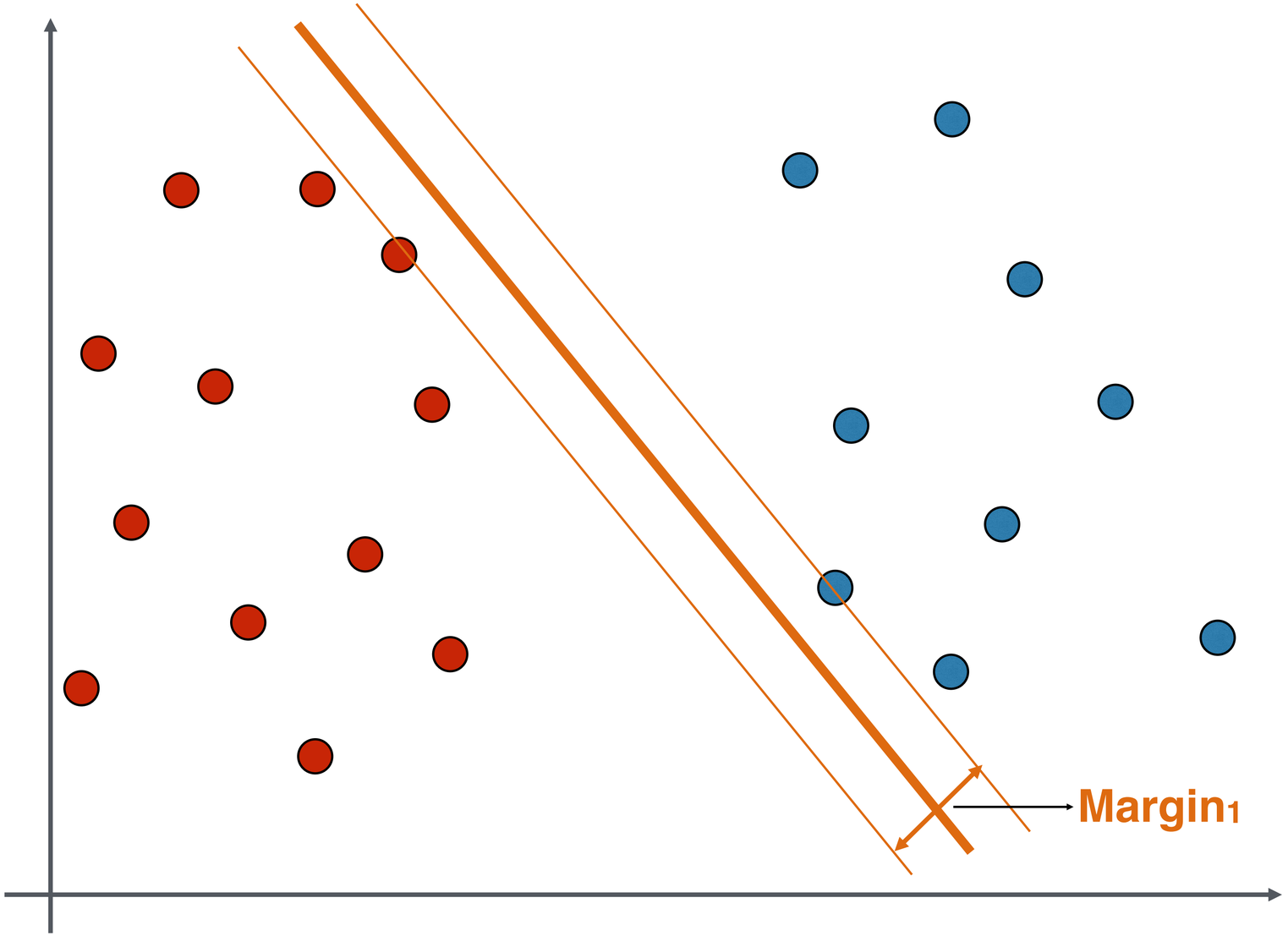
Vineet Gandhi
Centre for Visual Information Technology (CVIT)

Multiple solutions exist for linearly separable data



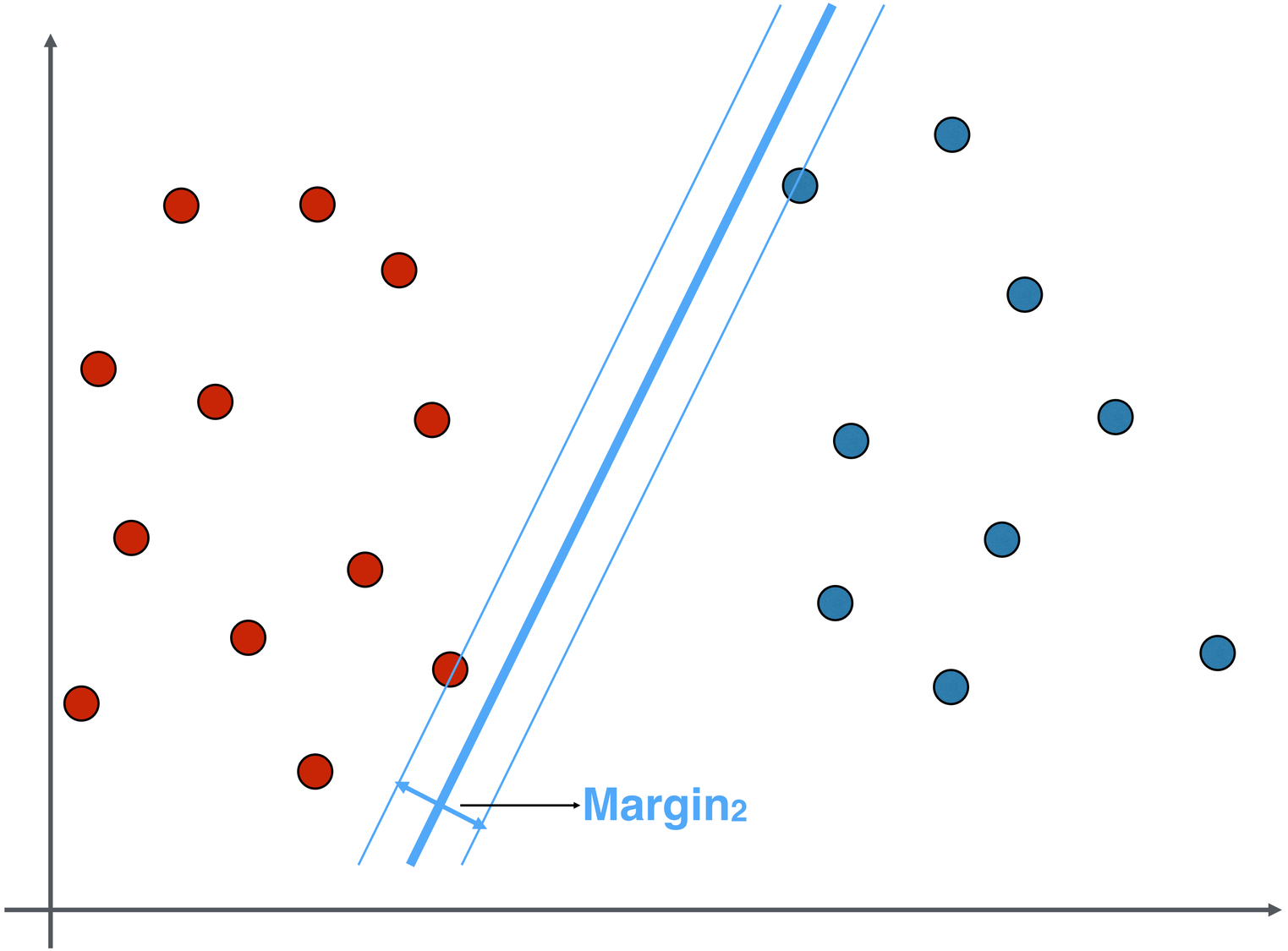
Are all solutions equally good?

Margin: No-mans Band



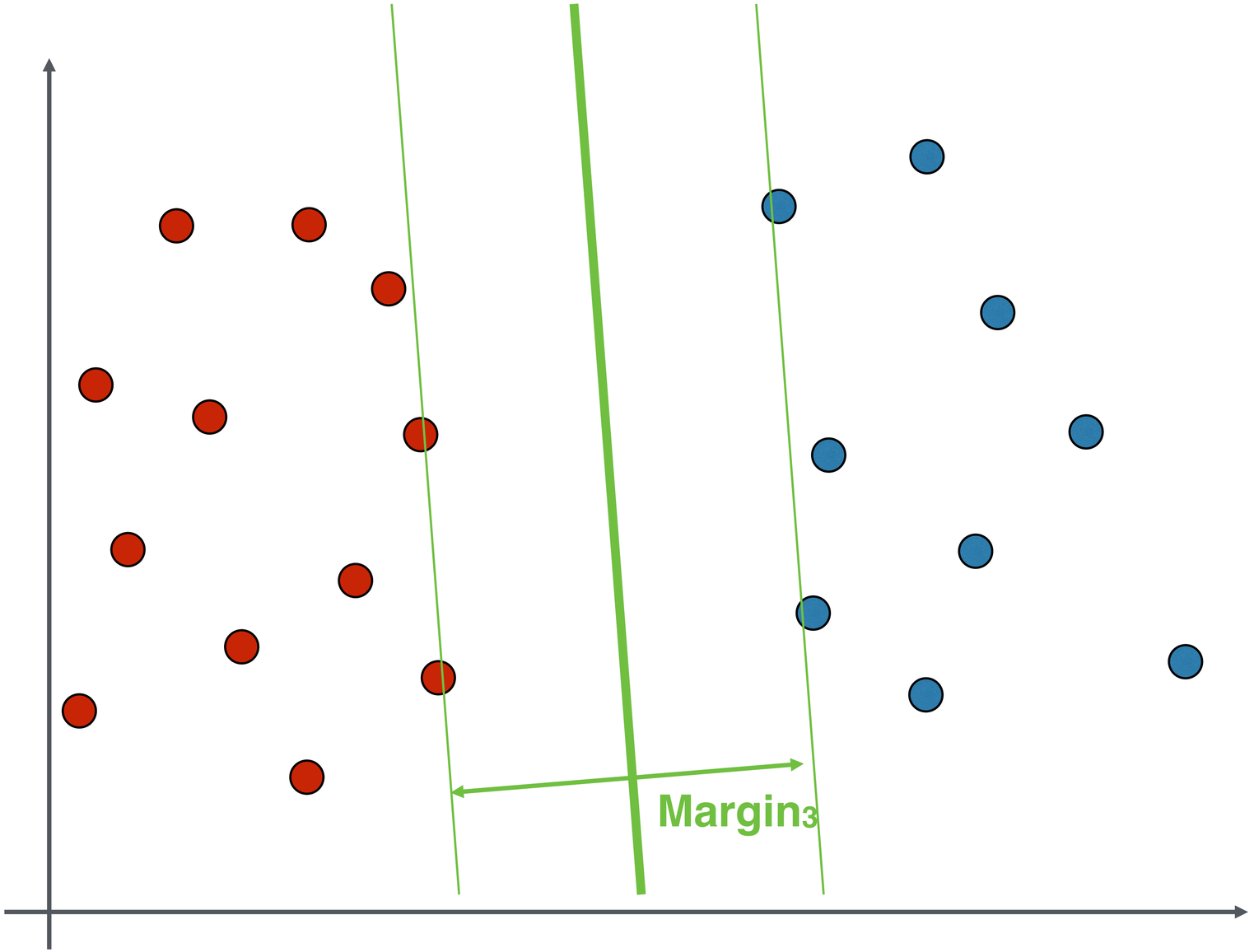
Margin: Width of a band around decision boundary without any training samples

Multiple solutions exist for linearly separable data



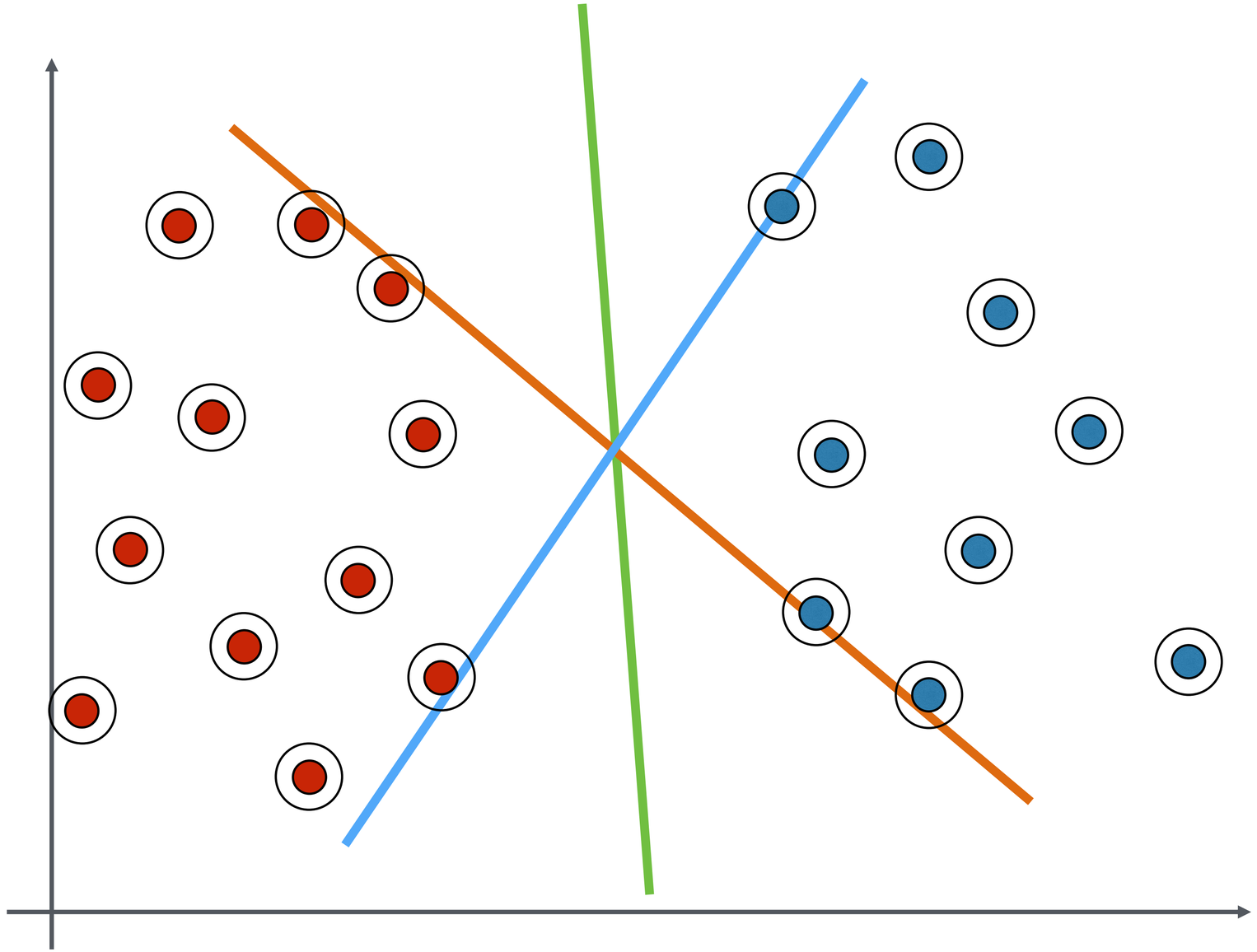
Margin: Width of a band around decision boundary without any training samples

Multiple solutions exist for linearly separable data



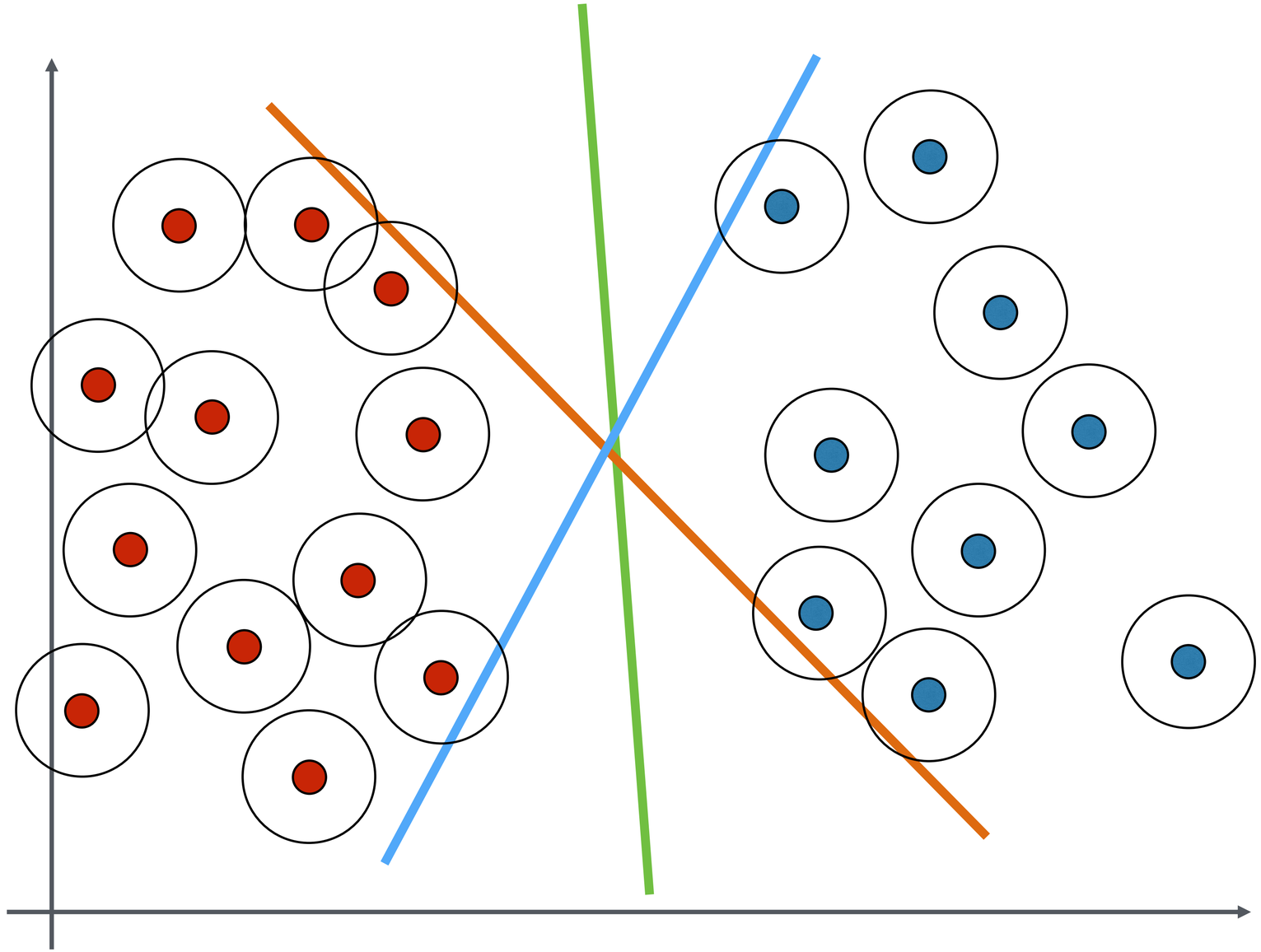
Is a Larger Margin better? Why?

Margin: Bubbles around samples



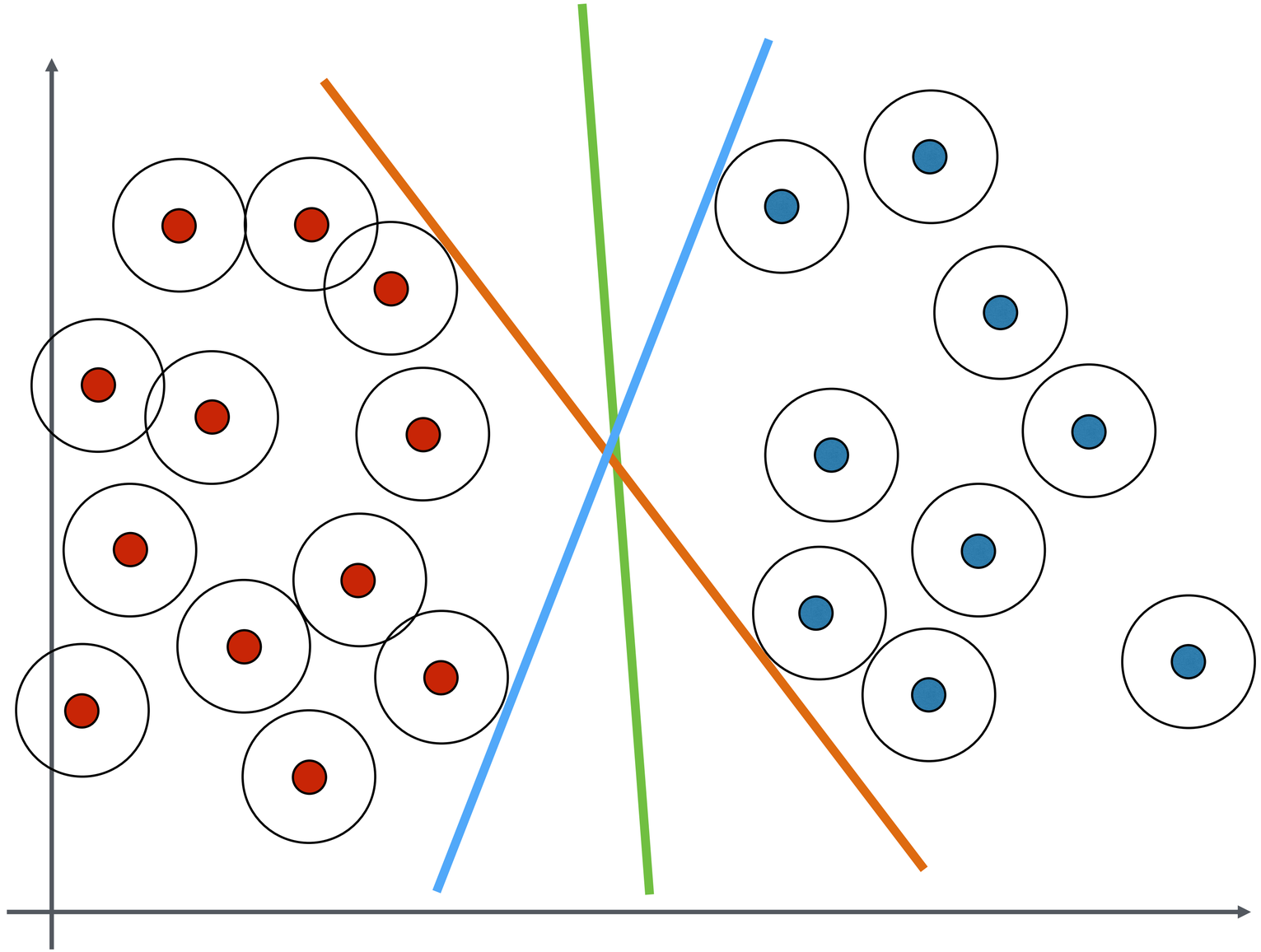
Margin: Radius of a region around each training sample, through which the decision boundary cannot pass

Margin: Bubbles around samples



Margin: Radius of a region around each training sample, through which the decision boundary cannot pass

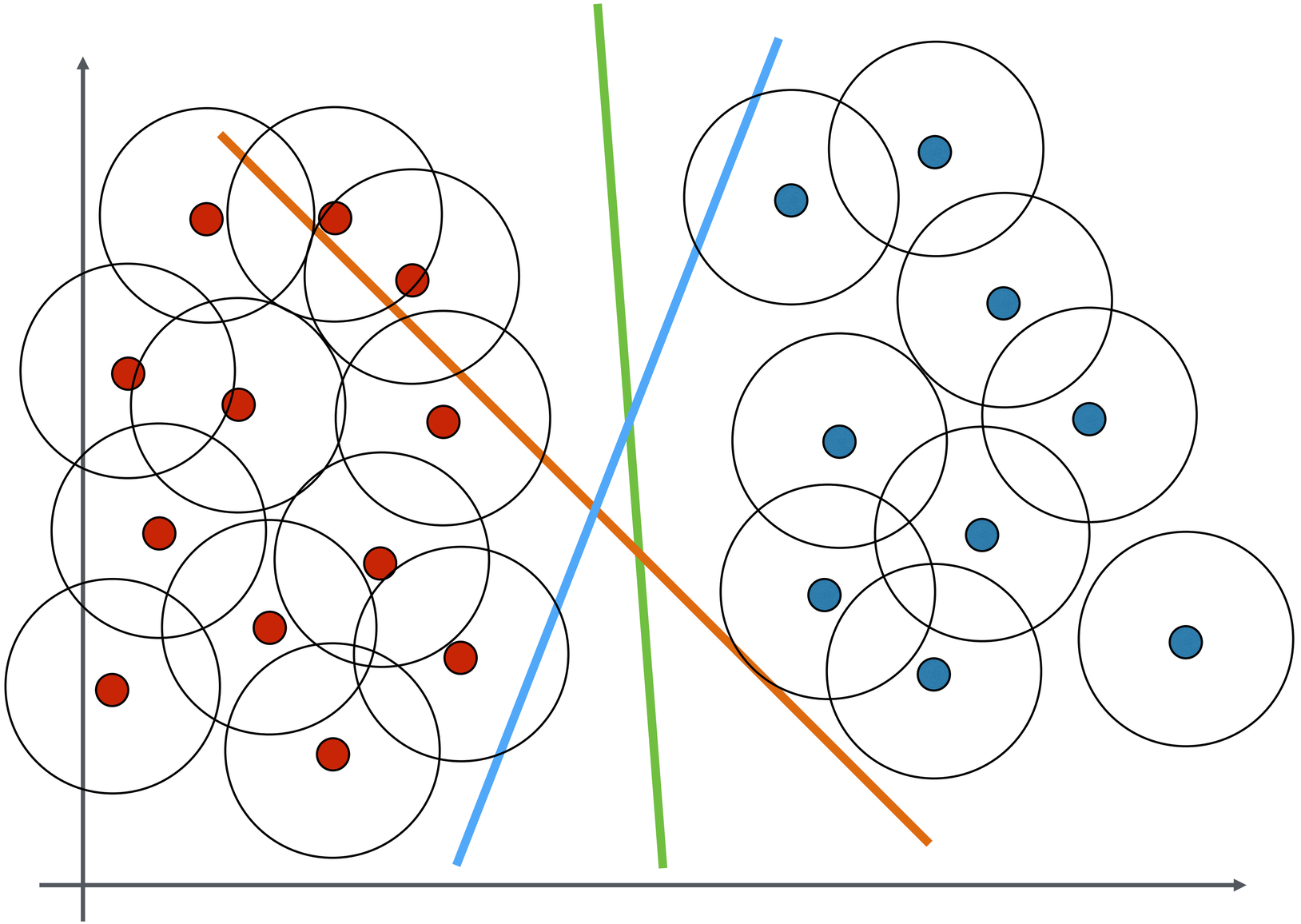
Margin: Bubbles around samples



Margin: Radius of a region around each training sample, through which the decision boundary cannot pass

As the margin increases, the feasible region reduces

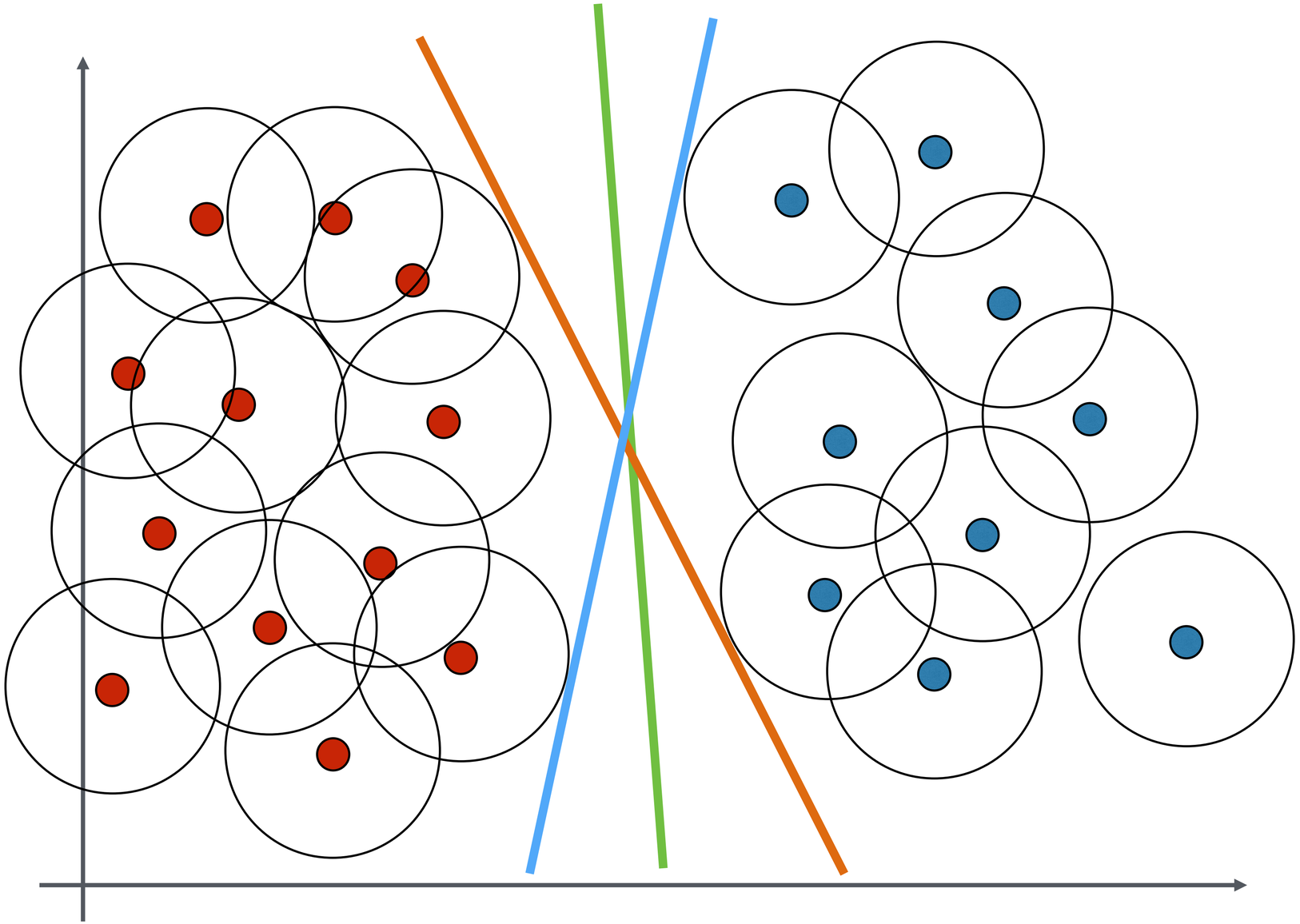
Margin: Bubbles around samples



Margin: Radius of a region around each training sample, through which the decision boundary cannot pass

As the margin increases, the feasible region reduces

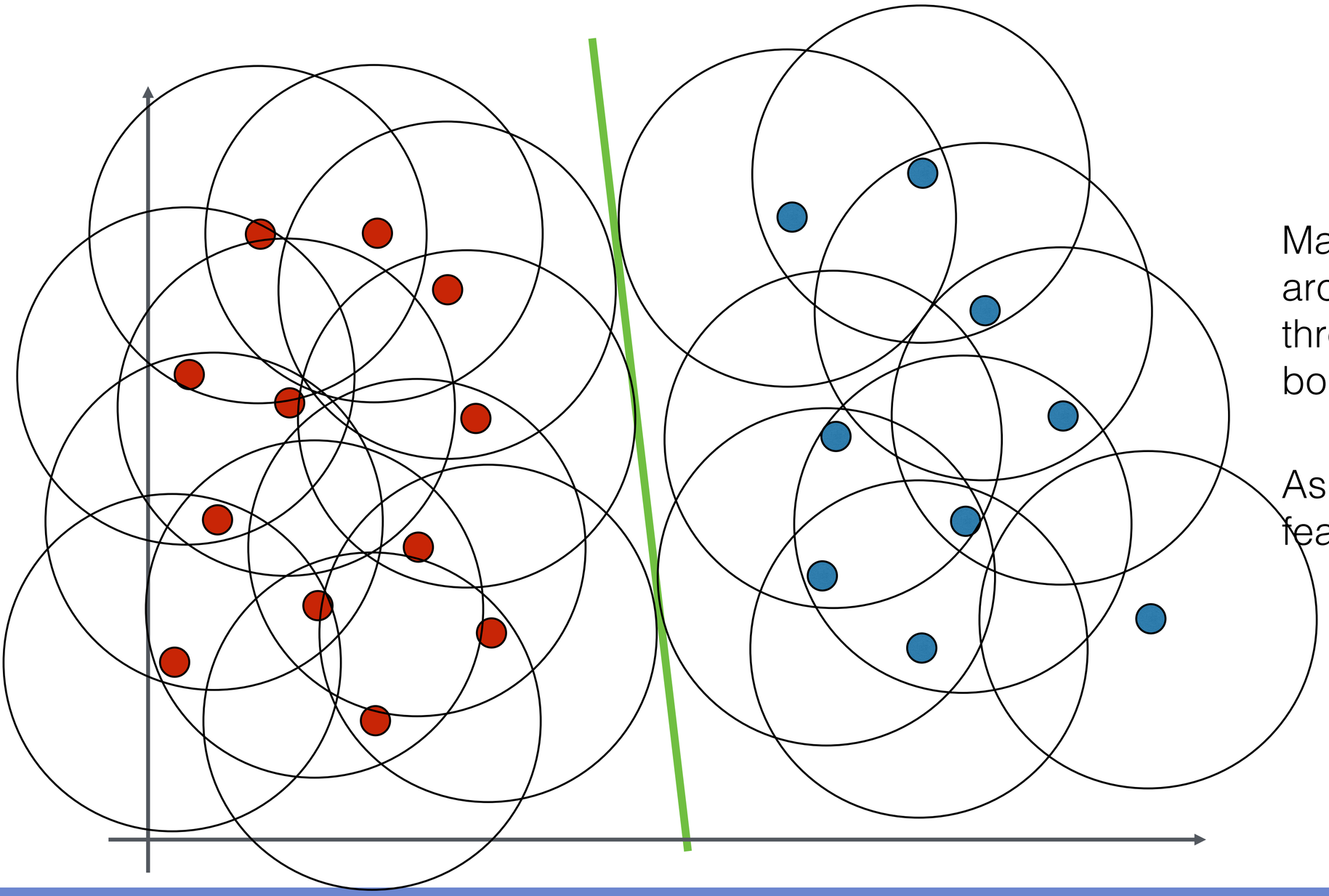
Margin: Bubbles around samples



Margin: Radius of a region around each training sample, through which the decision boundary cannot pass

As the margin increases, the feasible region reduces

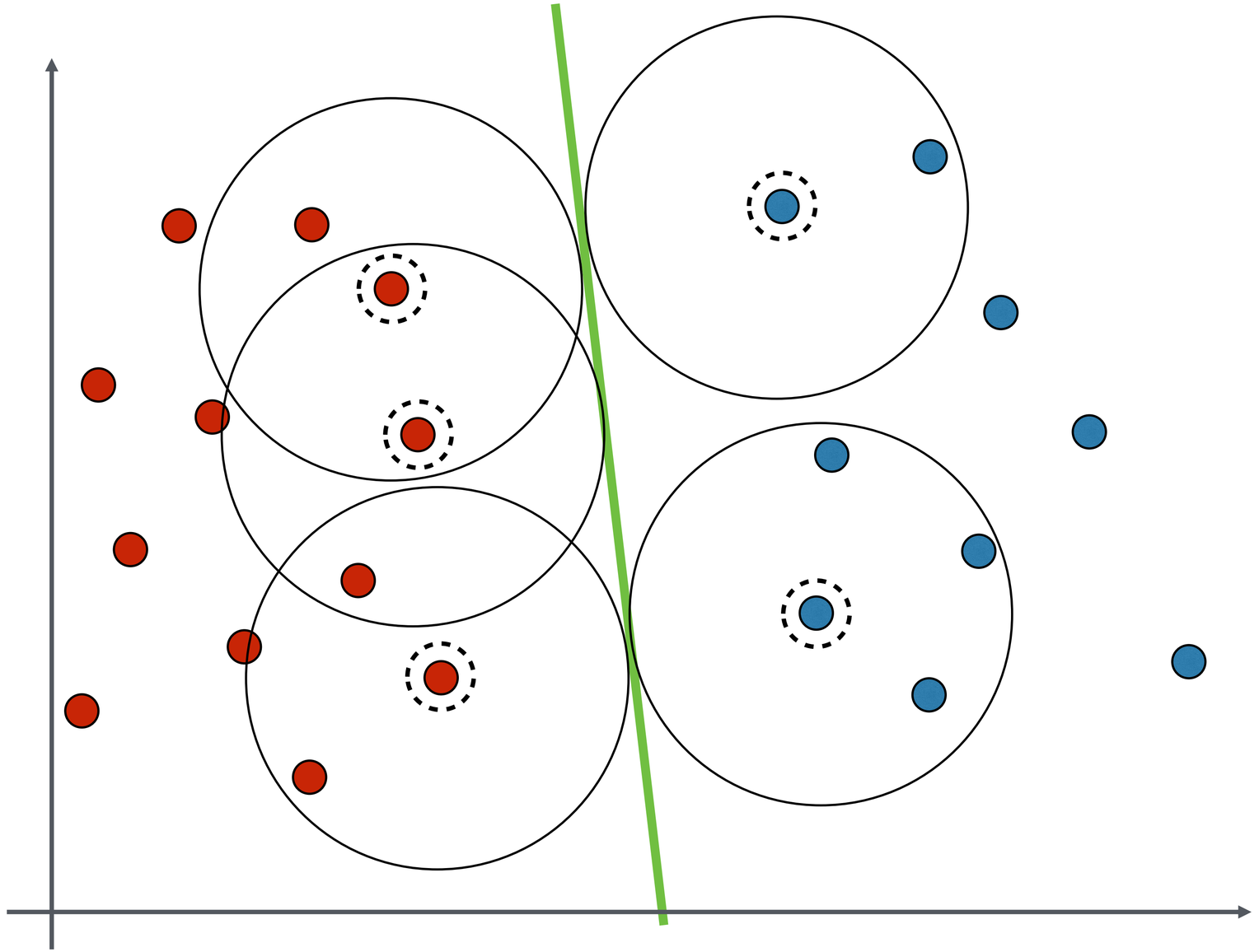
Margin: Bubbles around samples



Margin: Radius of a region around each training sample, through which the decision boundary cannot pass

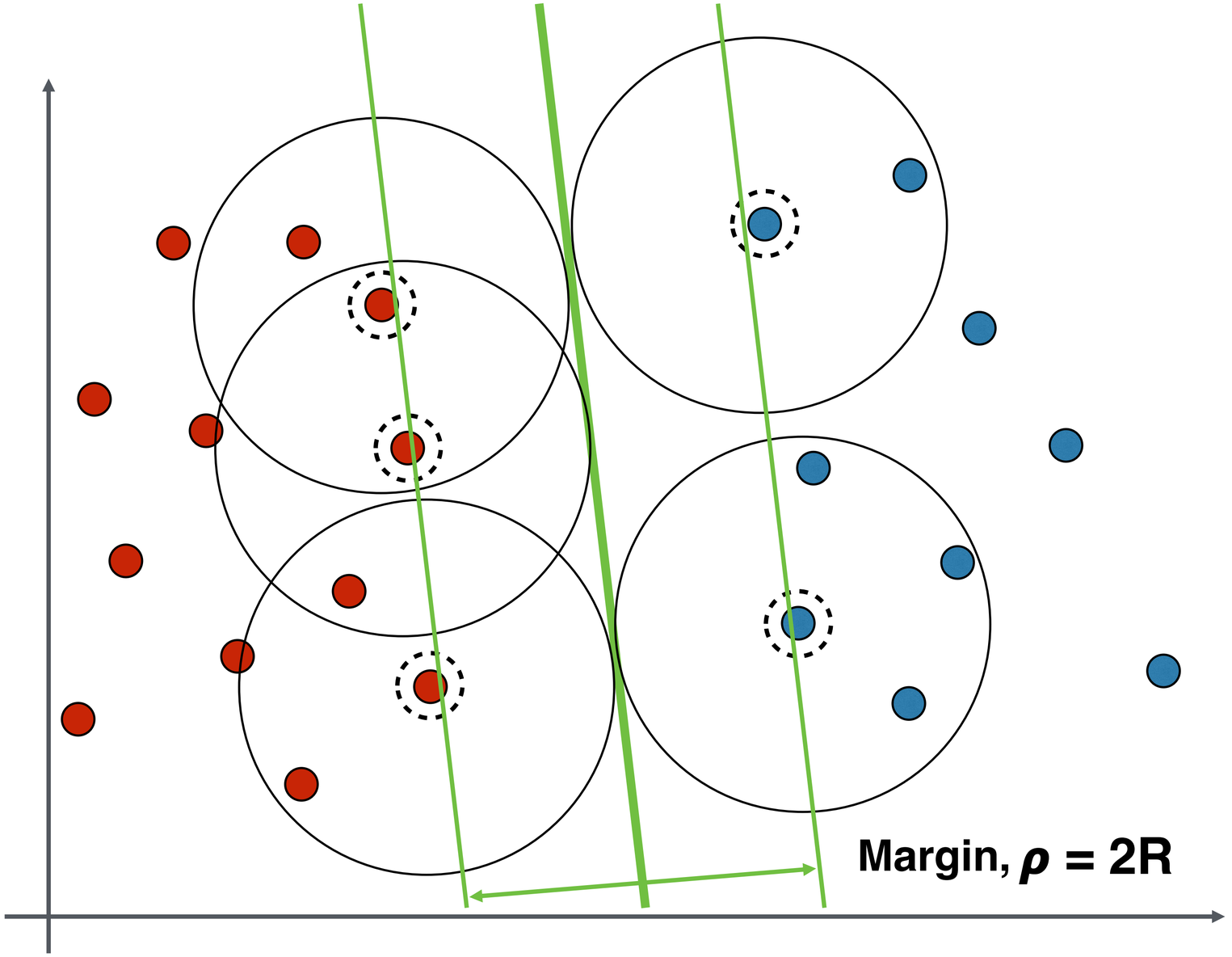
As the margin increases, the feasible region reduces

Margin: Bubbles around samples



A few samples control the Decision Boundary

Band vs. Bubbles



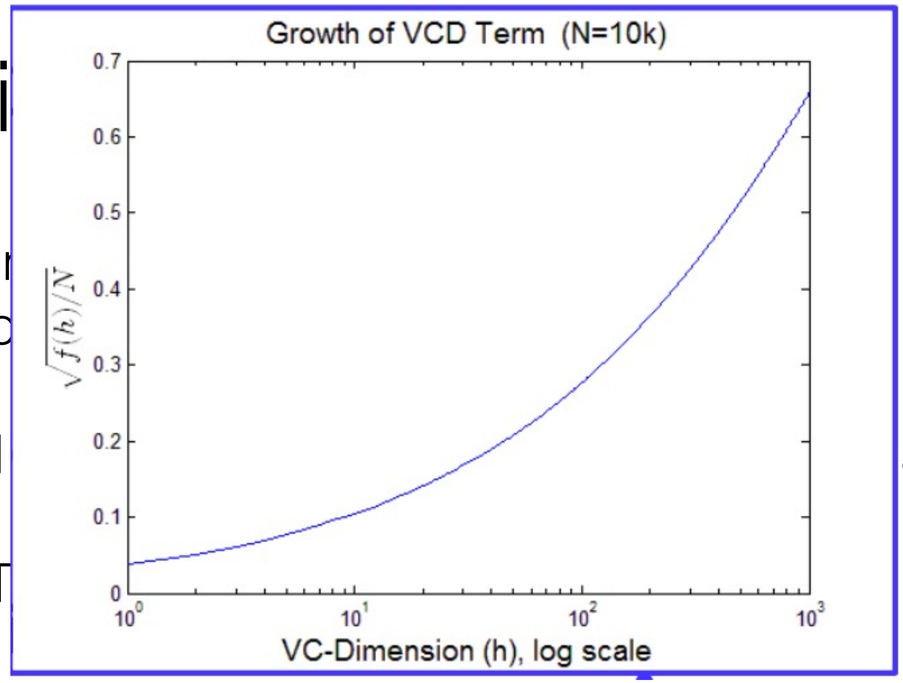
Samples that support the boundary are called ***Support Vectors***

Both interpretations lead yield the same decision boundary

Break though work from Vapni

1. Vapnik, Vladimir N., and A. Ya Chervonenkis. "On the u
frequencies of events to their probabilities." Measures c
Primenen., 1971, Volume 16, Issue 2, Pages 264–279
2. Vapnik, Vladimir N., Estimation of Dependences Based
Moscow.
3. Vapnik, Vladimir N., The Nature of Statistical Learning T

a,



Bound on expected loss:

$$R(\alpha) \leq R_{train}(\alpha) + \sqrt{f(h)/N}$$

h is the VC dimension, and $f(h)$ is given by:

$$f(h) = h + h\log(2N) - h\log(h) - c$$

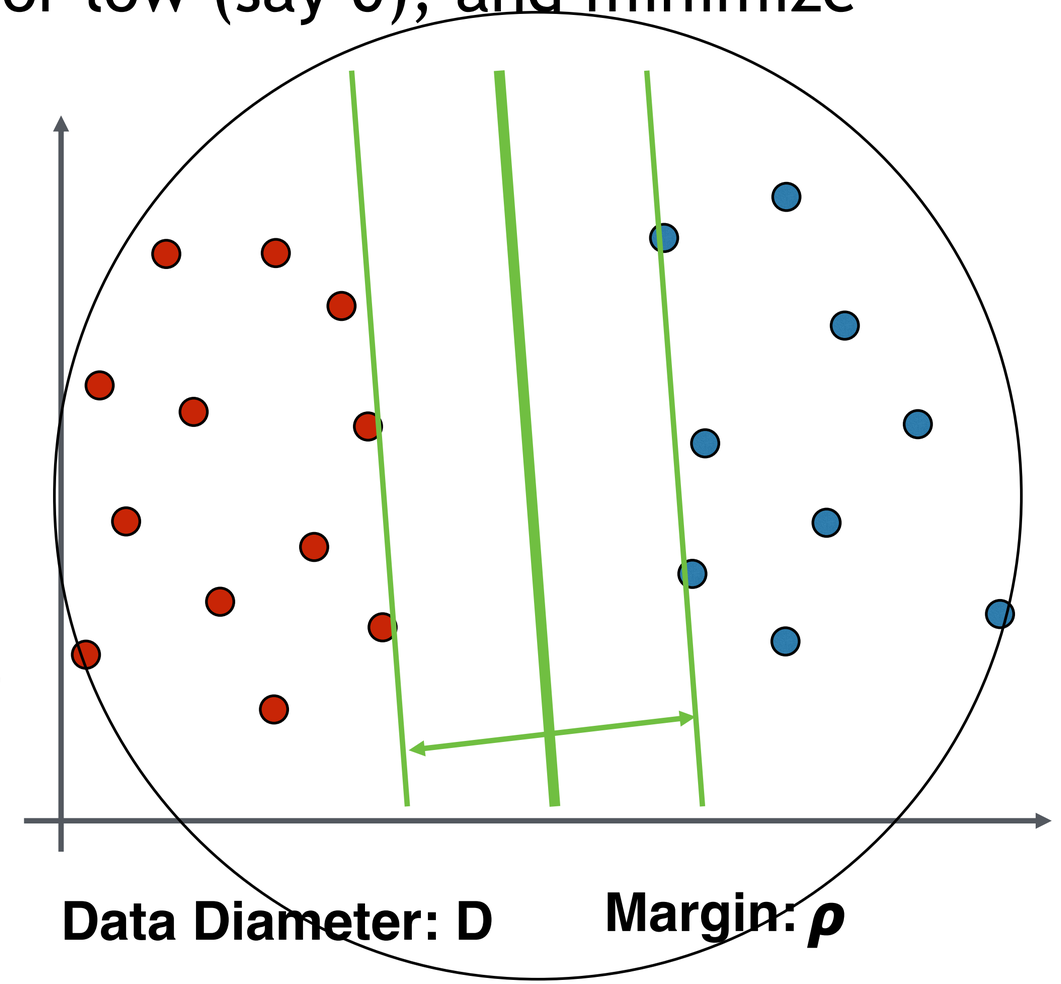
Why maximise the margin?

- To reduce test error, keep training error low (say 0), and minimize the V

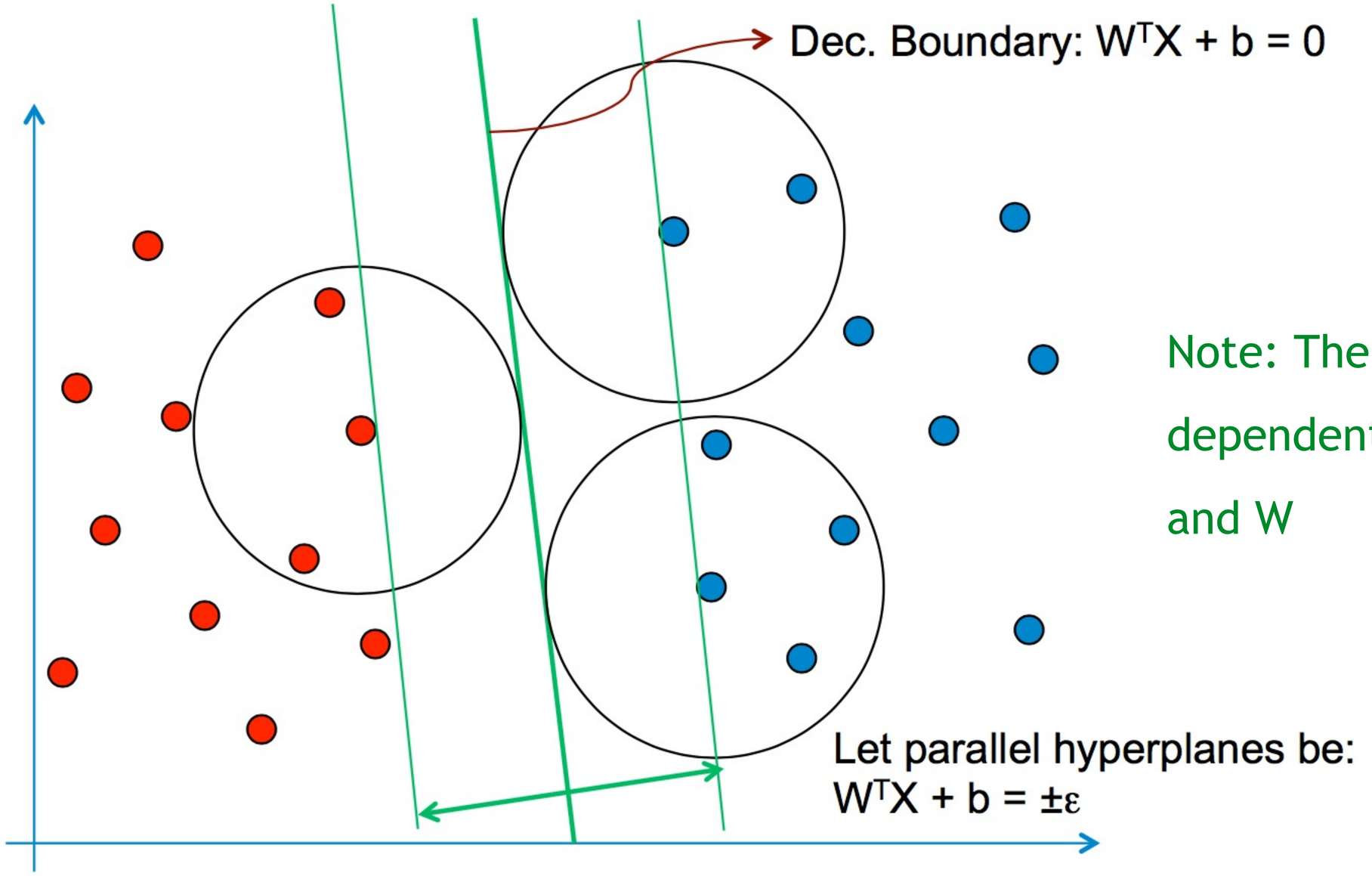
$$\text{Relative Margin}: \ \rho/\text{D}$$

$$\text{VC-D}, \ h \le \min\left\{d \ , \ \left\lceil \frac{D^2}{\rho^2} \right\rceil\right\} + 1$$

- Maximizing margin improves generalization.
- h can be made independent of the dimensionality: d.



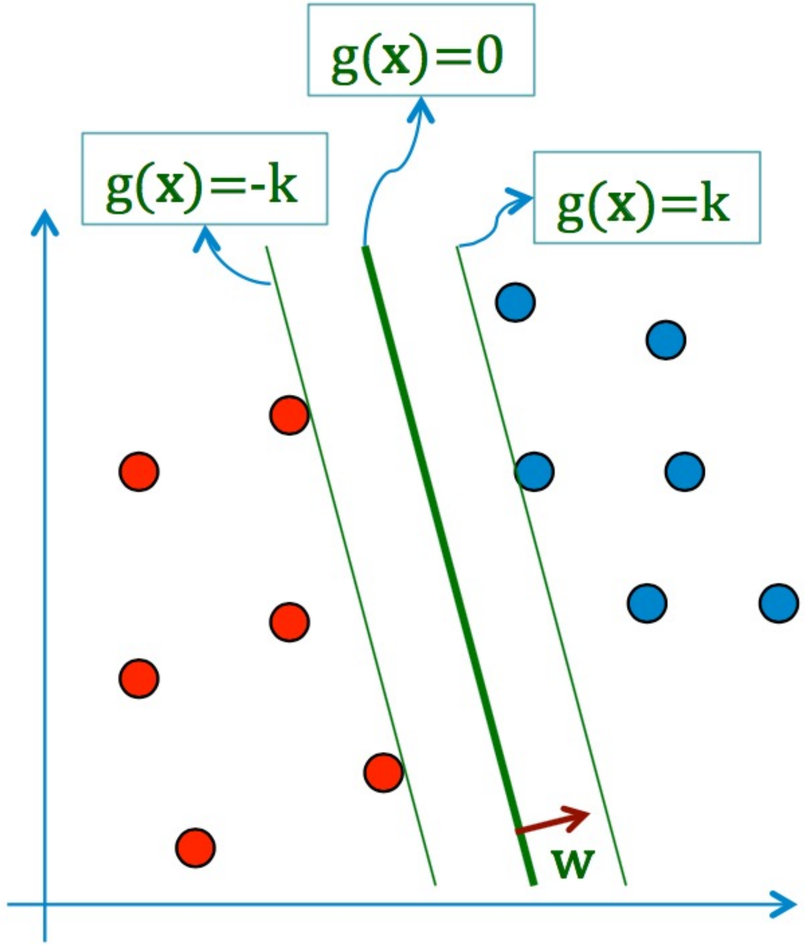
Formalizing the margin



Note: The value of $W^TX_i + b$ is dependent on the scale of X and W

Formulation

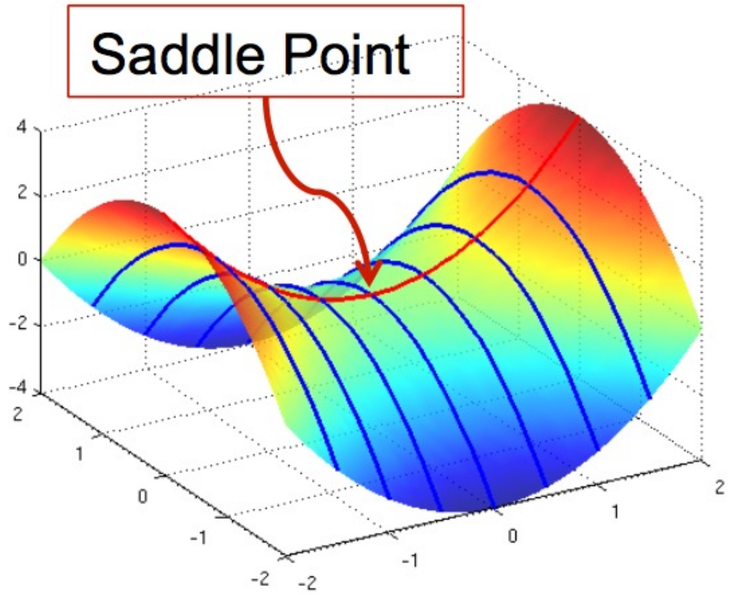
- Let $g(x)=w^Tx+b$.
- We want to maximize k such that:
  - $w^Tx_i+b \geq k$ for $d_i=1$
  - $w^Tx_i+b \leq -k$ for $d_i=-1$
- Value of $g(x)$ dependents on $\|w\|$:
  1. Keep $\|w\|=1$, and maximize $g(x)$, or
  2. Let $g(x)\geq 1$, and minimize $\|w\|$.
- We use approach (2) and formulate the problem as:
  - Minmize: $\frac{1}{2}w^Tw$
  - Subject to: $d_i(w^Tx_i+b) \geq 1$, for $i=1..N$



Optimization

$$\text{Minimize}: \Phi(\mathbf{w}) = \tfrac{1}{2}\mathbf{w}^T\mathbf{w}$$

$$\text{Subject to}: d_i(\mathbf{w}^T\mathbf{x_i} + b) - 1 \geq 0 \quad \forall i$$



Quadratic function: QP solvers

Lagrangian form:

$$\text{Minimize}: J(\mathbf{w}, b, \boldsymbol{\alpha}) = \tfrac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \alpha_i d_i(\mathbf{w}^T\mathbf{x_i} + b) + \sum_{i=1}^{N} \alpha_i$$

$$\text{Subject to}: \alpha_i \geq 0 \quad \forall i$$

Minimize J with respect to w and b, and maximize with respect to α.

Converting to Dual form

$$\text{Objective}: J(\mathbf{w},b,\boldsymbol{\alpha}) = \tfrac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N}\alpha_i d_i(\mathbf{w}^T\mathbf{x_i}+b) + \sum_{i=1}^{N}\alpha_i$$

At the optimum: $1: \dfrac{\partial J}{\partial \mathbf{w}} = 0$ and $2: \dfrac{\partial J}{\partial b} = 0$

$$1: \mathbf{w}_o = \sum_{i=1}^{N}\alpha_i d_i \mathbf{x_i} \qquad 2: \sum_{i=1}^{N}\alpha_i d_i = 0 \qquad 3: \alpha_i[d_i(\mathbf{w}_o{}^T\mathbf{x_i}+b_o)-1] = 0$$

$$\text{Obj}: J(\mathbf{w},b,\boldsymbol{\alpha}) = \sum_{i=1}^{N}\alpha_i + \tfrac{1}{2}\mathbf{w}^T\mathbf{w} - \mathbf{w}^T\sum_{i=1}^{N}\alpha_i d_i \mathbf{x_i} - b\sum_{i=1}^{N}\alpha_i d_i$$

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N}\alpha_i - \tfrac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j d_i d_j \mathbf{x_i}{}^T\mathbf{x}_j$$

Solving the Dual form

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \tfrac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \mathbf{x_i}^T \mathbf{x}_j$$

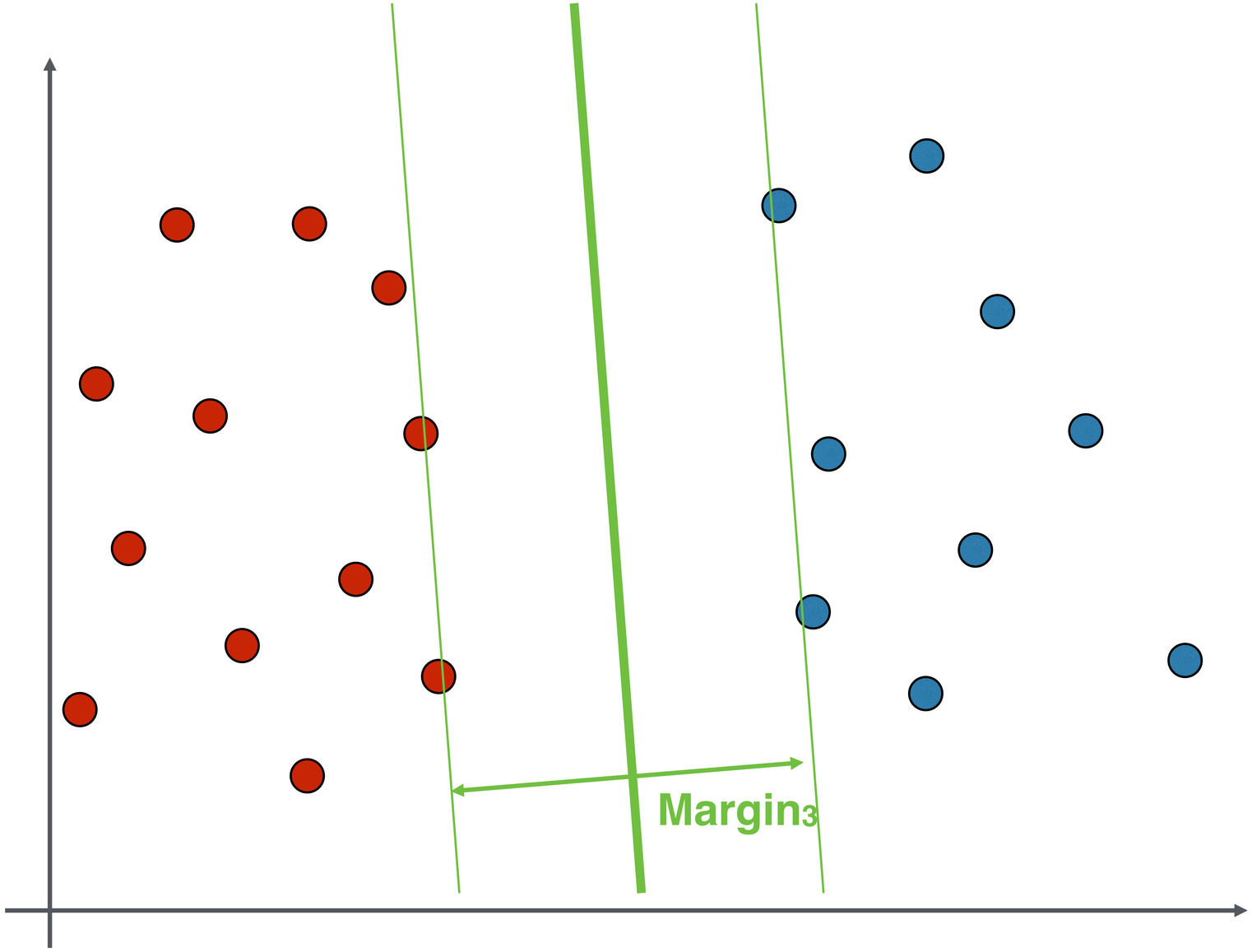
$$\text{Subject to } \alpha_i \geq 0 \quad \forall_i \quad \text{and} \quad \sum_{i=1}^{N} \alpha_i d_i = 0$$

QP Solver

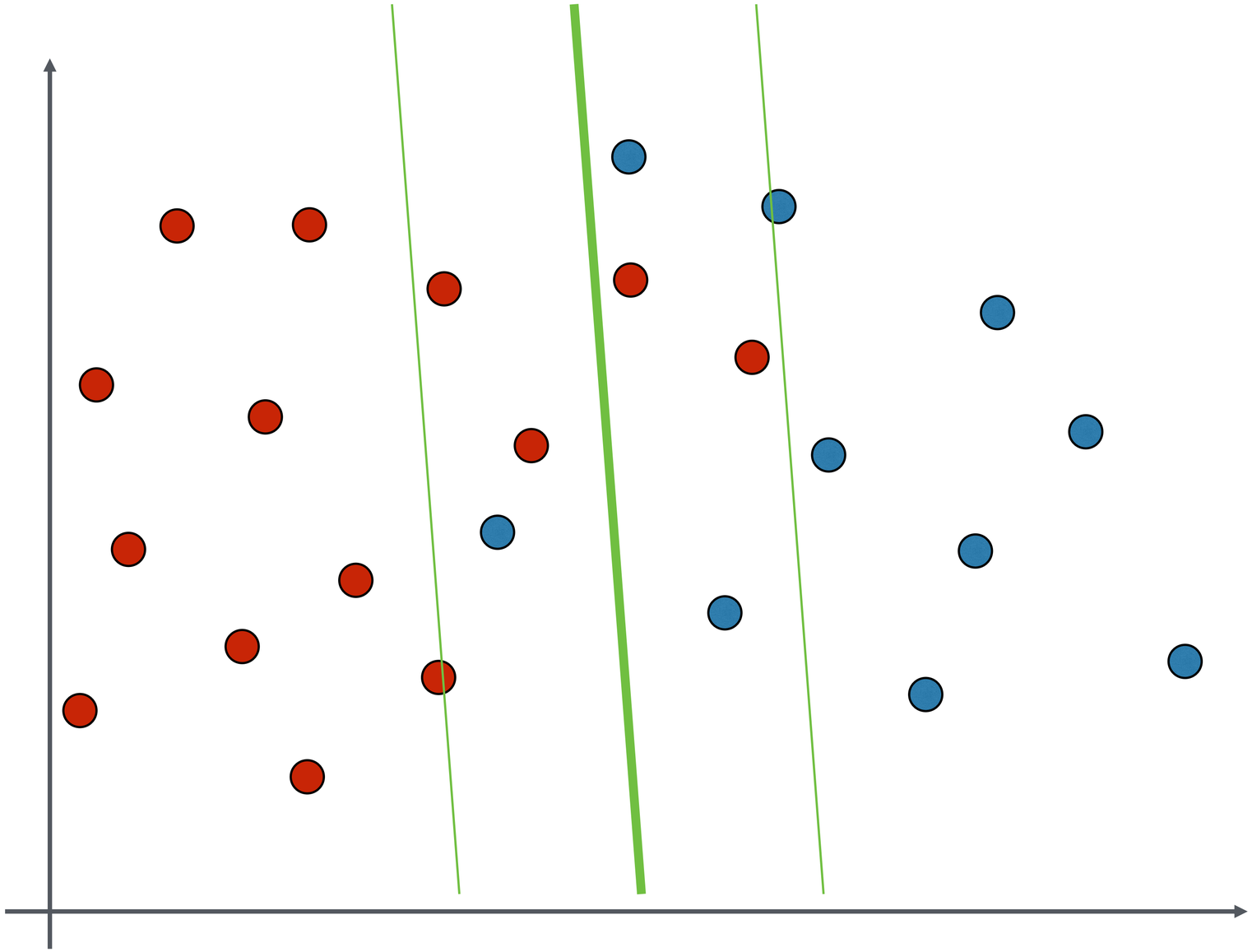
α_i

$$\mathbf{w}_o = \sum_{i=1}^{N} \alpha_i d_i \mathbf{x_i} \rightarrow \alpha_i [d_i (\mathbf{w}_o^T \mathbf{x_i} + b_o) - 1] = 0 \rightarrow b_o = 1 - \mathbf{w}_o^T \mathbf{x}_{s+}$$

Non Separable Data

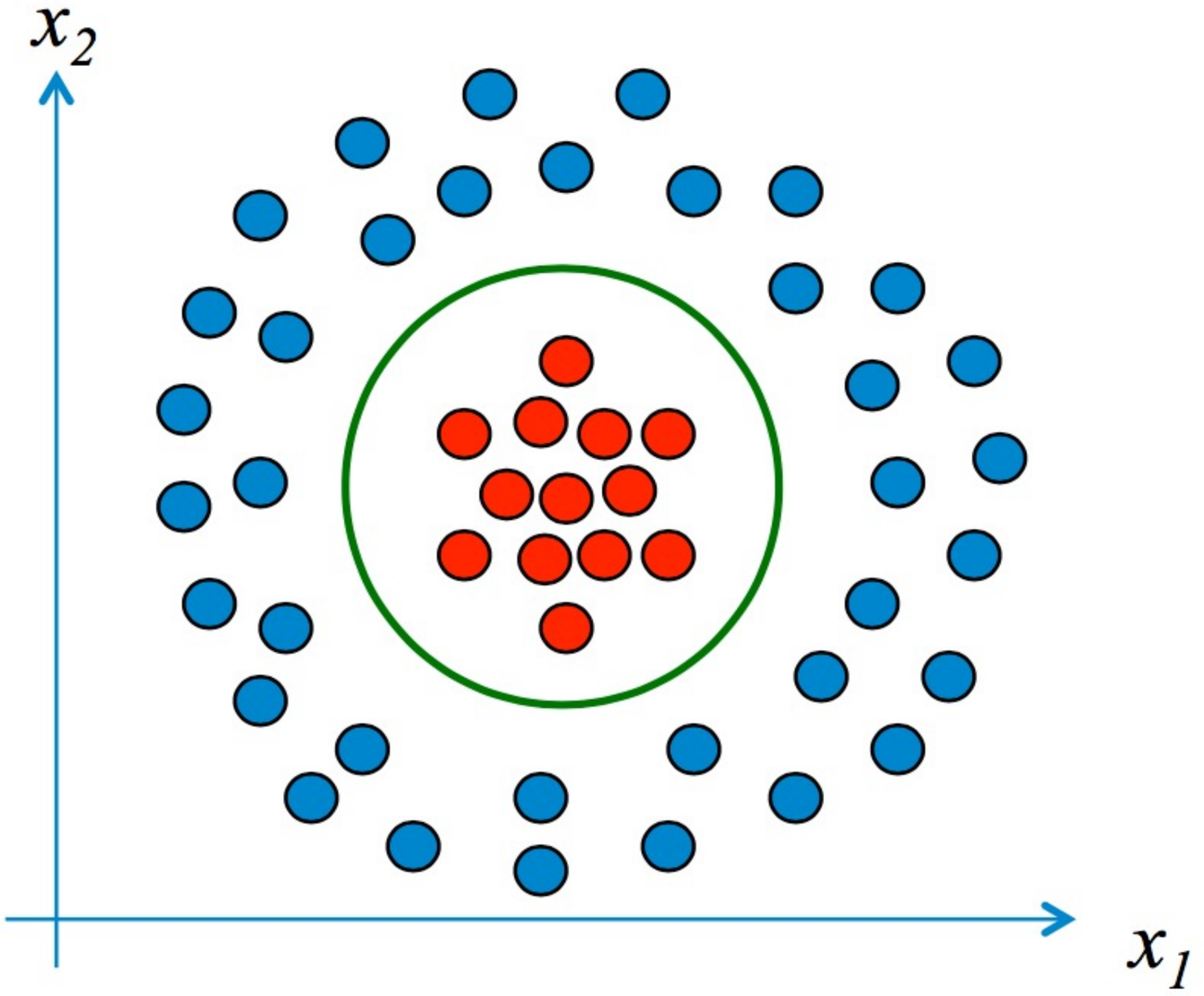


Non Separable Data



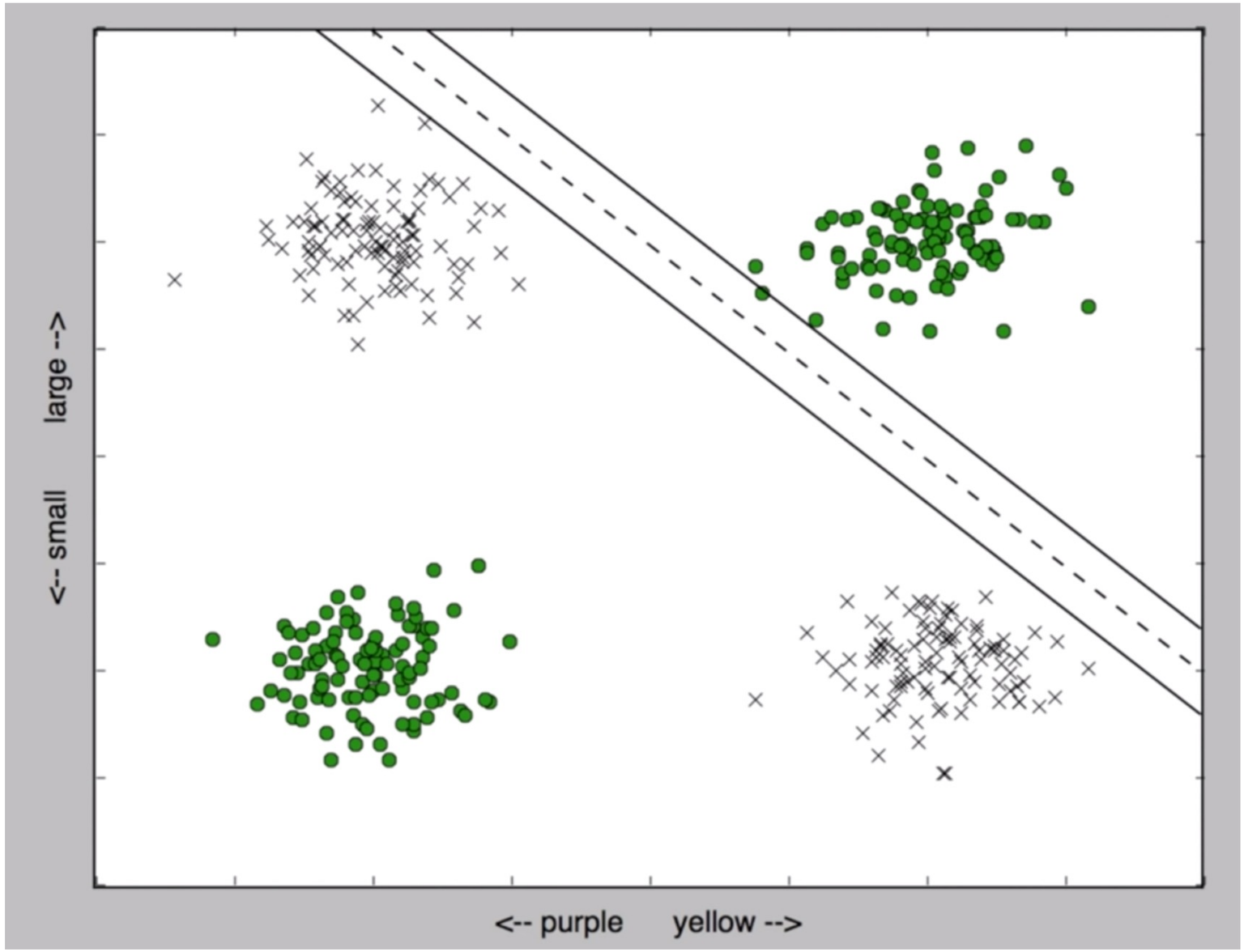
1. Noisy Data/ Bad Features

Non Separable Data



2. Non Linear Boundary

Non Separable Data



SVM with Noisy data

$$\text{Minimize}: \Phi(\mathbf{w}) = \tfrac{1}{2}\mathbf{w}^T\mathbf{w}$$

$$\text{Subject to}: d_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1 \quad \forall i$$

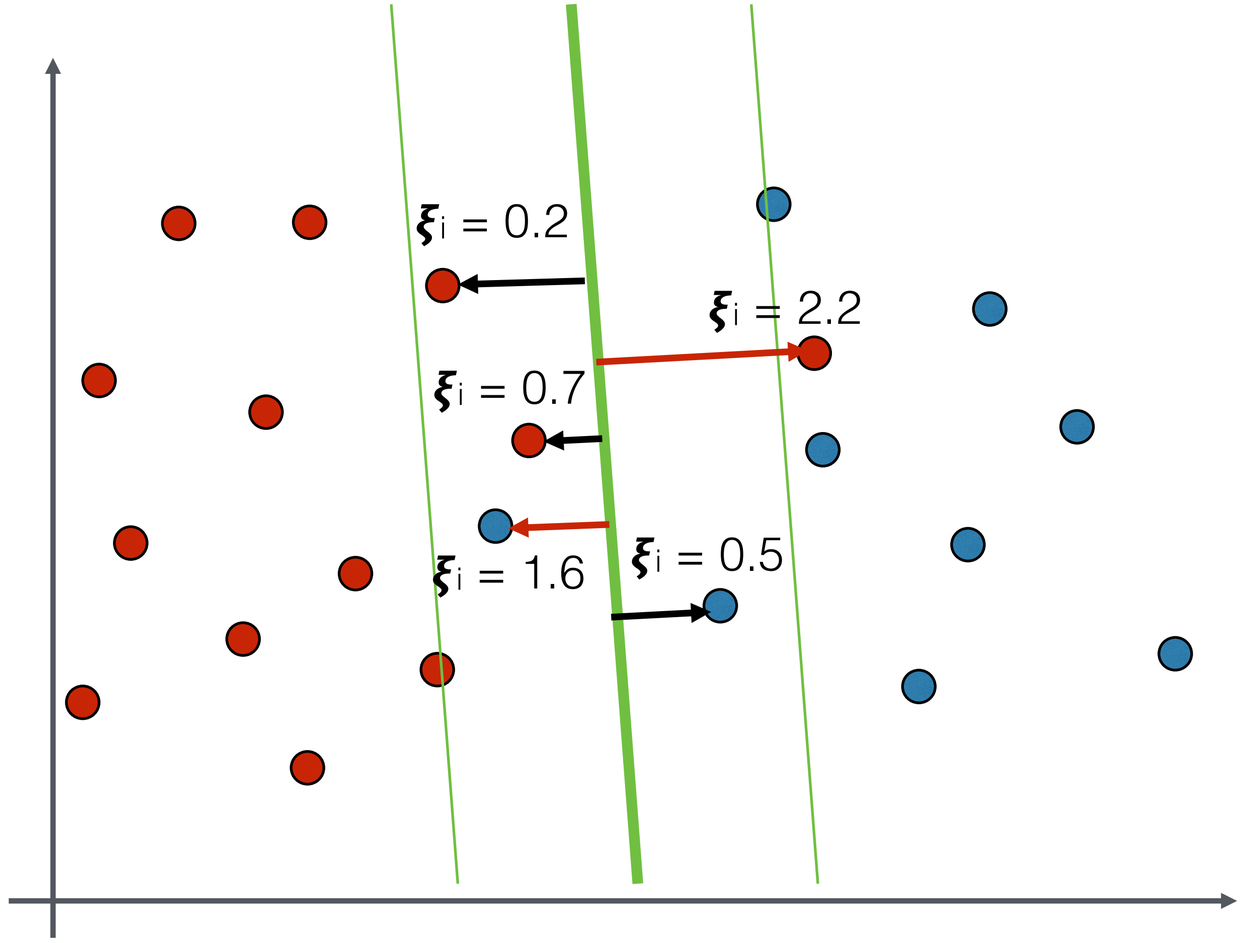
Introduce slack variables $\xi_i \geq 0$

$$\text{Minimize}: \Phi(\mathbf{w}) = \tfrac{1}{2}\mathbf{w}^T\mathbf{w}$$

$$\text{Subject to}: d_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1 - \xi_i \quad \forall i$$

$$\text{Minimize}: \Phi(\mathbf{w}) = \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i$$

$$\text{Subject to}: d_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1 - \xi_i; \quad \xi_i \geq 0, \quad \forall i$$



Dual Form with Slack variables

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \tfrac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \mathbf{x_i}^T \mathbf{x}_j$$

$$\text{Subject to } \alpha_i \geq 0 \quad \forall_i \quad \text{and} \quad \sum_{i=1}^{N} \alpha_i d_i = 0$$

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \tfrac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \mathbf{x_i}^T \mathbf{x}_j$$

$$\text{Subject to } 0 \leq \alpha_i \leq C \quad \forall_i \quad \text{and} \quad \sum_{i=1}^{N} \alpha_i d_i = 0$$

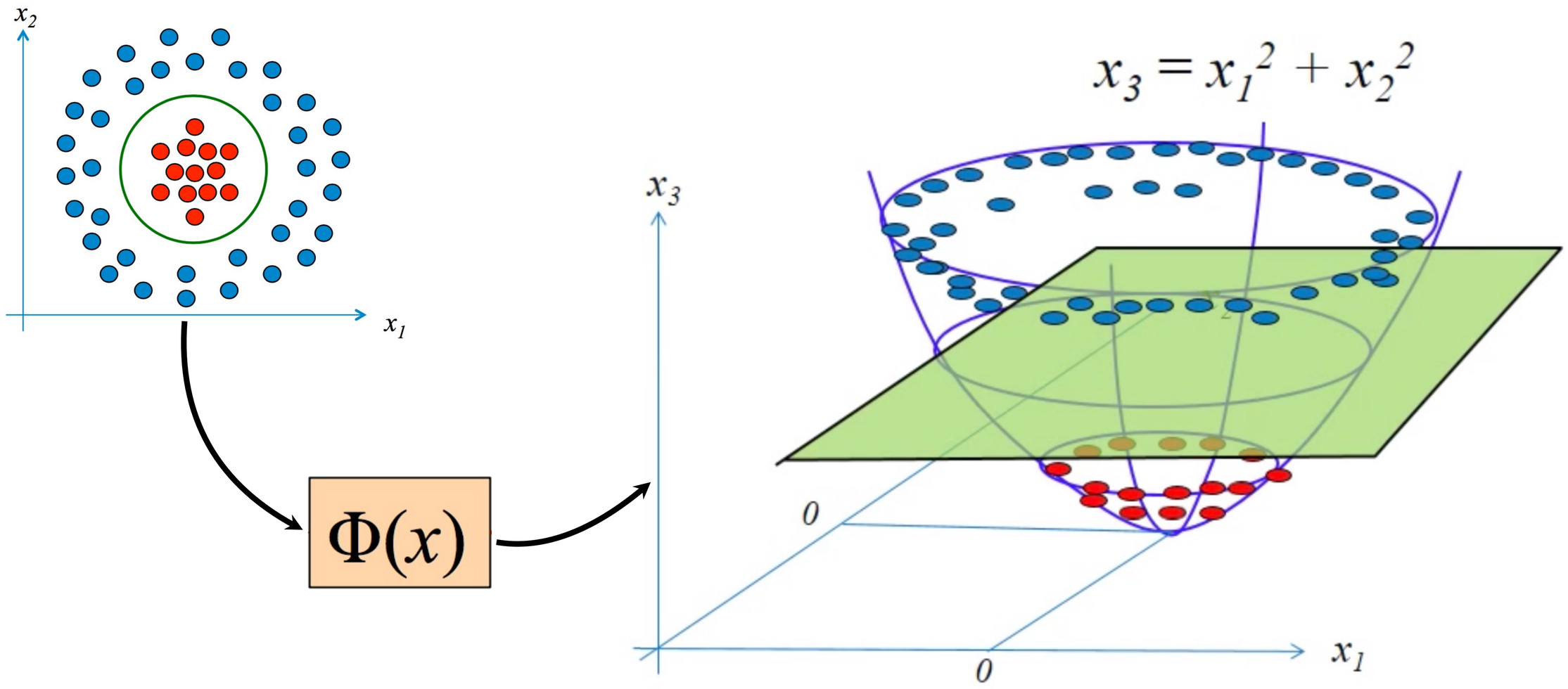
Dual Form with Slack variables

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \mathbf{x_i}^T \mathbf{x}_j$$

$$\text{Subject to } 0 \le \alpha_i \le C \quad \forall_i \quad \text{and} \quad \sum_{i=1}^{N} \alpha_i d_i = 0$$

- Note that neither the slack variables, nor their Lagrange multipliers appear in the dual
- The only change is the additional constraint on α_i
- The parameter C controls the relative weight between training error and the VC dimension

Non Linear Boundaries



Non linear mapping from into a higher dimensional space

SVM post mapping

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \, \Phi(\mathbf{x}_i) \bullet \Phi(\mathbf{x}_j)$$

$$\text{Subject to} \quad 0 \le \alpha_i \le C \quad \forall_i \quad \text{and} \quad \sum_{i=1}^{N} \alpha_i d_i = 0$$

Training

$$\text{Label} = sign\left(\mathbf{w}_o \bullet \Phi(\mathbf{x}_{test}) + b_o\right)$$

$$\mathbf{w}_o = \sum_{i=1}^{N} \alpha_i d_i \, \Phi(\mathbf{x}_i)$$

$$\therefore \text{Label} = sign\left(\sum_{i=1}^{N} \left(\alpha_i d_i \, \Phi(\mathbf{x}_i) \bullet \Phi(\mathbf{x}_{test})\right) + b_o\right)$$

Testing

SVM post mapping

- Data vectors occur only as dot products in SVM-learning and testing
- If we can find a function K(X,Y), which is equivalent to Φ(X)·Φ(Y), we can avoid explicit mapping to high dimensions.

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \, \mathrm{K}(\mathbf{x_i}, \mathbf{x}_j)$$

$$\text{Label} = sign\left(\sum_{i=1}^{N} \left(\alpha_i d_i \, \mathrm{K}(\mathbf{x_i}, \mathbf{x_{test}}) \right) + b_o \right)$$

What do we gain by using K?

$$\text{Let } \Phi(\mathbf{X}) = \Phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_2x_1 \\ x_2x_2 \end{bmatrix}$$

We can compute $K(X,Y)=(X.Y)^2$ instead of mapping with Φ explicitly and then computing dot product.

$$\text{Let } K(\mathbf{X},\mathbf{Y}) = \Phi(\mathbf{X}) \bullet \Phi(\mathbf{Y}) = \begin{bmatrix} x_1^2 \\ x_1x_2 \\ x_2x_1 \\ x_2^2 \end{bmatrix} \bullet \begin{bmatrix} y_1^2 \\ y_1y_2 \\ y_2y_1 \\ y_2^2 \end{bmatrix}$$

$$= x_1^2y_1^2 + 2x_1x_2y_1y_2 + x_2^2y_2^2 = (x_1y_1 + x_2y_2)^2 = (\mathbf{X} \bullet \mathbf{Y})^2$$

What do we gain by using K?

- Original Space: 2-dimensional

Let $K(\mathbf{X},\mathbf{Y}) = (\mathbf{X}\bullet\mathbf{Y})^3 = (x_1y_1 + x_2y_2)^3$

$$\Phi(\mathbf{X}) = \Phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_1^3 \\ x_1^2x_2 \\ x_1x_2^2 \\ x_2^3 \end{bmatrix}$$

What do we gain by using K?

- Original Space: 3-dimensional

$$\text{Let } K(\mathbf{X},\mathbf{Y}) = (\mathbf{X}\bullet\mathbf{Y})^3 = (x_1y_1 + x_2y_2 + x_3y_3)^3$$

$$\Phi(\mathbf{X}) = \Phi\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right) = \begin{bmatrix} x_1^3 \\ x_2^3 \\ x_3^3 \\ x_1^2x_2 \\ x_1x_2^2 \\ x_1^2x_3 \\ x_1x_3^2 \\ x_2^2x_3 \\ x_2x_3^2 \\ x_1x_2x_3 \end{bmatrix}$$

A Generic Polynomial Kernel

- Adding two Kernels gives you a new Kernel:

$$K(\mathbf{X},\mathbf{Y}) = K_1(\mathbf{X},\mathbf{Y}) + K_2(\mathbf{X},\mathbf{Y}) \quad : \quad \Phi(\mathbf{X}) = \begin{bmatrix} \Phi_1(\mathbf{X}) \\ \Phi_2(\mathbf{X}) \end{bmatrix}$$

$$K(\mathbf{X},\mathbf{Y}) = \Phi(\mathbf{X}) \bullet \Phi(\mathbf{Y}) = \Phi_1(\mathbf{X}) \bullet \Phi_1(\mathbf{Y}) + \Phi_2(\mathbf{X}) \bullet \Phi_2(\mathbf{Y})$$

$$K_p(\mathbf{X},\mathbf{Y}) = (1 + \mathbf{X} \bullet \mathbf{Y})^p = 1 + \mathbf{X} \bullet \mathbf{Y} + (\mathbf{X} \bullet \mathbf{Y})^2 + \cdots + (\mathbf{X} \bullet \mathbf{Y})^p$$

Just adding a 1 and raising to the power of p maps the input vector into a space containing all original dimensions, all 2-products, 3-products,..,p-products.

Mercer's Theorem

- Using Kernels, we avoid explicit mapping with Φ
- In fact, we do not even have to know what Φ is as long as we are sure there exists a valid Φ
- Mercer's Theorem:

Any given kernel can be expanded as a series :

$$K(\mathbf{X},\mathbf{Y}) = \sum_{i=1}^{\infty} \lambda_i \, \Phi_i(\mathbf{X}) \bullet \Phi_i(\mathbf{Y}), \quad \lambda_i > 0; \; iff$$

$K(\mathbf{X},\mathbf{Y})$ satisfies the Mercer's conditions

(symmetric, continuous, positive semi - definite)

Popular Kernels

- Polynomial:

$$K_p(\mathbf{X}, \mathbf{Y}) = (1 + \mathbf{X} \bullet \mathbf{Y})^p$$

- Radial Basis Function (RBF) or Gaussian:

$$K_r(\mathbf{X}, \mathbf{Y}) = e^{-\frac{1}{2\sigma^2}\|\mathbf{X}-\mathbf{Y}\|_2^2}$$

- Hyperbolic Tangent:

$$K_s(\mathbf{X}, \mathbf{Y}) = \tanh(\beta_0 \mathbf{X} \bullet \mathbf{Y} + \beta_1)$$