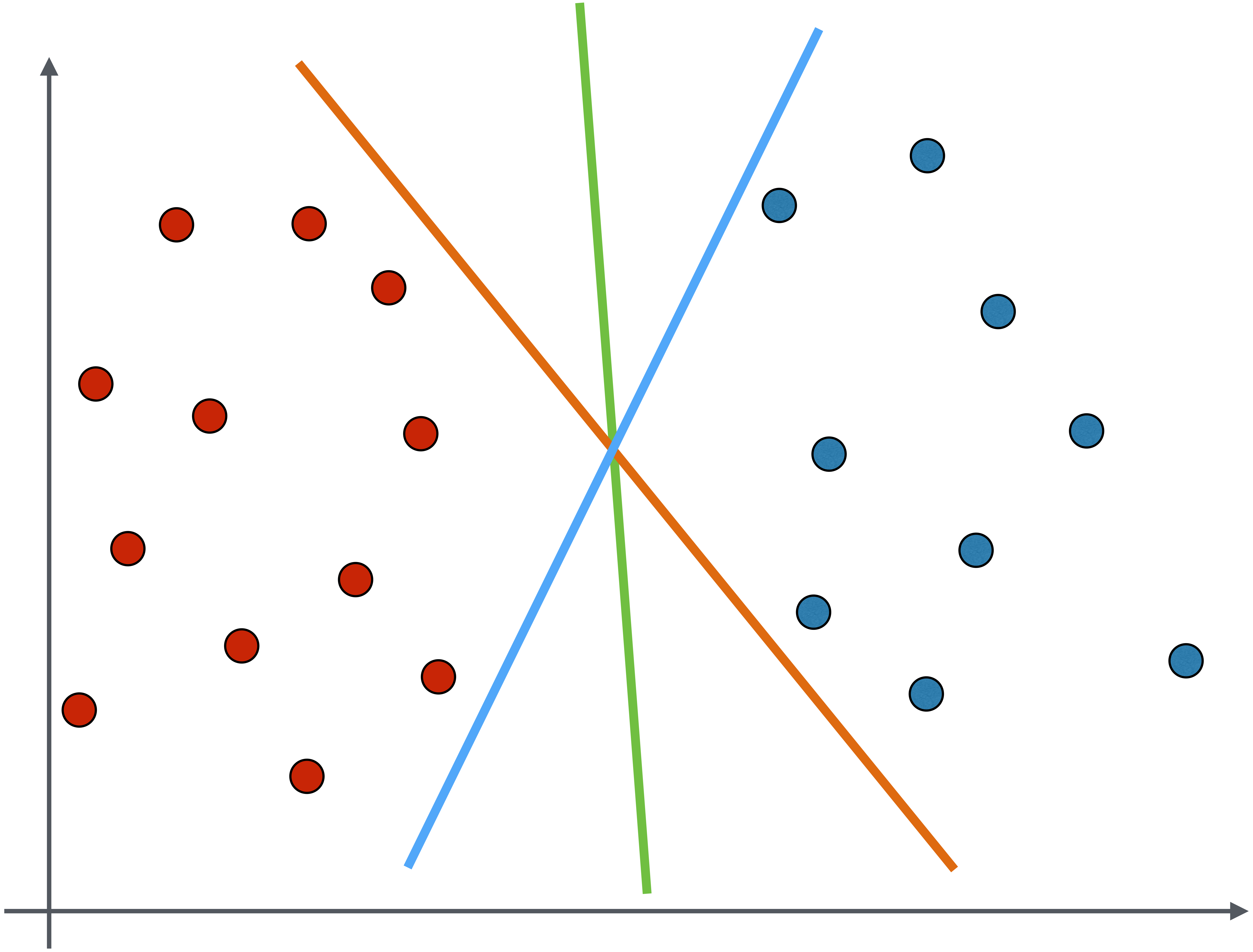
Statistical Methods in AI (CSE 471)
Lecture 6: Support Vector Machines

Vineet Gandhi
Centre for Visual Information Technology (CVIT)

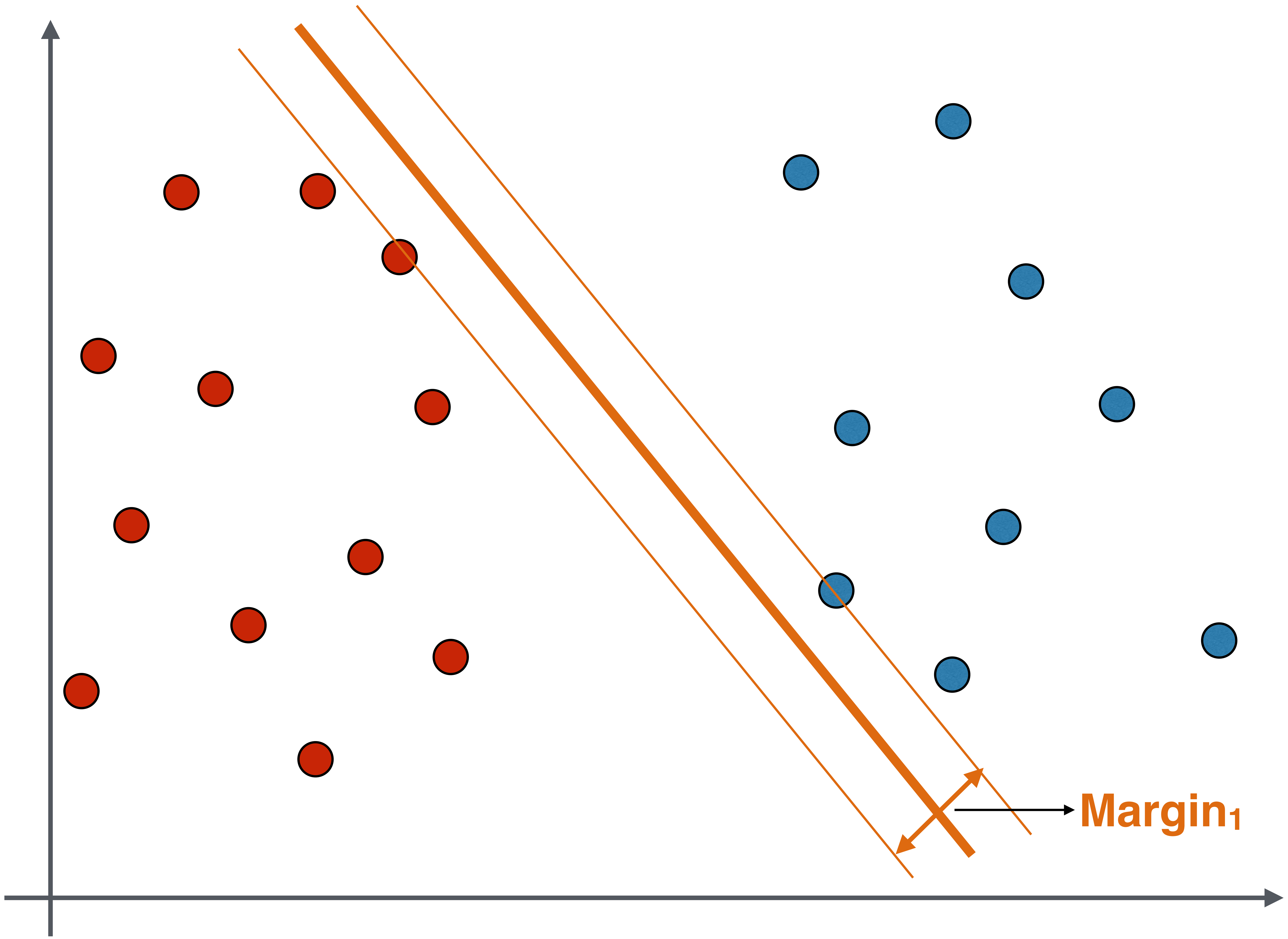
INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Multiple solutions exist for linearly separable data



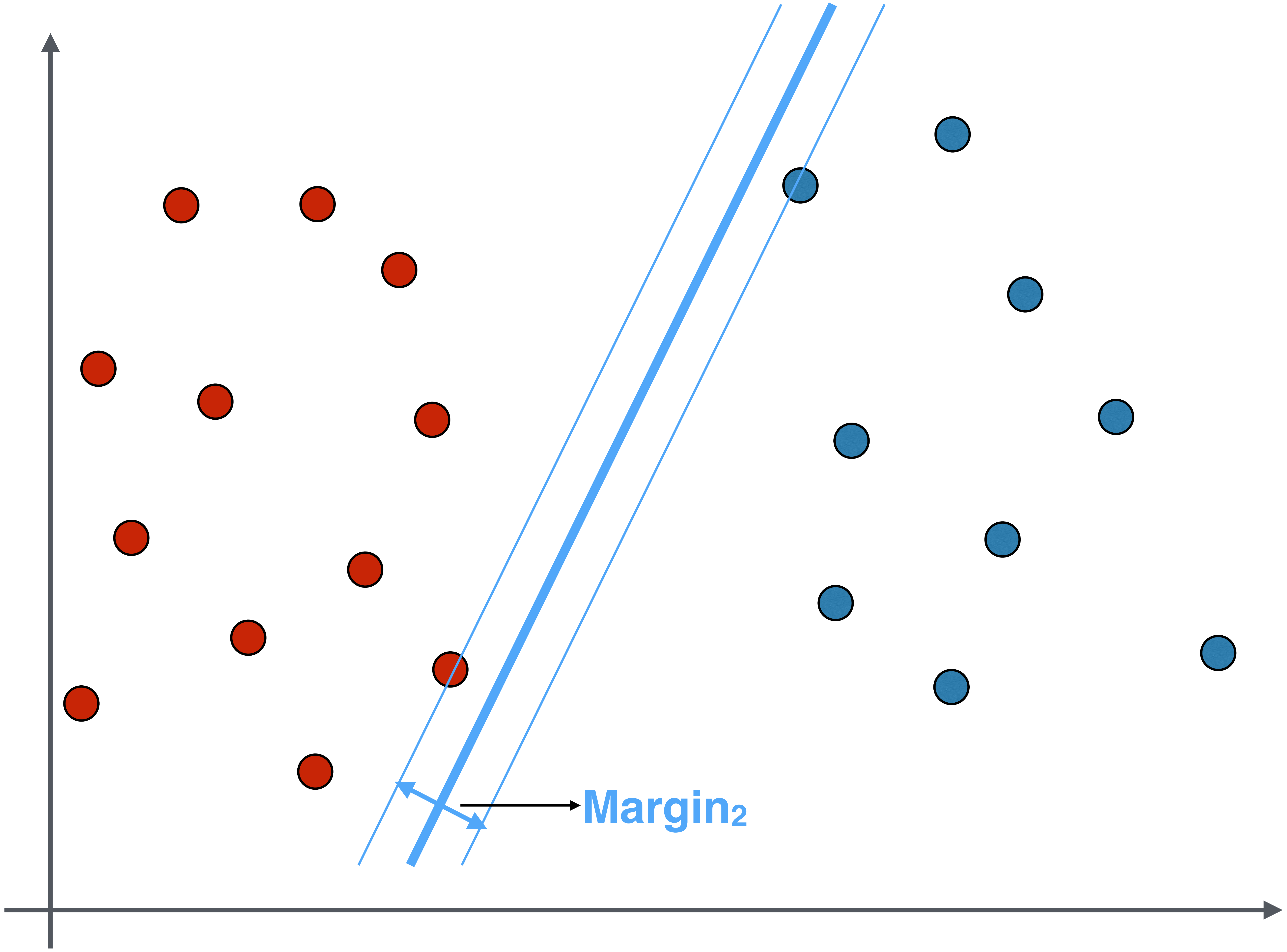
Are all solutions equally good?

Margin: No-mans Band



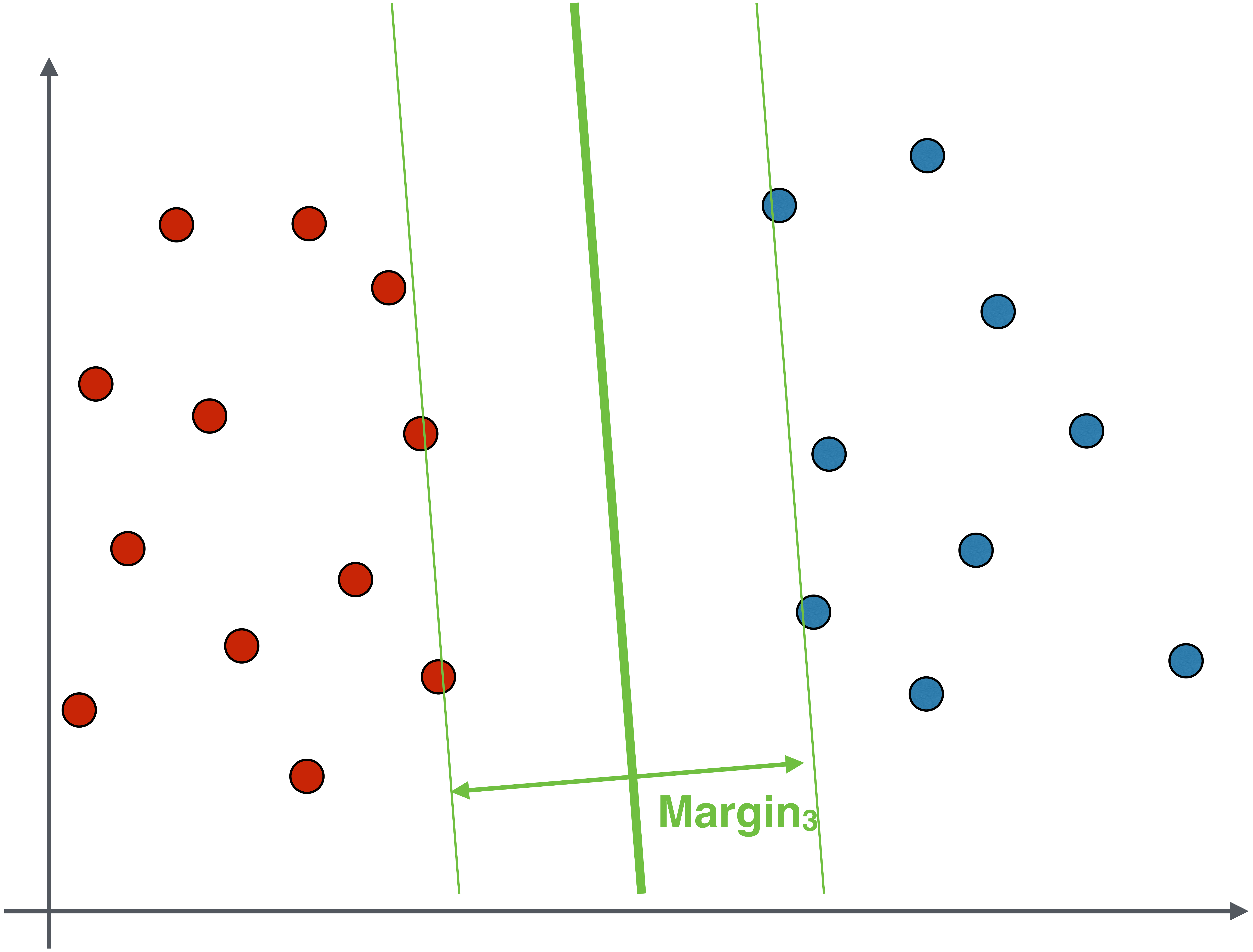
Margin: Width of a band around decision boundary without any training samples

Multiple solutions exist for linearly separable data



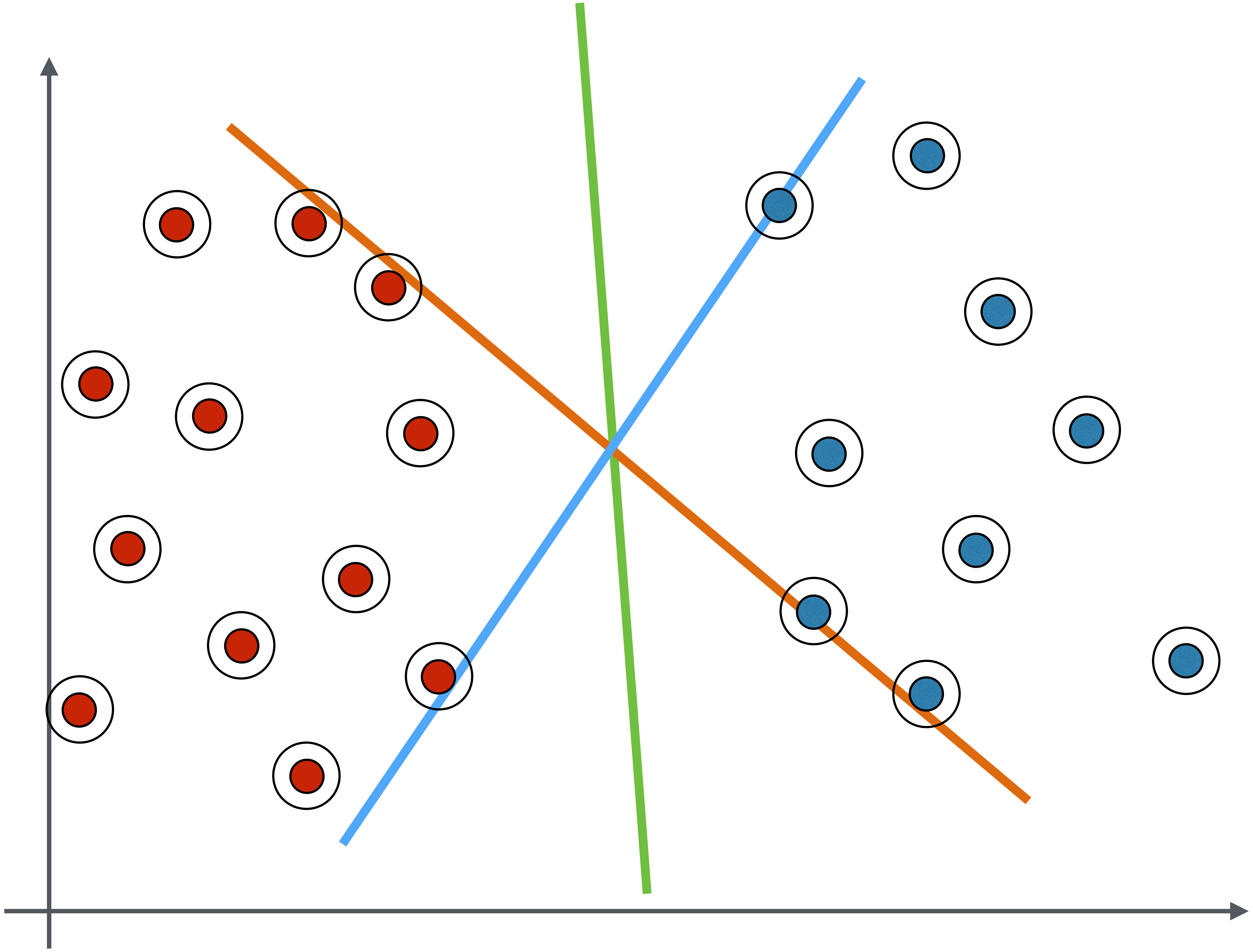
Margin: Width of a band around decision boundary without any training samples

Multiple solutions exist for linearly separable data



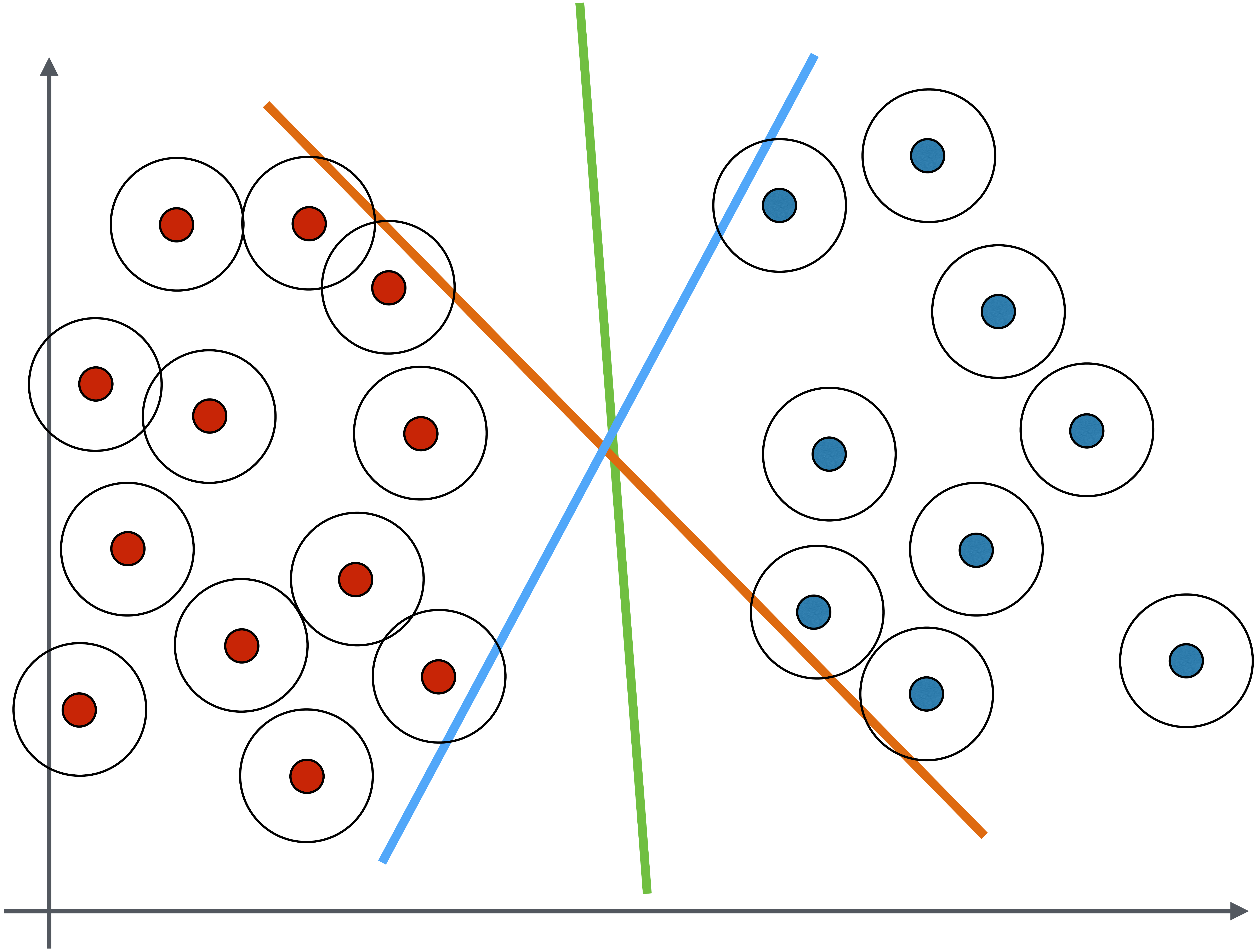
Is a Larger Margin better? Why?

Margin: Bubbles around samples



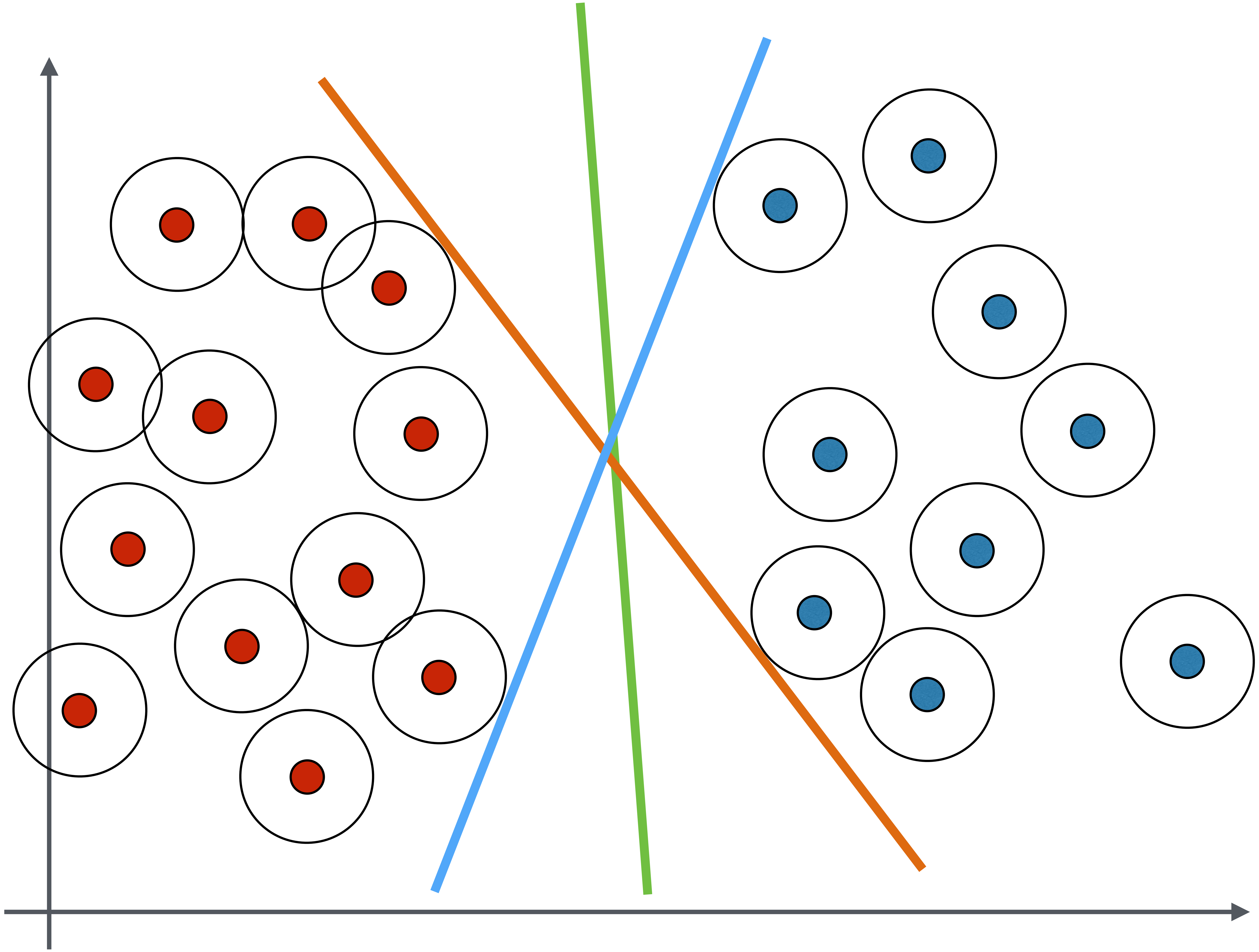
Margin: Radius of a region around each training sample, through which the decision boundary cannot pass

Margin: Bubbles around samples



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

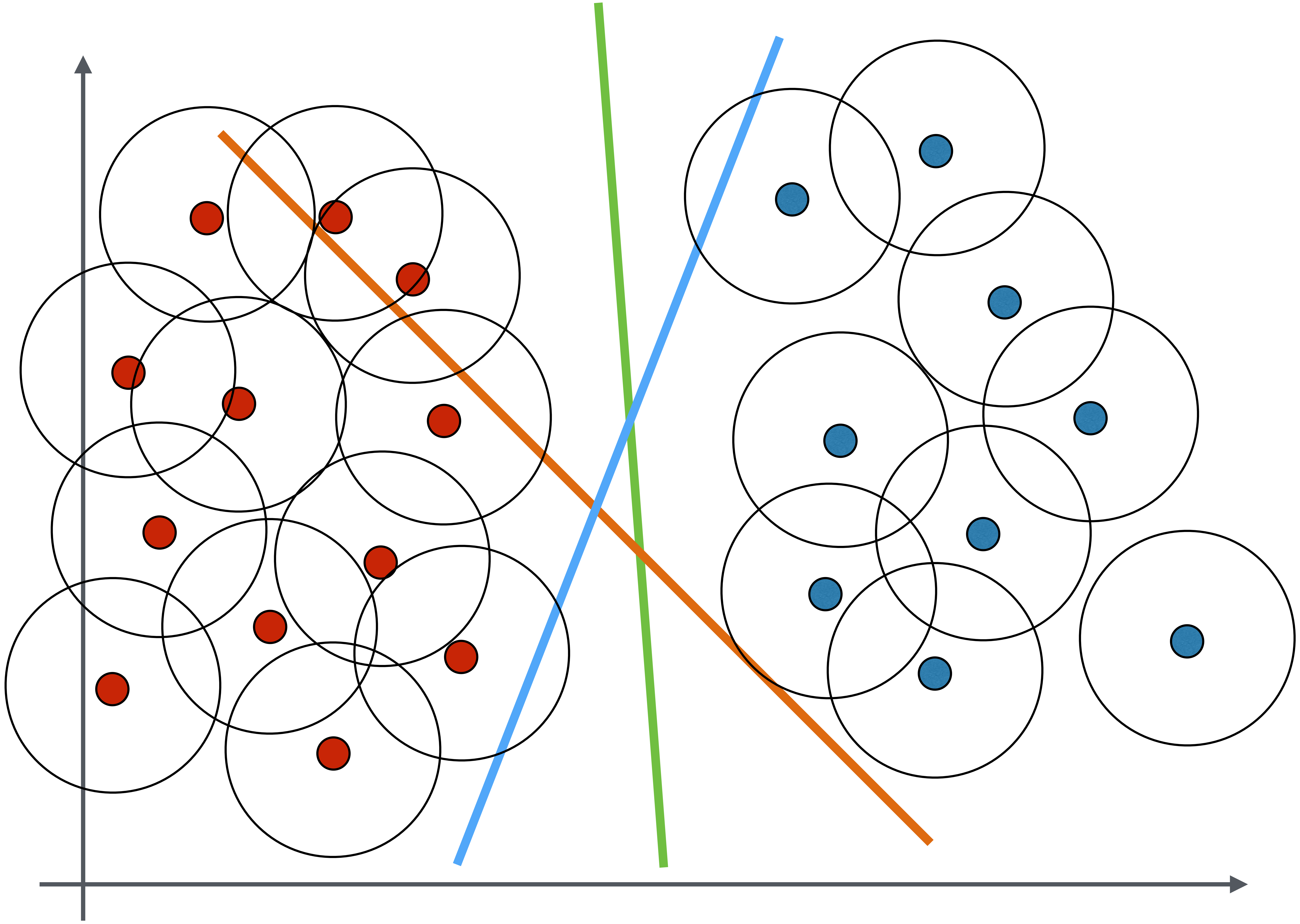
Margin: Bubbles around samples



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

As the margin increases, the feasible region reduces

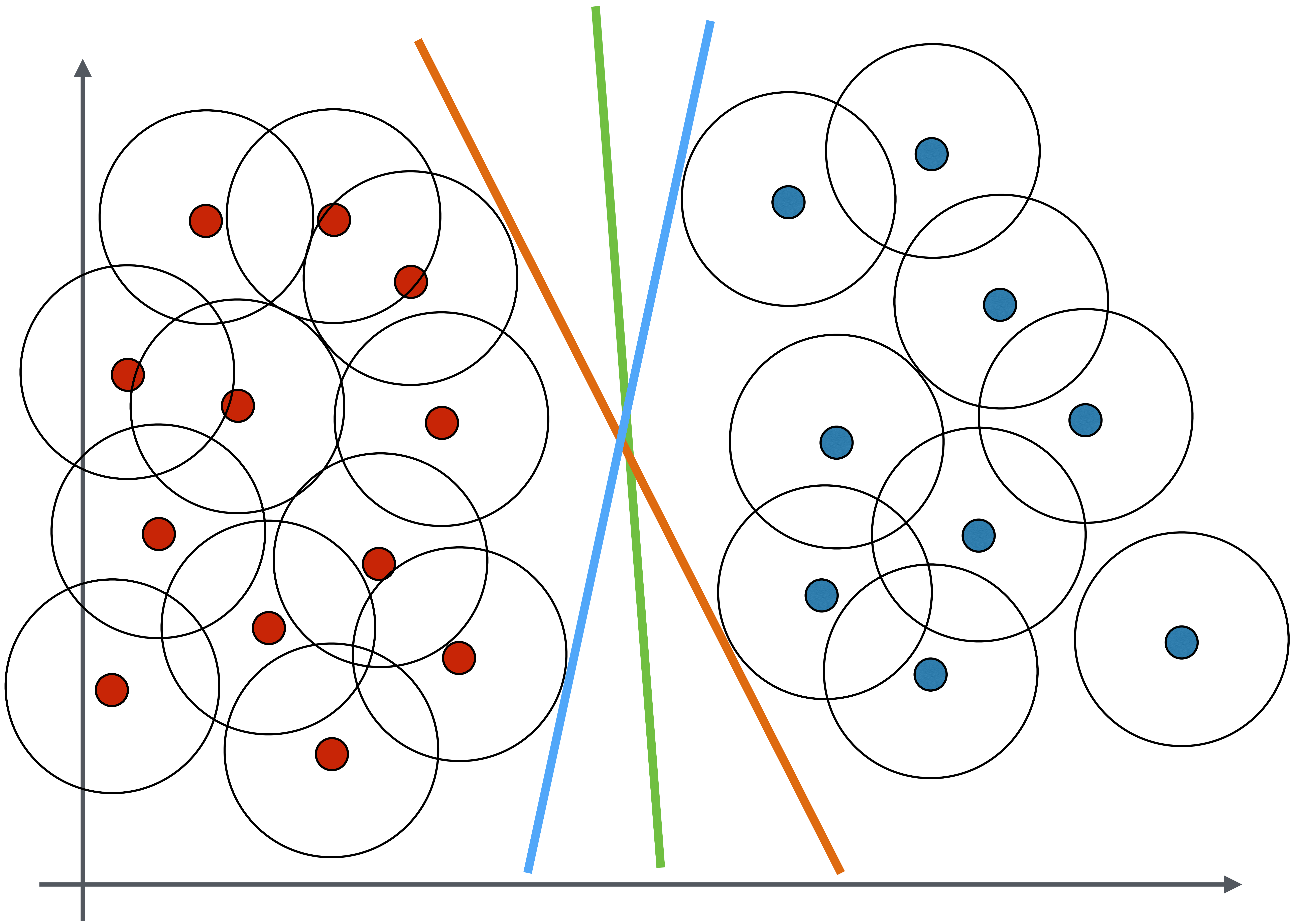
Margin: Bubbles around samples



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

As the margin increases, the feasible region reduces

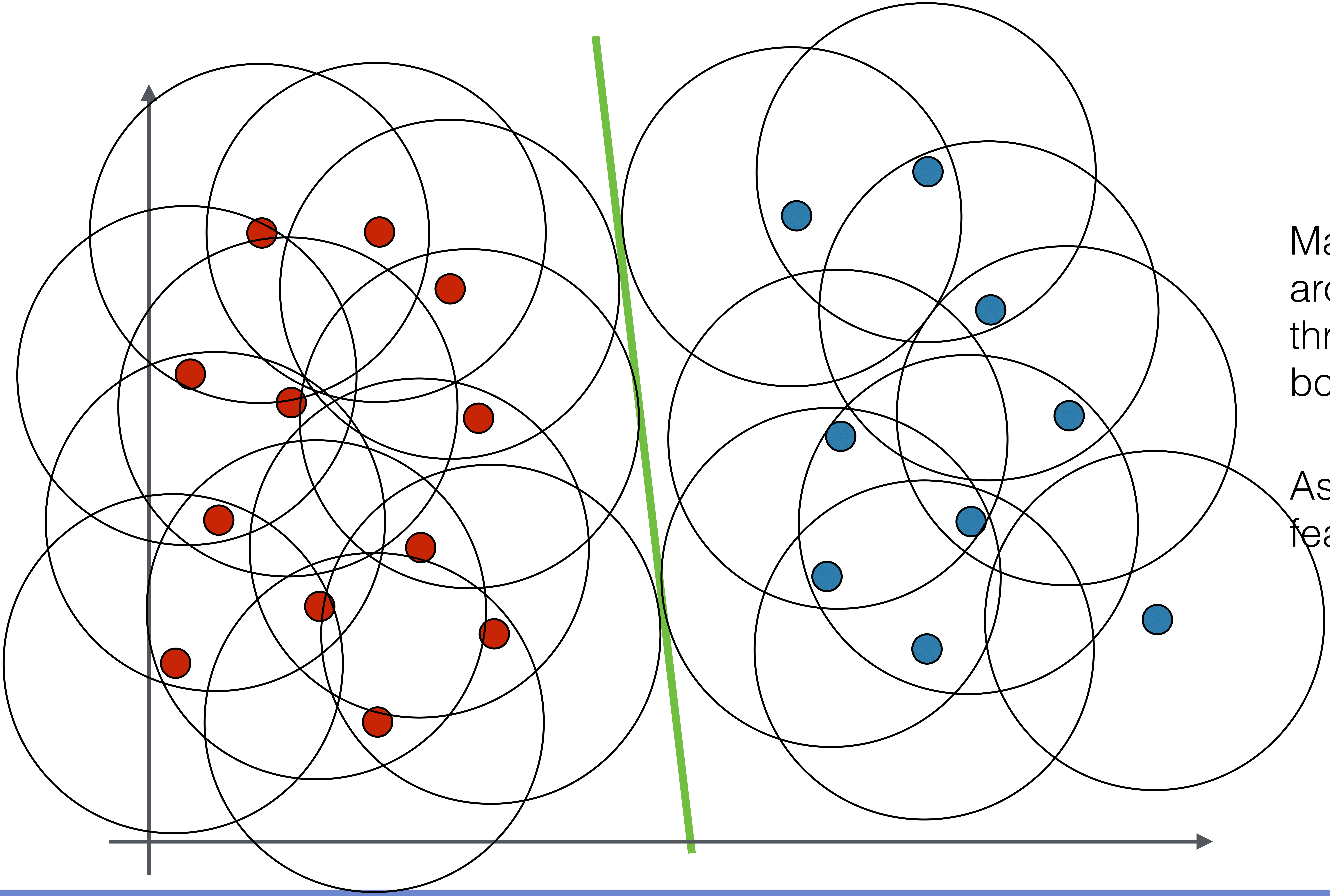
Margin: Bubbles around samples



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

As the margin increases, the feasible region reduces

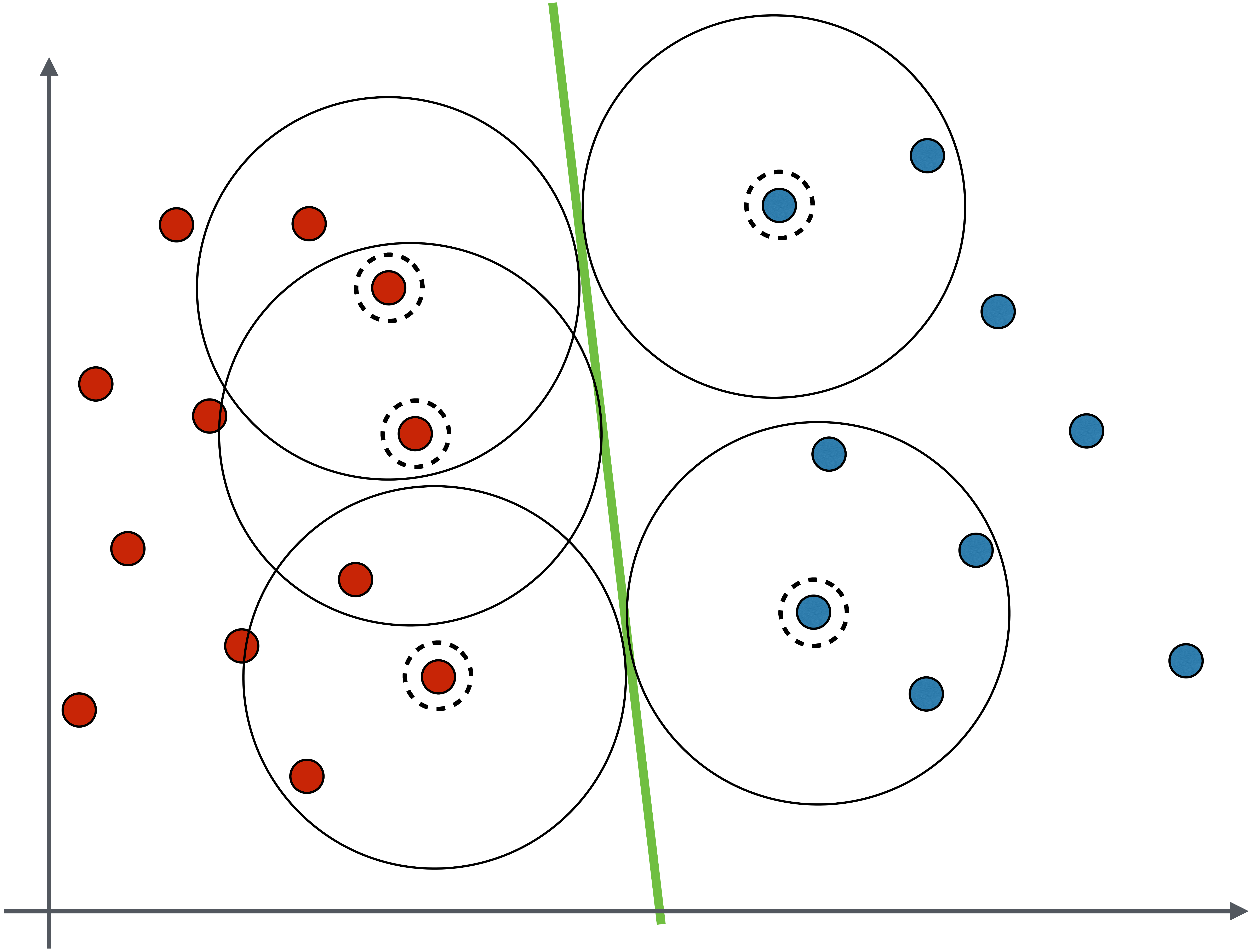
Margin: Bubbles around samples



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

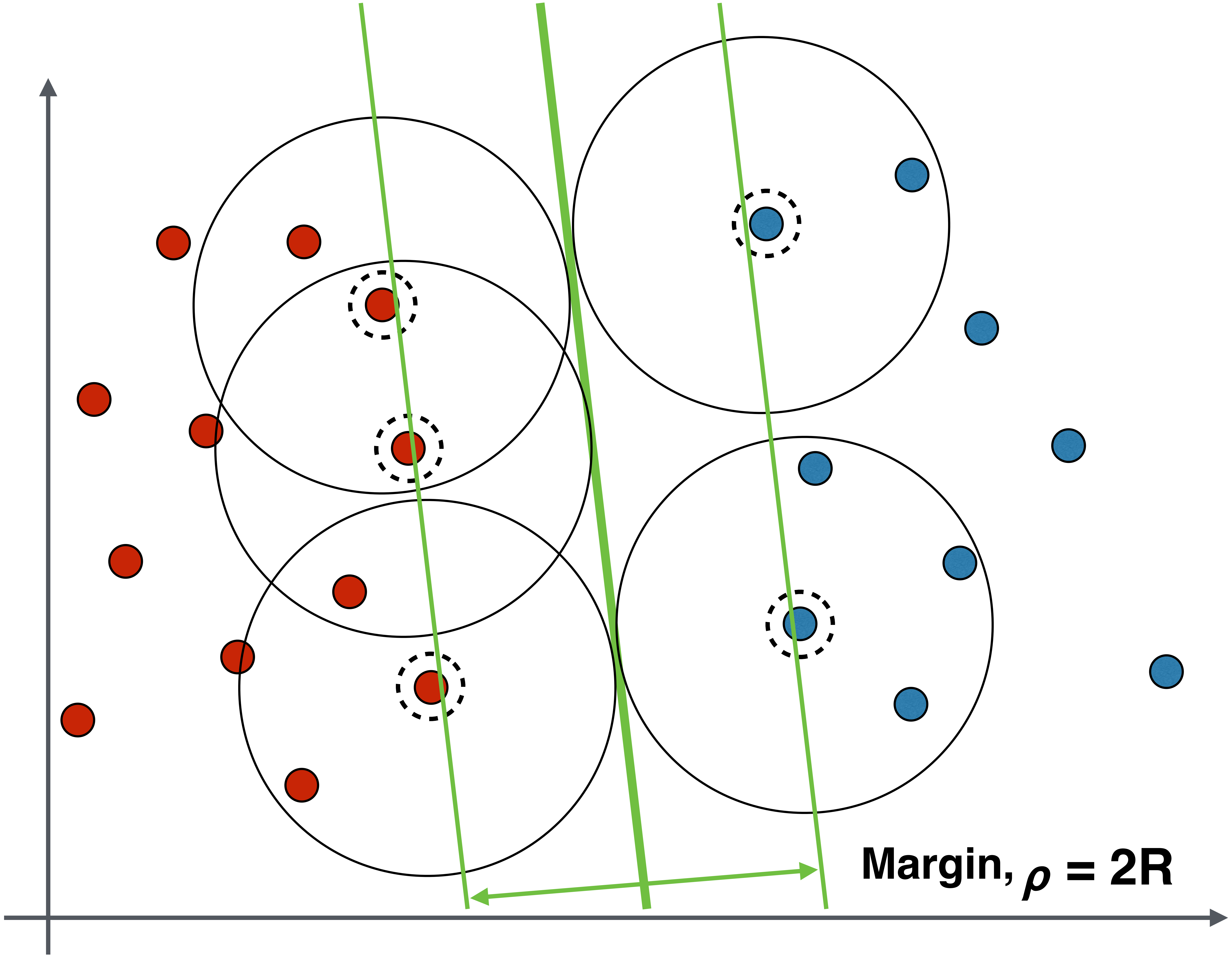
As the margin increases, the feasible region reduces

Margin: Bubbles around samples



A few samples control the Decision Boundary

Band vs. Bubbles

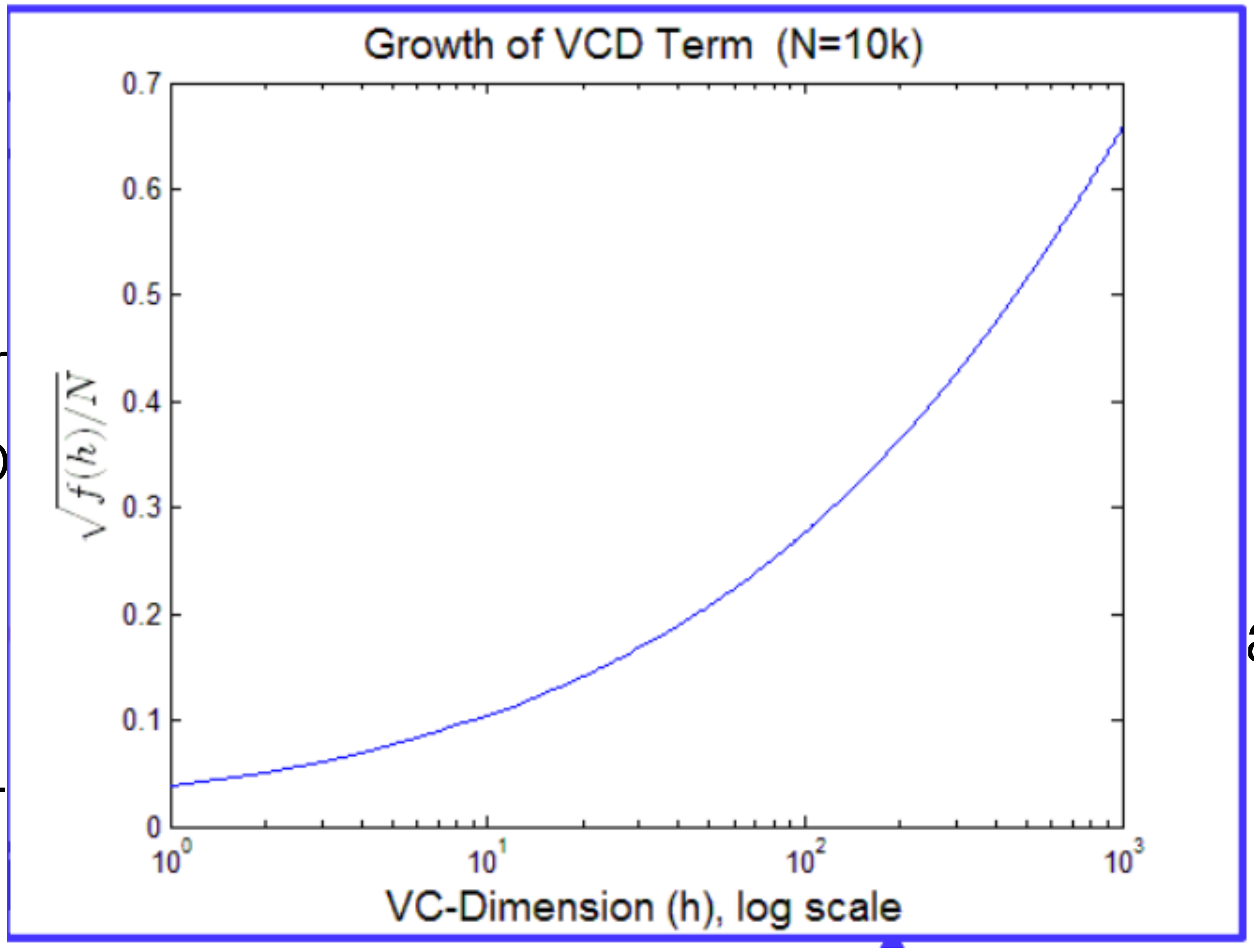


Samples that support the boundary are called ***Support Vectors***

Both interpretations lead yield the same decision boundary

Break though work from Vapni

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



Bound on expected loss: $$R(\alpha) \le R_{train}(\alpha) + \boxed{\sqrt{f(h)/N}}$$

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

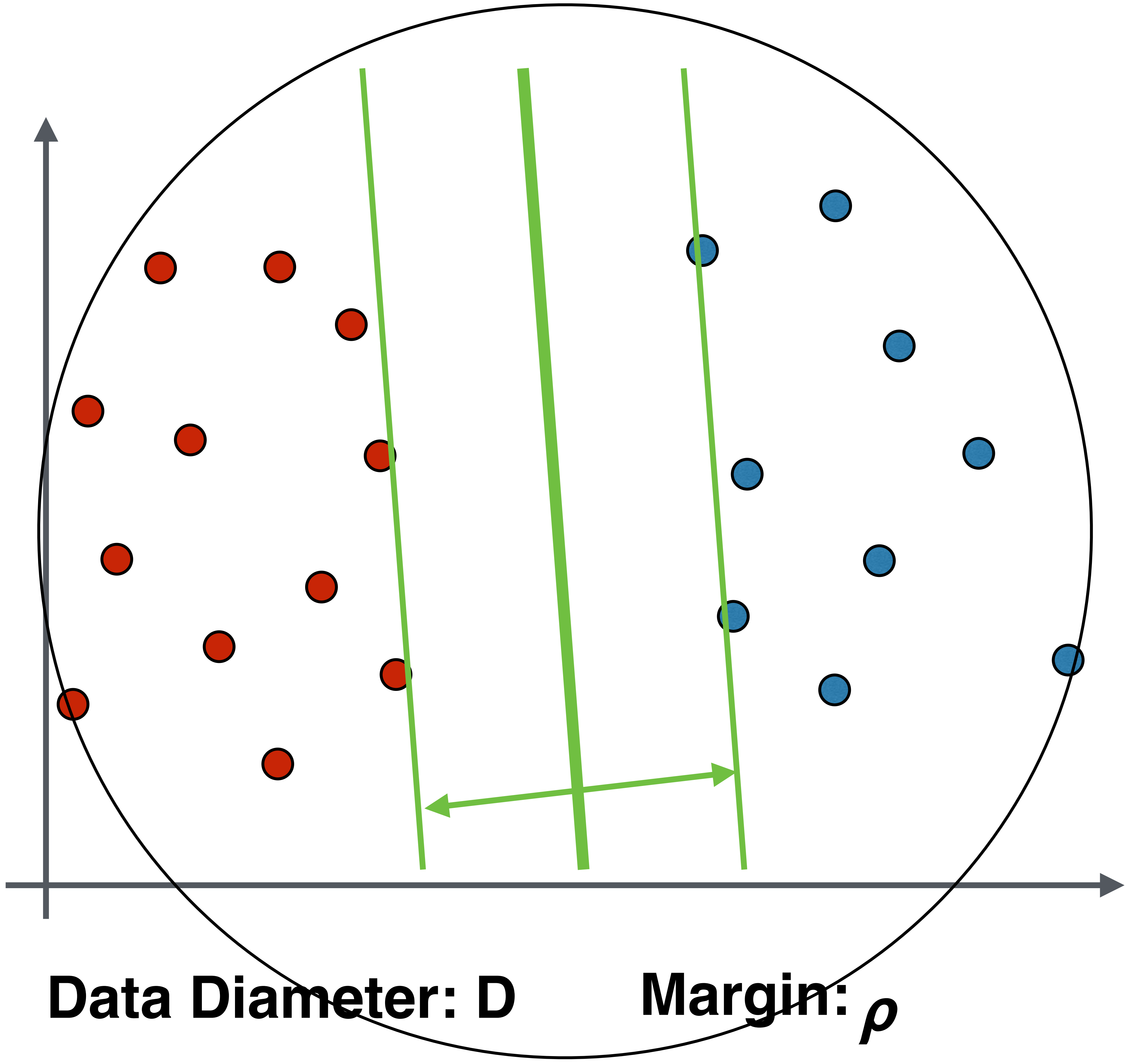
$$f(h) = h + h\log(2N) - h\log(h) - c$$

Why maximise the margin?

- To reduce test error, keep training error low (say 0), and minimize the VC-dimension, h.

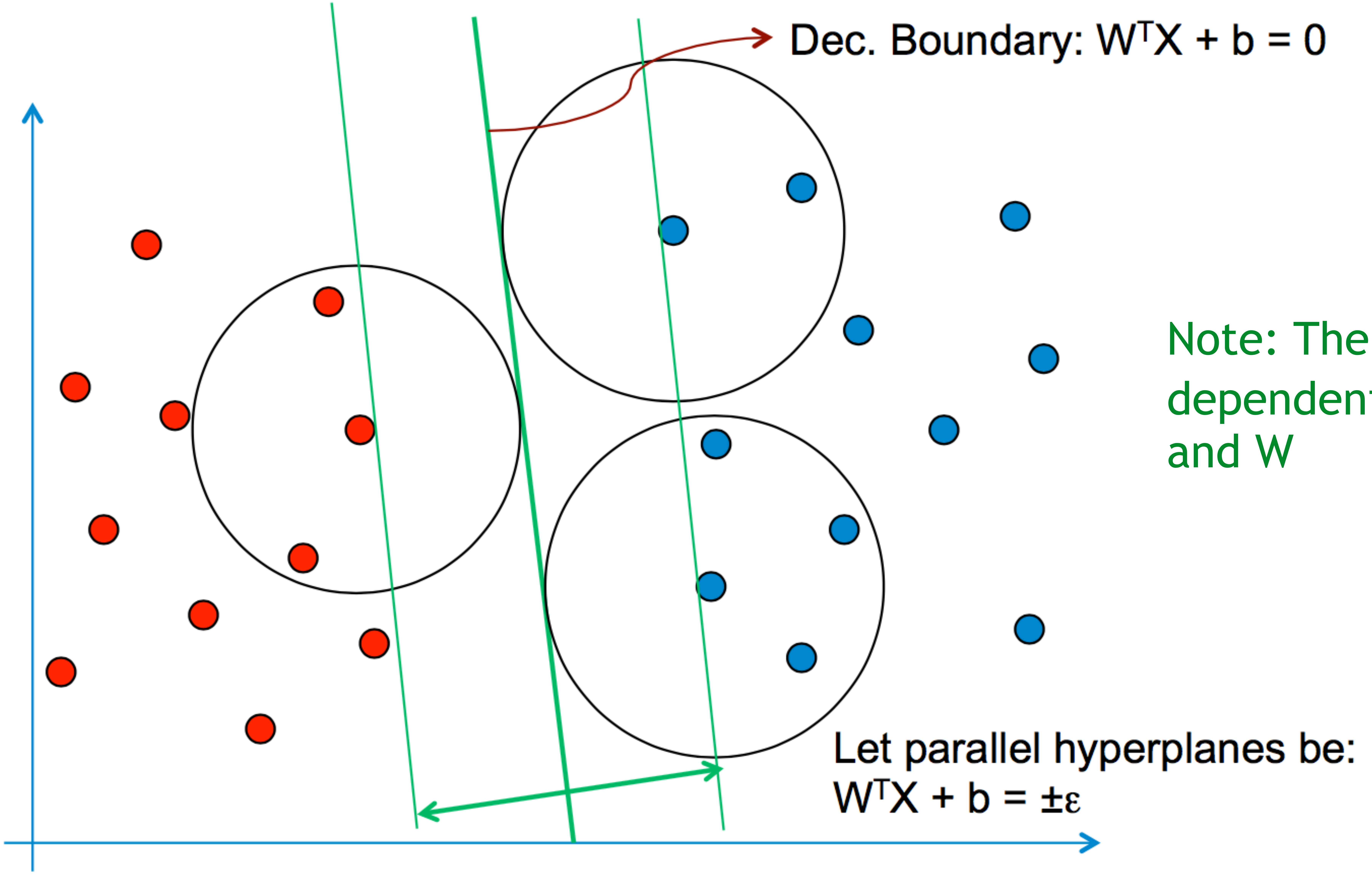
$$\text{Relative Margin}: \ {}^{\rho}/_{\text{D}}$$

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



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

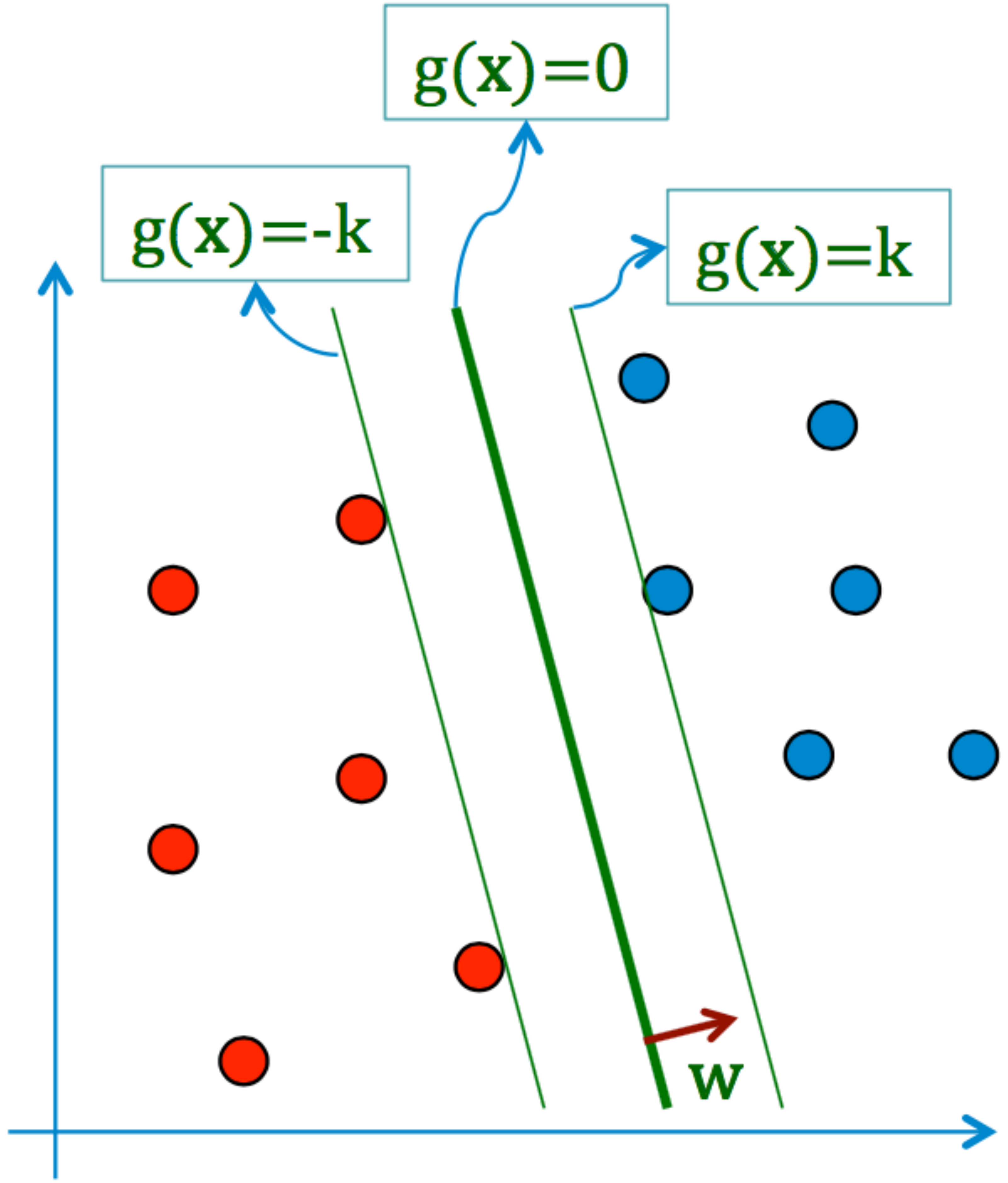
Formalizing the margin



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

Formulation

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

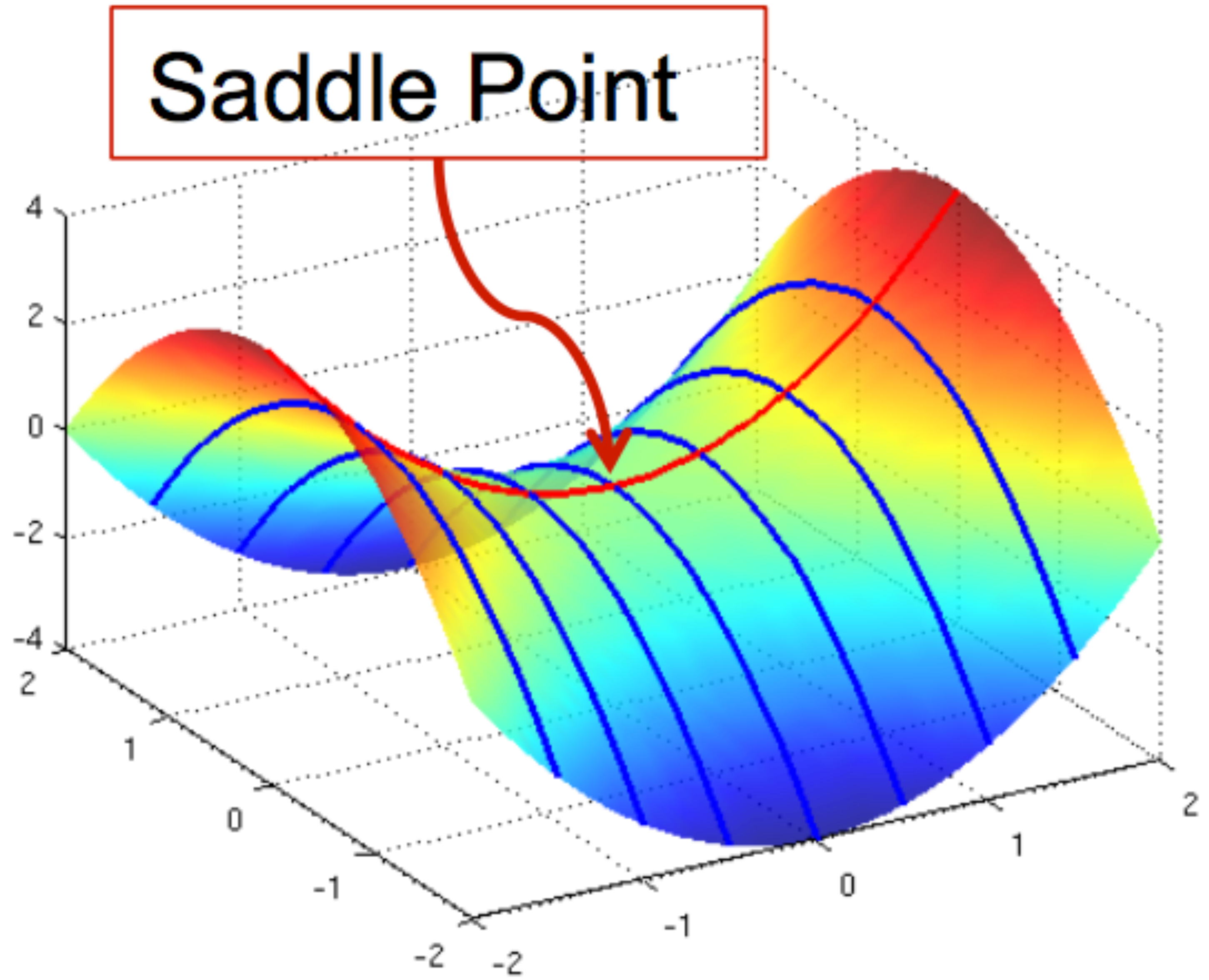


Optimization

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

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

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: $\boxed{1: \frac{\partial J}{\partial \mathbf{w}} = 0}$ and $\boxed{2: \frac{\partial J}{\partial b} = 0}$

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

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

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

Solving the Dual form

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

Subject to $\alpha_i \geq 0 \quad \forall_{\mathrm{i}}$ and $\sum_{\mathrm{i}=1}^{N} \alpha_{\mathrm{i}} d_i = 0$

QP Solver

$\downarrow$

$\boxed{\alpha_i}$

$\downarrow$

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