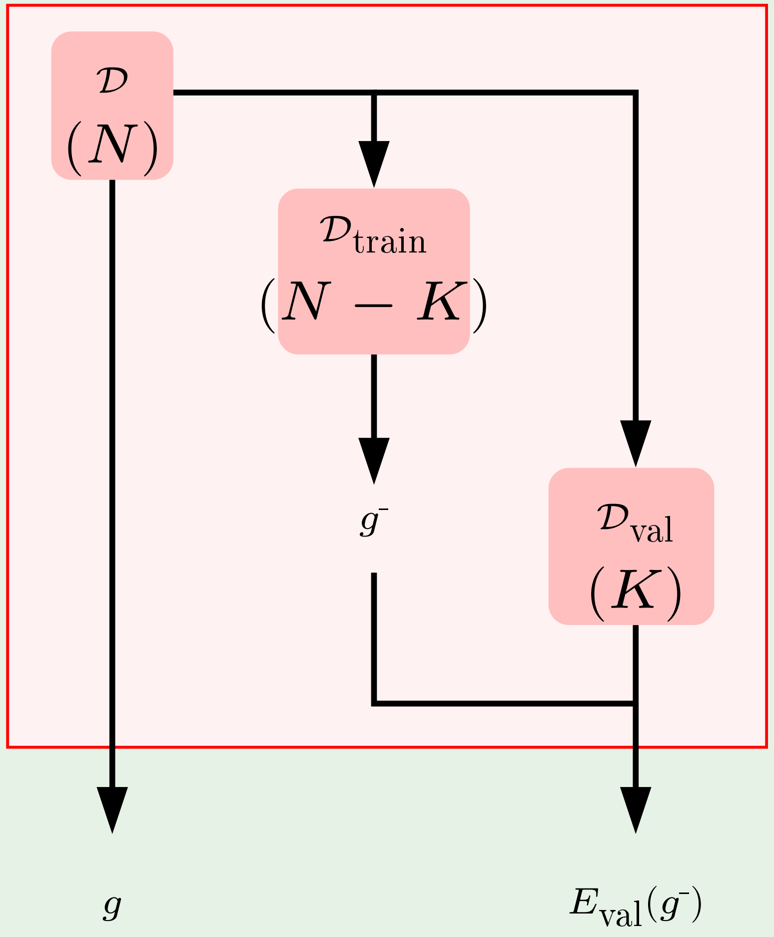
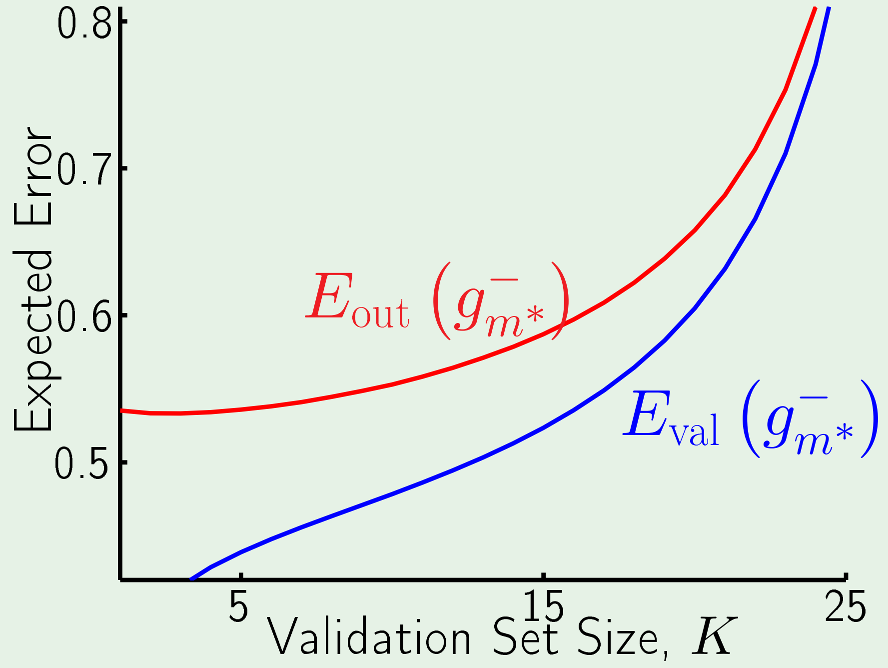
Review of Lecture 13

- Validation



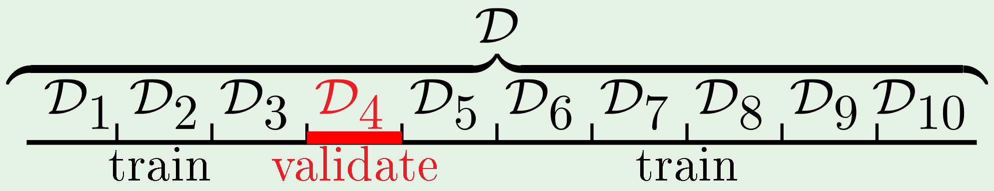
$E_{\text{val}}(g^-)$ estimates $E_{\text{out}}(g)$

- Data contamination



$\mathcal{D}_{\text{val}}$ slightly contaminated

- Cross validation



10-fold cross validation

Learning From Data

Yaser S. Abu-Mostafa
California Institute of Technology

Lecture 14: Support Vector Machines

Sponsored by Caltech's Provost Office, E&AS Division, and IST • Thursday, May 17, 2012

Outline

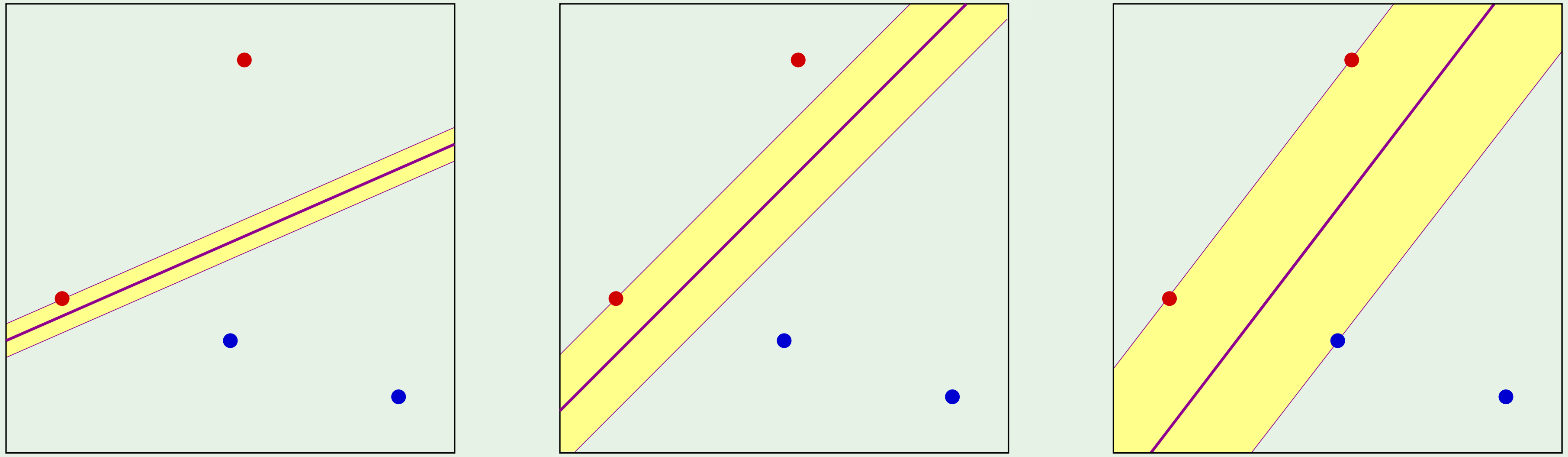
- Maximizing the margin
- The solution
- Nonlinear transforms

Better linear separation

Linearly separable data

Different separating lines

Which is best?



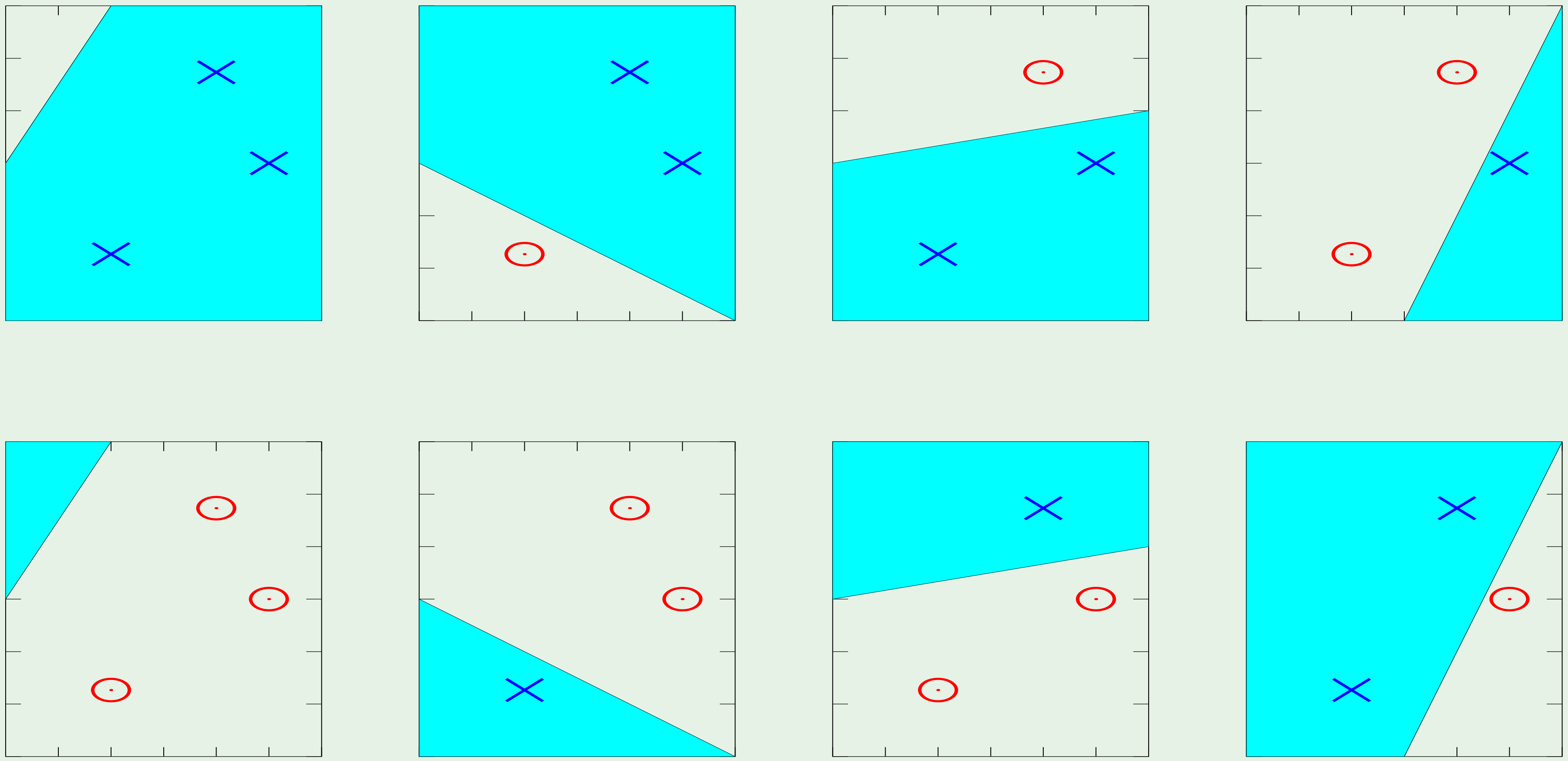
Two questions:

1. Why is bigger margin better?

2. Which $\mathbf{w}$ maximizes the margin?

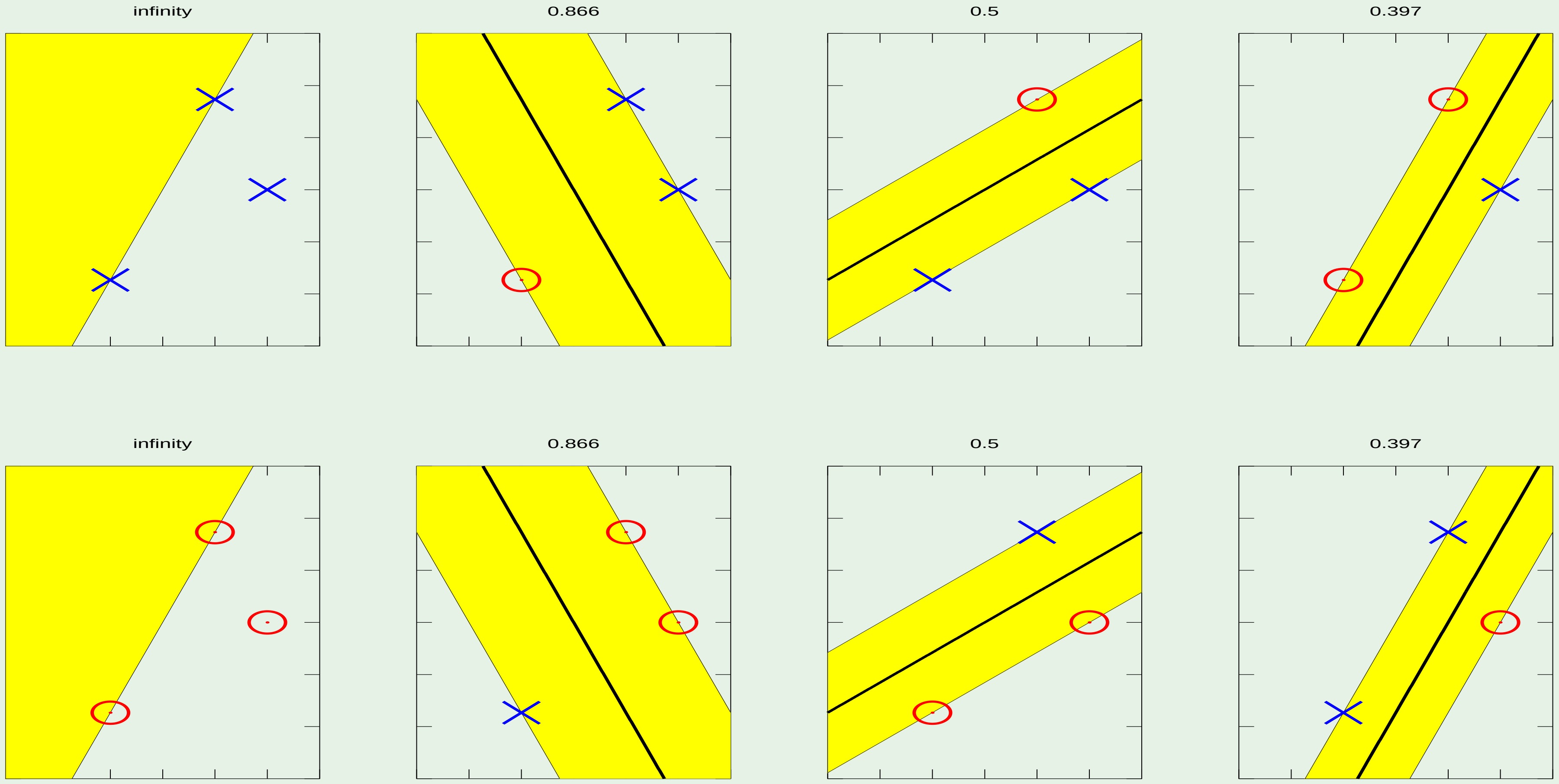
Remember the growth function?

All dichotomies with any line:



Dichotomies with fat margin

Fat margins imply fewer dichotomies



Finding $\mathbf{w}$ with large margin

Let $\mathbf{x}_n$ be the nearest data point to the plane $\mathbf{w}^\intercal\mathbf{x} = 0$. How far is it?

2 preliminary technicalities:

1. **Normalize $\mathbf{w}$:**

   $$\left|\mathbf{w}^\intercal\mathbf{x}_n\right| = 1$$

2. **Pull out w_0:**

   $\mathbf{w} = (w_1, \cdots, w_d)$ apart from b

   The plane is now $\boxed{\mathbf{w}^\intercal\mathbf{x} + b = 0}$ (no x_0)

Computing the distance

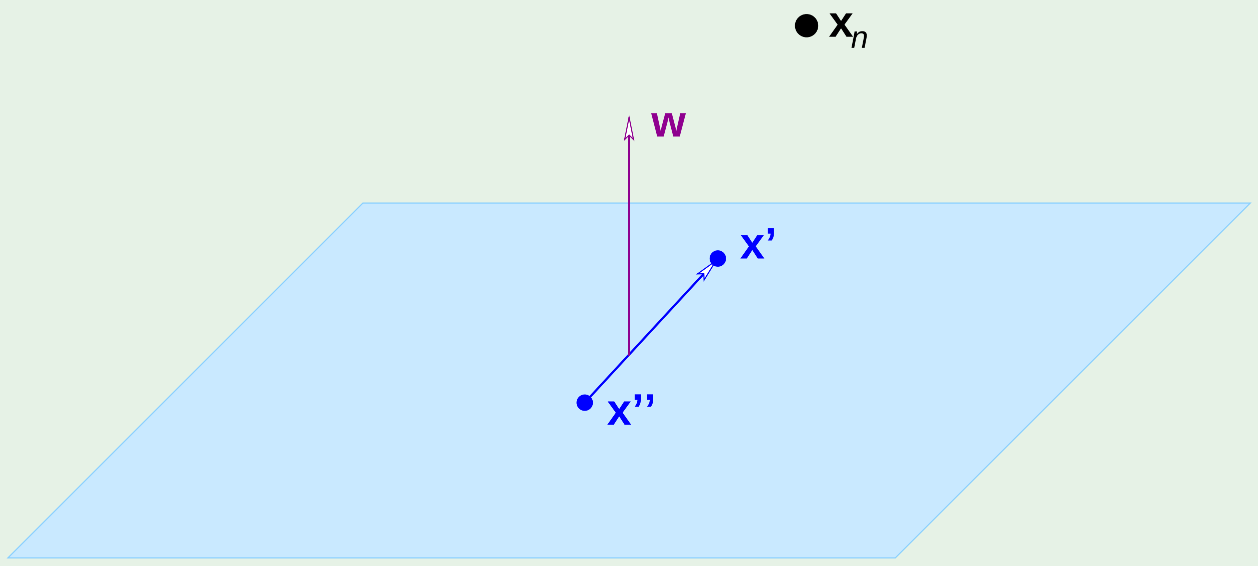
The distance between $\mathbf{x}_n$ and the plane $\mathbf{w}^\intercal\mathbf{x} + b = 0$ where $|\mathbf{w}^\intercal\mathbf{x}_n + b| = 1$

The vector $\mathbf{w}$ is $\perp$ to the plane in the $\mathcal{X}$ space:

Take $\mathbf{x}'$ and $\mathbf{x}''$ on the plane

$$\mathbf{w}^\intercal\mathbf{x}' + b = 0 \quad \text{and} \quad \mathbf{w}^\intercal\mathbf{x}'' + b = 0$$

$$\Longrightarrow \quad \mathbf{w}^\intercal(\mathbf{x}' - \mathbf{x}'') = 0$$

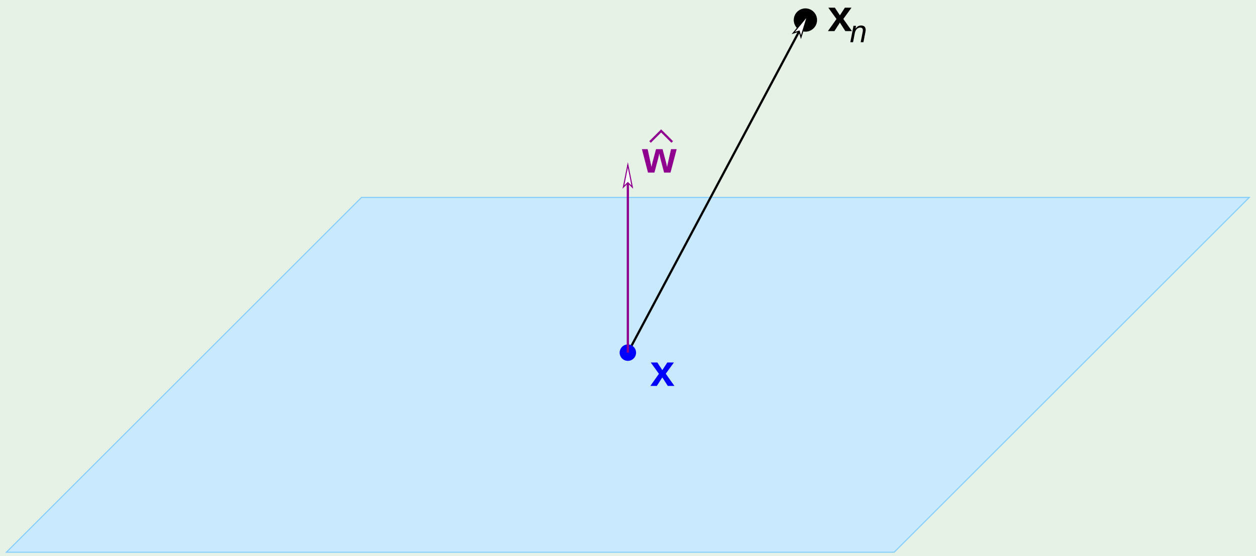


and the distance is ...

Distance between $\mathbf{x}_n$ and the plane:

Take any point $\mathbf{x}$ on the plane

Projection of $\mathbf{x}_n - \mathbf{x}$ on $\mathbf{w}$



$$\hat{\mathbf{w}} = \frac{\mathbf{w}}{\|\mathbf{w}\|} \implies \text{distance} = \left|\hat{\mathbf{w}}^{\intercal}(\mathbf{x}_n - \mathbf{x})\right|$$

$$\text{distance} = \frac{1}{\|\mathbf{w}\|}\left|\mathbf{w}^{\intercal}\mathbf{x}_n - \mathbf{w}^{\intercal}\mathbf{x}\right| = \frac{1}{\|\mathbf{w}\|}\left|\mathbf{w}^{\intercal}\mathbf{x}_n + b - \mathbf{w}^{\intercal}\mathbf{x} - b\right| = \frac{1}{\|\mathbf{w}\|}$$

The optimization problem

Maximize $\dfrac{1}{\|\mathbf{w}\|}$

subject to $\displaystyle\min_{n=1,2,\ldots,N} |\mathbf{w}^{\intercal}\mathbf{x}_n + b| \ = \ 1$

Notice: $|\mathbf{w}^{\intercal}\mathbf{x}_n + b| \ = y_n \left(\mathbf{w}^{\intercal}\mathbf{x}_n + b\right)$

Minimize $\dfrac{1}{2}\,\mathbf{w}^{\intercal}\mathbf{w}$

subject to $y_n \left(\mathbf{w}^{\intercal}\mathbf{x}_n + b\right) \geq 1 \quad$ for $\quad n = 1, 2, \ldots, N$

Outline

- Maximizing the margin
- The solution
- Nonlinear transforms

Constrained optimization

Minimize $\quad \frac{1}{2}\,\mathbf{w}^\intercal\mathbf{w}$

subject to $\quad y_n\left(\mathbf{w}^\intercal\mathbf{x}_n + b\right) \geq 1 \quad$ for $\quad n = 1, 2, \ldots, N$

$\mathbf{w} \in \mathbb{R}^d,\ b \in \mathbb{R}$

Lagrange? inequality constraints $\Longrightarrow$ KKT

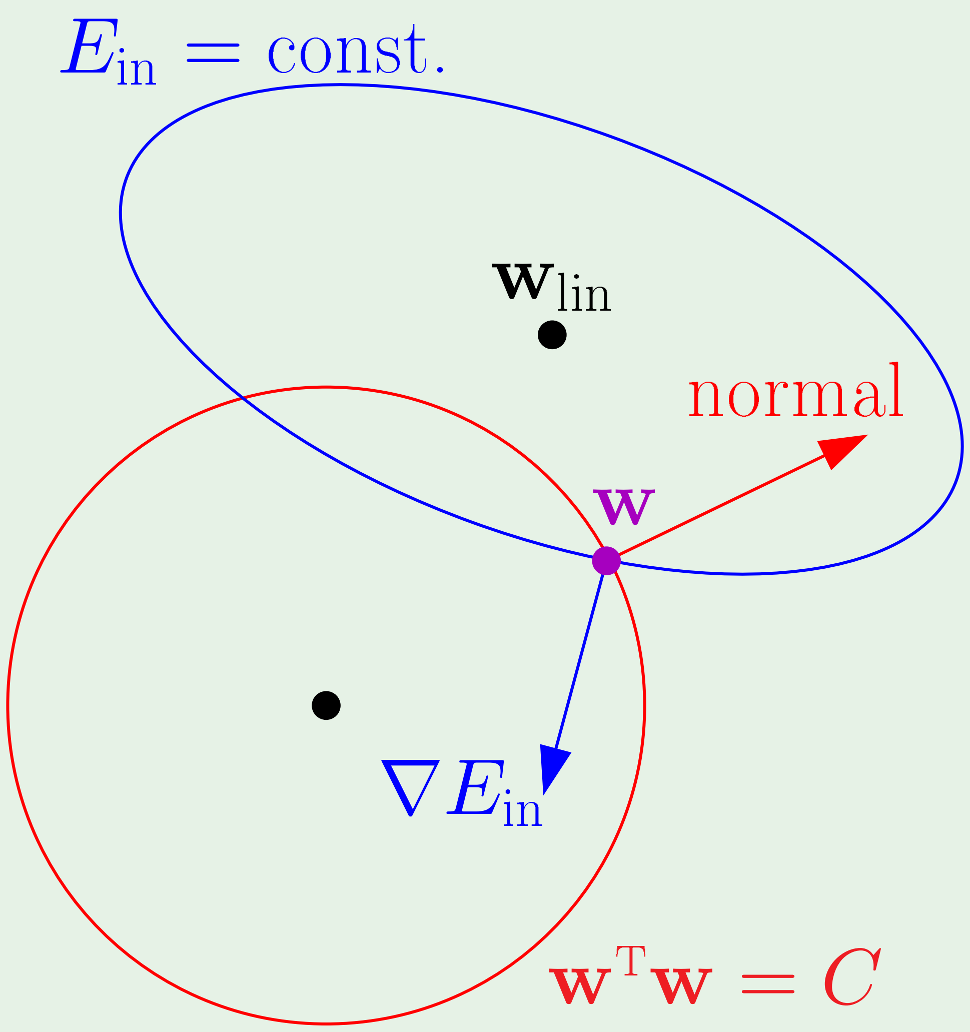
We saw this before

Remember regularization?

Minimize $\quad E_{\text{in}}(\mathbf{w}) \;=\; \frac{1}{N}\,(\mathrm{Z}\mathbf{w}-\mathbf{y})^{\intercal}(\mathrm{Z}\mathbf{w}-\mathbf{y})$

subject to: $\quad \mathbf{w}^{\intercal}\mathbf{w} \le C$

∇E_{in} normal to constraint



	optimize	constrain
Regularization:	E_{in}	$\mathbf{w}^{\intercal}\mathbf{w}$
SVM:	$\mathbf{w}^{\intercal}\mathbf{w}$	E_{in}

Lagrange formulation

Minimize $\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{w} - \sum_{n=1}^{N} \alpha_n (y_n (\mathbf{w}^{\mathsf{T}}\mathbf{x}_n + b) - 1)$

w.r.t. $\mathbf{w}$ and b and maximize w.r.t. each $\alpha_n \geq 0$

$$\nabla_{\mathbf{w}}\mathcal{L} = \mathbf{w} - \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n = \mathbf{0}$$

$$\frac{\partial \mathcal{L}}{\partial b} = - \sum_{n=1}^{N} \alpha_n y_n = 0$$

Substituting ...

$$\mathbf{w} \;=\; \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n \qquad \text{and} \qquad \sum_{n=1}^{N} \alpha_n y_n \;=\; 0$$

in the Lagrangian

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\,\mathbf{w}^{\mathsf{T}}\mathbf{w} \;-\; \sum_{n=1}^{N} \alpha_n \left(y_n \left(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n + b\right) - 1\right)$$

we get

$$\mathcal{L}(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} y_n y_m \, \alpha_n \alpha_m \, \mathbf{x}_n^{\mathsf{T}} \mathbf{x}_m$$

Maximize w.r.t. to $\boldsymbol{\alpha}$ subject to $\alpha_n \geq 0$ for $n = 1, \cdots, N$ **and** $\sum_{n=1}^{N} \alpha_n y_n = 0$

The solution - quadratic programming

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\,\boldsymbol{\alpha}^\intercal \underbrace{\begin{bmatrix} y_1y_1\ \mathbf{x}_1^\intercal\mathbf{x}_1 & y_1y_2\ \mathbf{x}_1^\intercal\mathbf{x}_2 & \dots & y_1y_N\ \mathbf{x}_1^\intercal\mathbf{x}_N \\ y_2y_1\ \mathbf{x}_2^\intercal\mathbf{x}_1 & y_2y_2\ \mathbf{x}_2^\intercal\mathbf{x}_2 & \dots & y_2y_N\ \mathbf{x}_2^\intercal\mathbf{x}_N \\ \dots & \dots & \dots & \dots \\ y_Ny_1\ \mathbf{x}_N^\intercal\mathbf{x}_1 & y_Ny_2\ \mathbf{x}_N^\intercal\mathbf{x}_2 & \dots & y_Ny_N\ \mathbf{x}_N^\intercal\mathbf{x}_N \end{bmatrix}}_{\text{quadratic coefficients}} \boldsymbol{\alpha} \;+\; \underbrace{\left(-\mathbf{1}^\intercal\right)}_{\text{linear}}\boldsymbol{\alpha}$$

subject to $\underbrace{\mathbf{y}^\intercal\boldsymbol{\alpha} = 0}_{\text{linear constraint}}$

$$\underbrace{\mathbf{0}}_{\text{lower bounds}} \quad \leq \quad \boldsymbol{\alpha} \quad \leq \quad \underbrace{\infty}_{\text{upper bounds}}$$

QP hands us $\boldsymbol{\alpha}$

Solution: $\boldsymbol{\alpha} = \alpha_1, \cdots, \alpha_N$

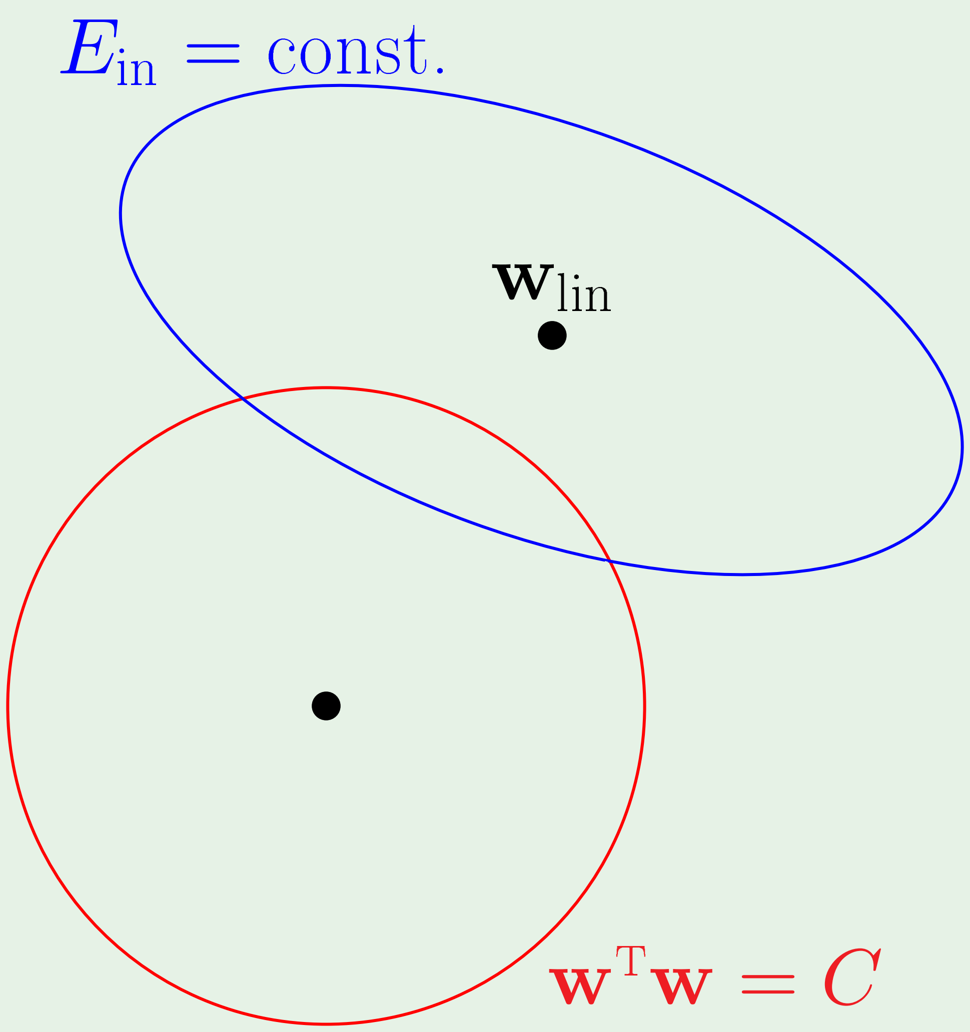
$$\Longrightarrow \quad \mathbf{w} \; = \; \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n$$

KKT condition: For $n = 1, \cdots, N$

$$\alpha_n \left(y_n \left(\mathbf{w}^{\mathsf{T}} \mathbf{x}_n + b \right) - 1 \right) \; = \; 0$$

We saw this before!

$\alpha_n > 0 \implies \mathbf{x}_n$ is a **support vector**



Support vectors

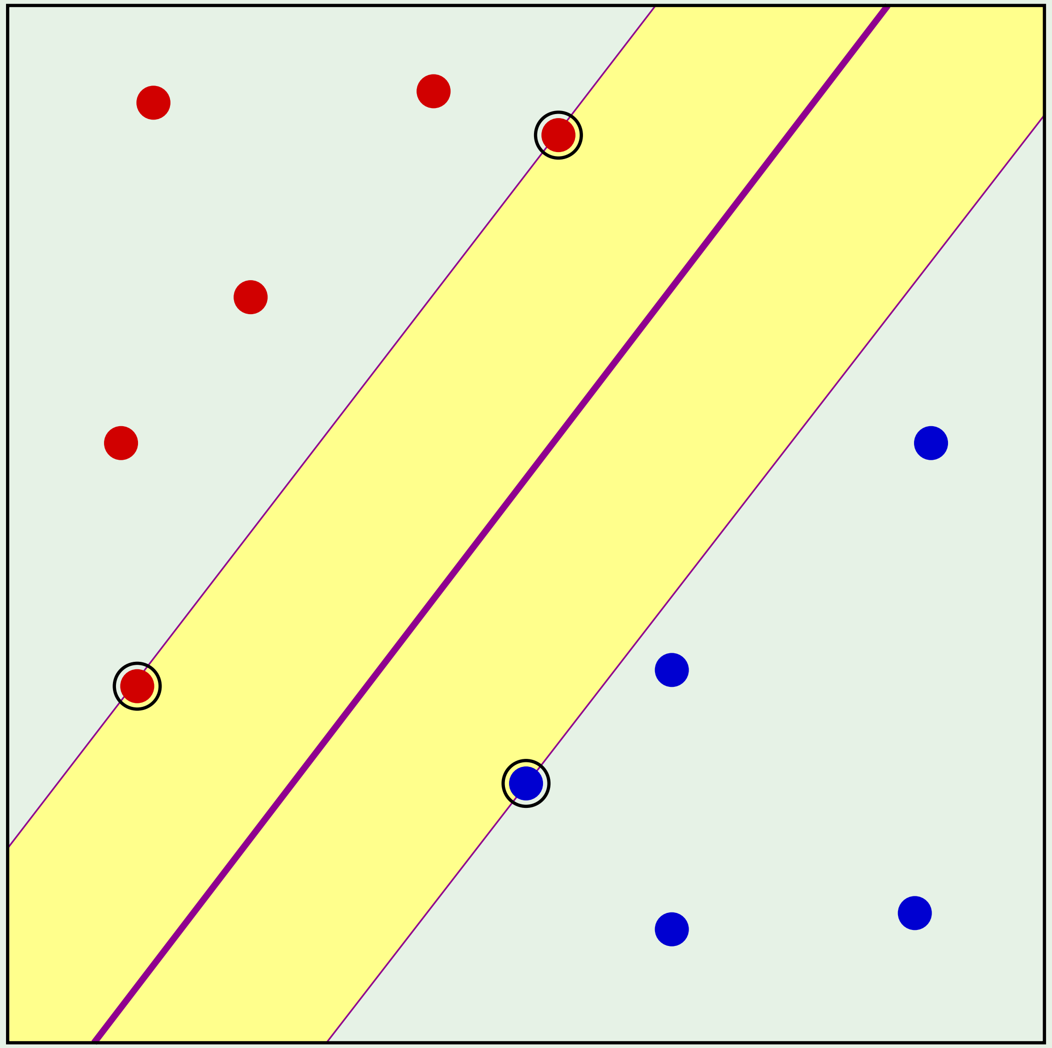
Closest $\mathbf{x}_n$'s to the plane: achieve the margin

$$\Longrightarrow \quad y_n\left(\mathbf{w}^\intercal\mathbf{x}_n + b\right) = 1$$

$$\mathbf{w} = \sum_{\mathbf{x}_n \text{ is SV}} \alpha_n y_n \mathbf{x}_n$$

Solve for b using any SV:

$$y_n\left(\mathbf{w}^\intercal\mathbf{x}_n + b\right) = 1$$



Outline

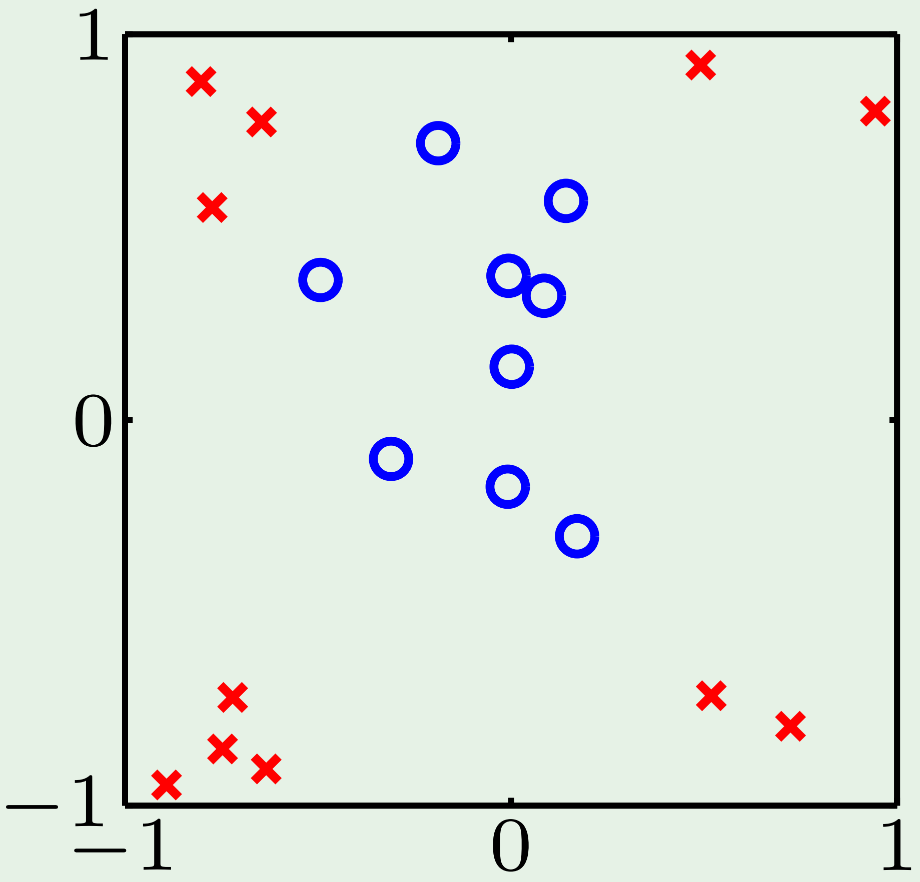
- Maximizing the margin

- The solution

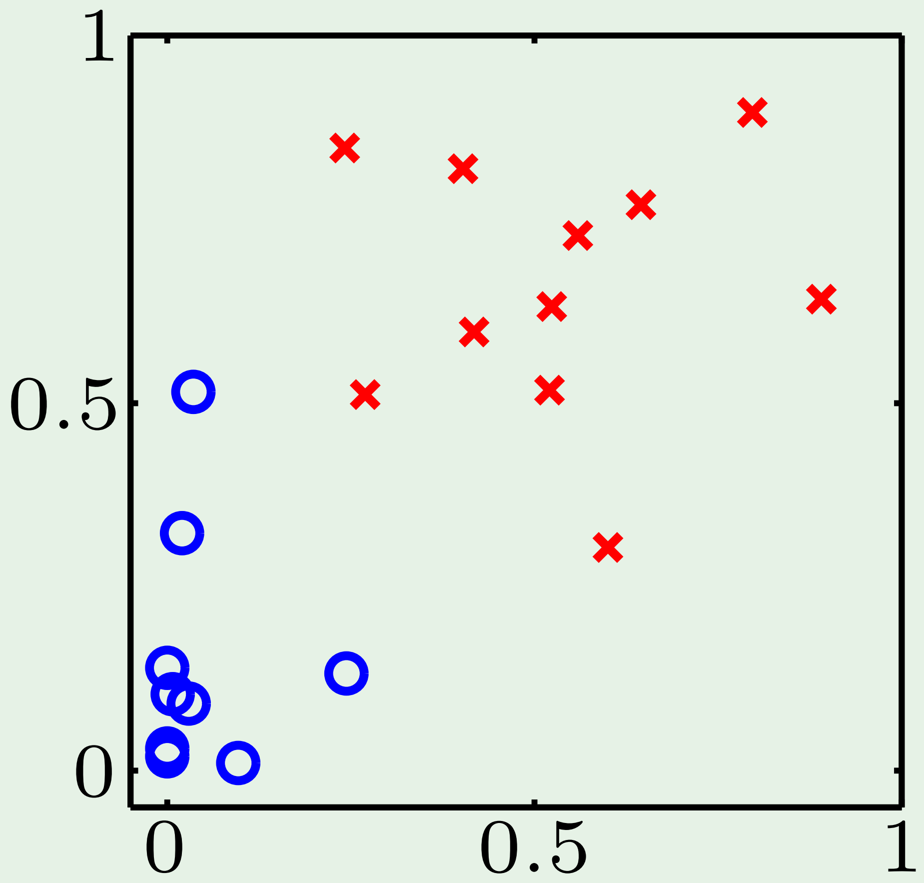
- Nonlinear transforms

z instead of x

$$\mathcal{L}(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} y_n y_m \, \alpha_n \alpha_m \, \mathbf{z}_n^{\mathsf{T}} \mathbf{z}_m$$



$\mathcal{X} \longrightarrow \mathcal{Z}$

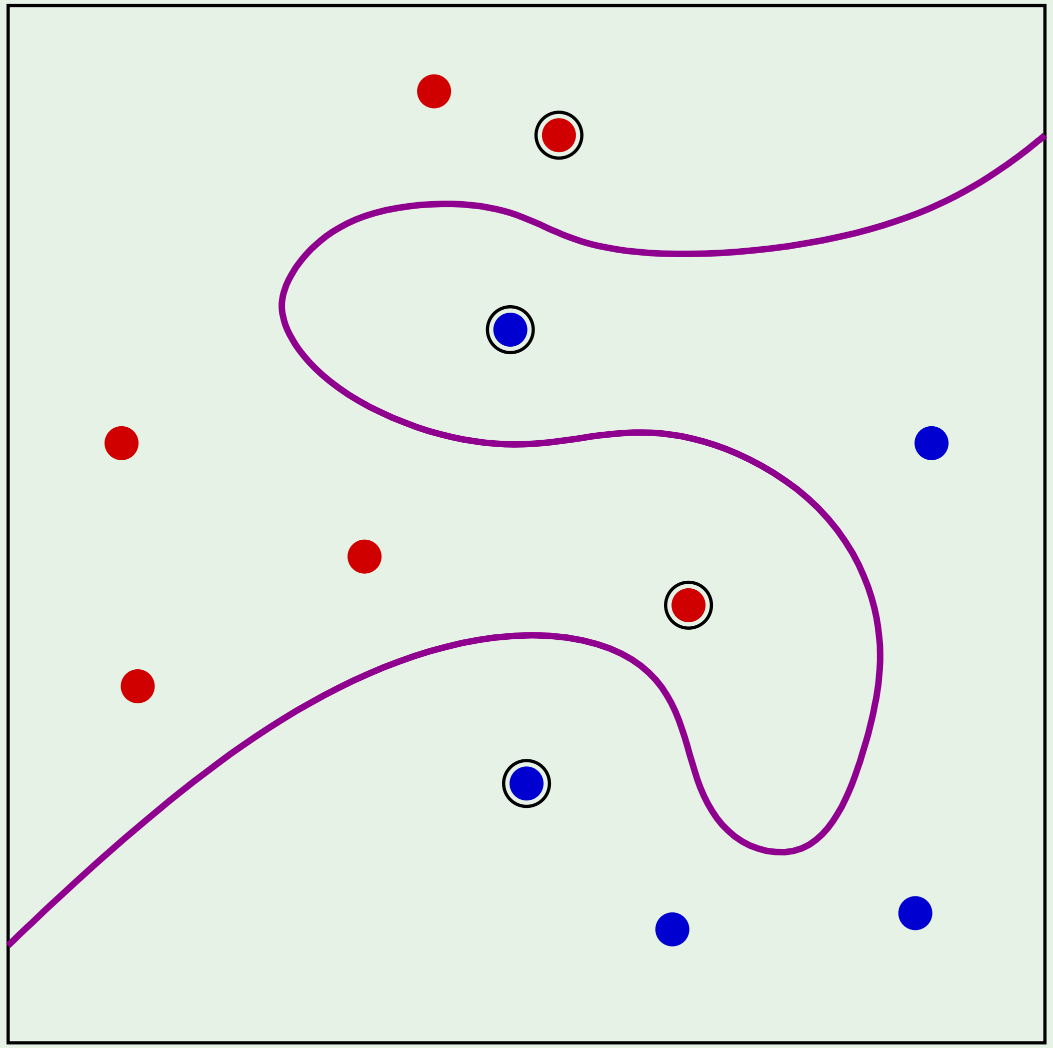


“Support vectors” in $\mathcal{X}$ space

Support vectors live in $\mathcal{Z}$ space

In $\mathcal{X}$ space, “pre-images” of support vectors

The margin is maintained in $\mathcal{Z}$ space



Generalization result

$$\mathbb{E}[E_{\text{out}}] \leq \frac{\mathbb{E}[\#\text{ of SV's}]}{N-1}$$