Geometry of Support Vector Machines

Mittagsseminar

Martin Jaggi, 28.2.2008

Computers are able to diagnose Alzheimer's disease faster and more accurately than experts

Medical Research News
Published: Sunday, 24-Feb-2008

Printer Friendly Email to a Friend

Computers are able to diagnose Alzheimer's disease faster and more accurately than experts, according to research published in the journal Brain.

The findings may help ensure that patients are diagnosed earlier, increasing treatment options.

According to the Alzheimer's Research Trust, there are over 700,000 people currently living in the UK with dementia, of which Alzheimer's disease, a neurodegenerative disease, is the most common form.

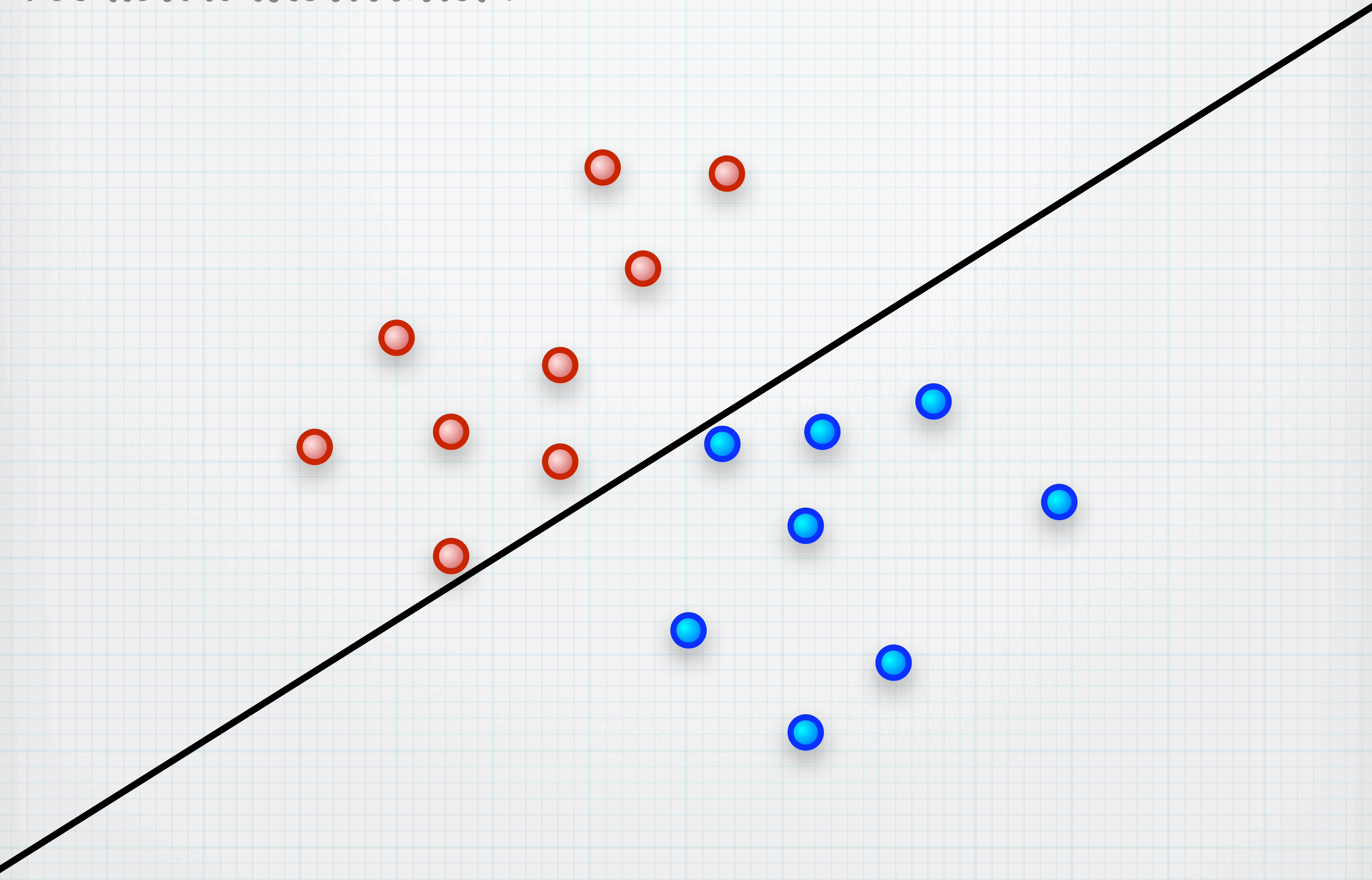
Alzheimer's is caused by the build up in the brain of plaques and neurofibrillary tangles (tangles of brain tissue filaments), leading the brain to atrophy. Definitive diagnosis is usually only possible after death, but Alzheimer's is usually diagnosed using a combination of brain scans, blood tests and interviews carried out by a trained clinician. The tests are time consuming, and distinguishing the disease from other forms of dementia can be difficult. The accuracy of diagnosis is only about 85%

Now, a team of researchers led by scientists at the Wellcome Trust Centre for Neuroimaging at University College London, has shown that scans of patients with Alzheimer's can be distinguished from those of healthy individuals and patients with other forms of dementia. Computers can identify the characteristic damage of Alzheimer's disease with an accuracy as high as 96%.

Contents

- Introduction to Support Vector Machines
  - Kernel trick
  - allowing outliers
- A new geometric solution algorithm
- SVM as an LP-type problem
- SVM as a smallest enclosing ball problem

Geometric Motivation:



Linear Classification

Given: points $x_i \in \mathbb{R}^n$

class labels $y_i \in \pm 1$

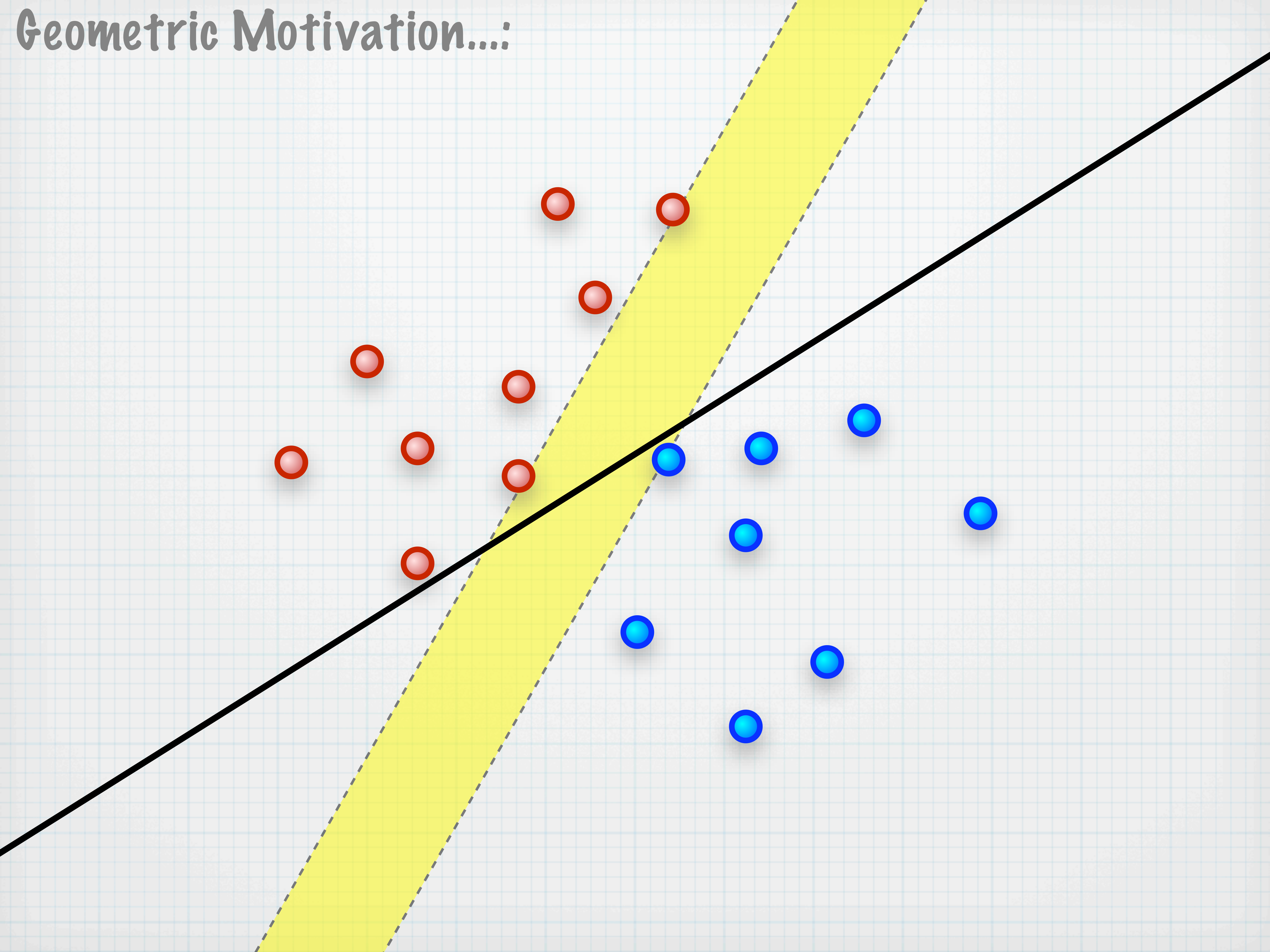
$$I_+ := \{i \mid y_i = +1\}$$
$$I_- := \{i \mid y_i = -1\}$$

Goal:

Find a linear classification function that best fits the data

Find $\omega \in \mathbb{R}^n, b \in \mathbb{R}$ **s.t.** $sign(\omega^T x_i + b) \approx y_i$

Geometric Motivation...:



Maximum Margin Separation

PRIMAL

$$\min_{\omega,b,\rho} \quad \|\omega\|^2 - 2\rho$$

$$s.t. \quad y_i(\omega^T x_i + b) \geq \rho \quad \forall i$$

$$\omega = \frac{1}{2} \sum_i \alpha_i y_i x_i$$

DUAL

$$\min_{\alpha} \quad \|\textstyle\sum_{I_+} \alpha_i x_i - \sum_{I_-} \alpha_i x_i\|^2$$

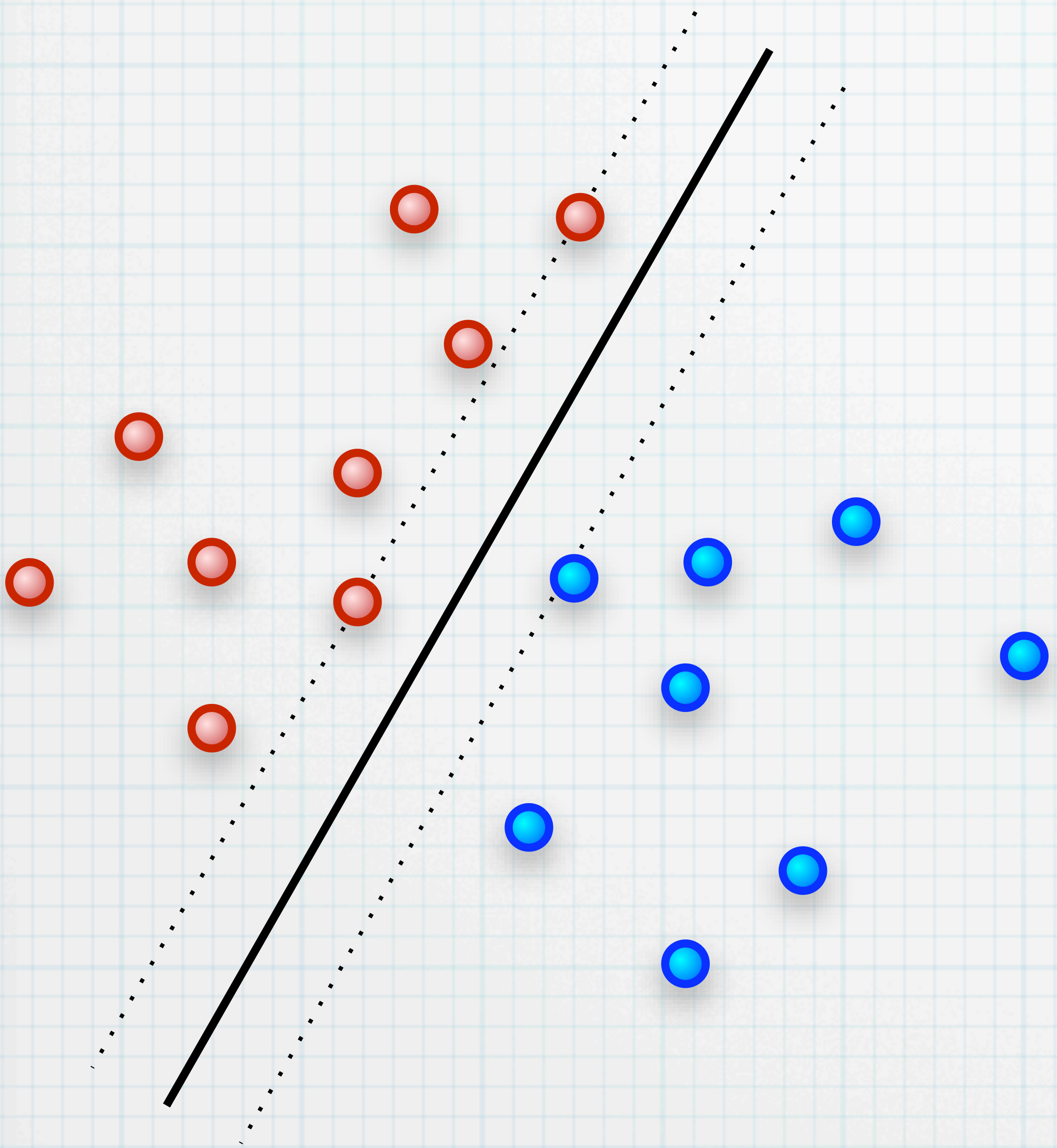
$$s.t. \quad \alpha_i \geq 0 \quad \forall i$$

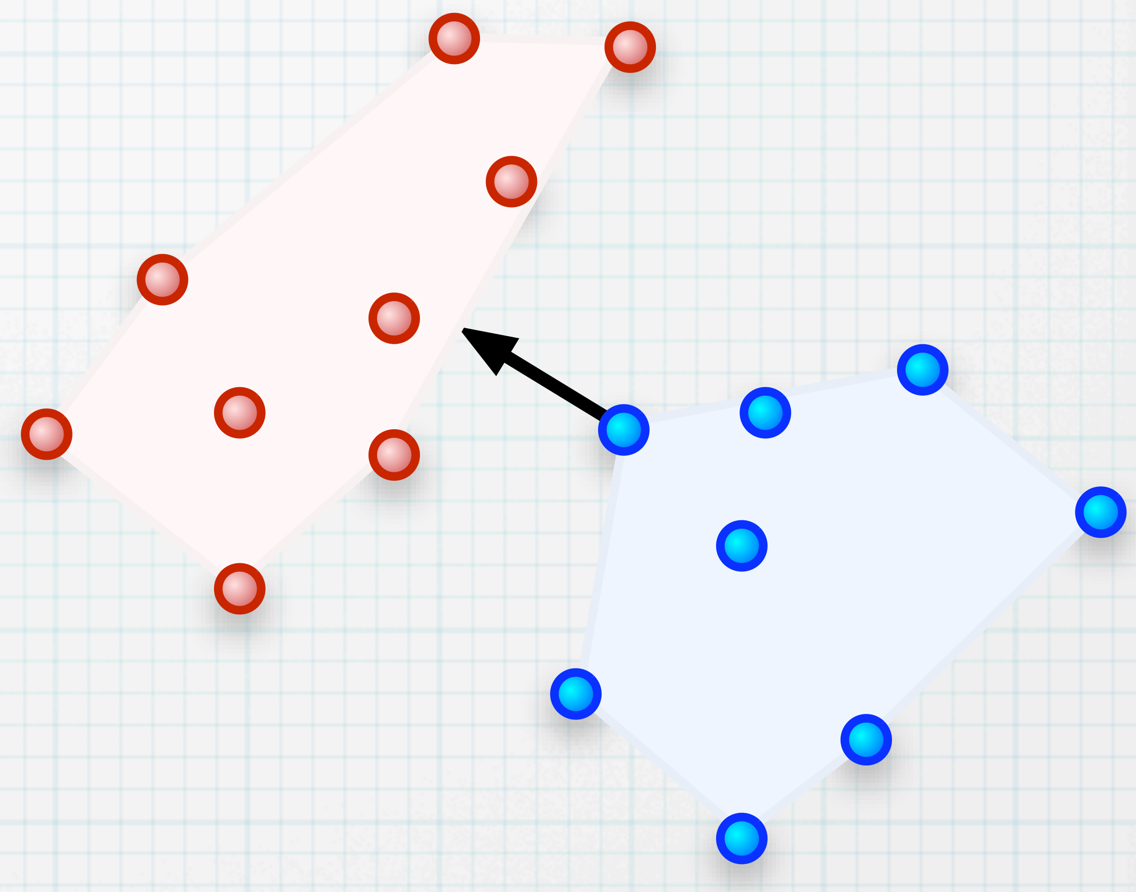
$$\textstyle\sum_{I_+} \alpha_i = 1$$

$$\textstyle\sum_{I_-} \alpha_i = 1$$

Polytope Distance

Geometric Interpretation:





$$\min_{\alpha} \quad \sum_{i,j} \alpha_i \alpha_j \, y_i y_j \, x_i^T x_j$$

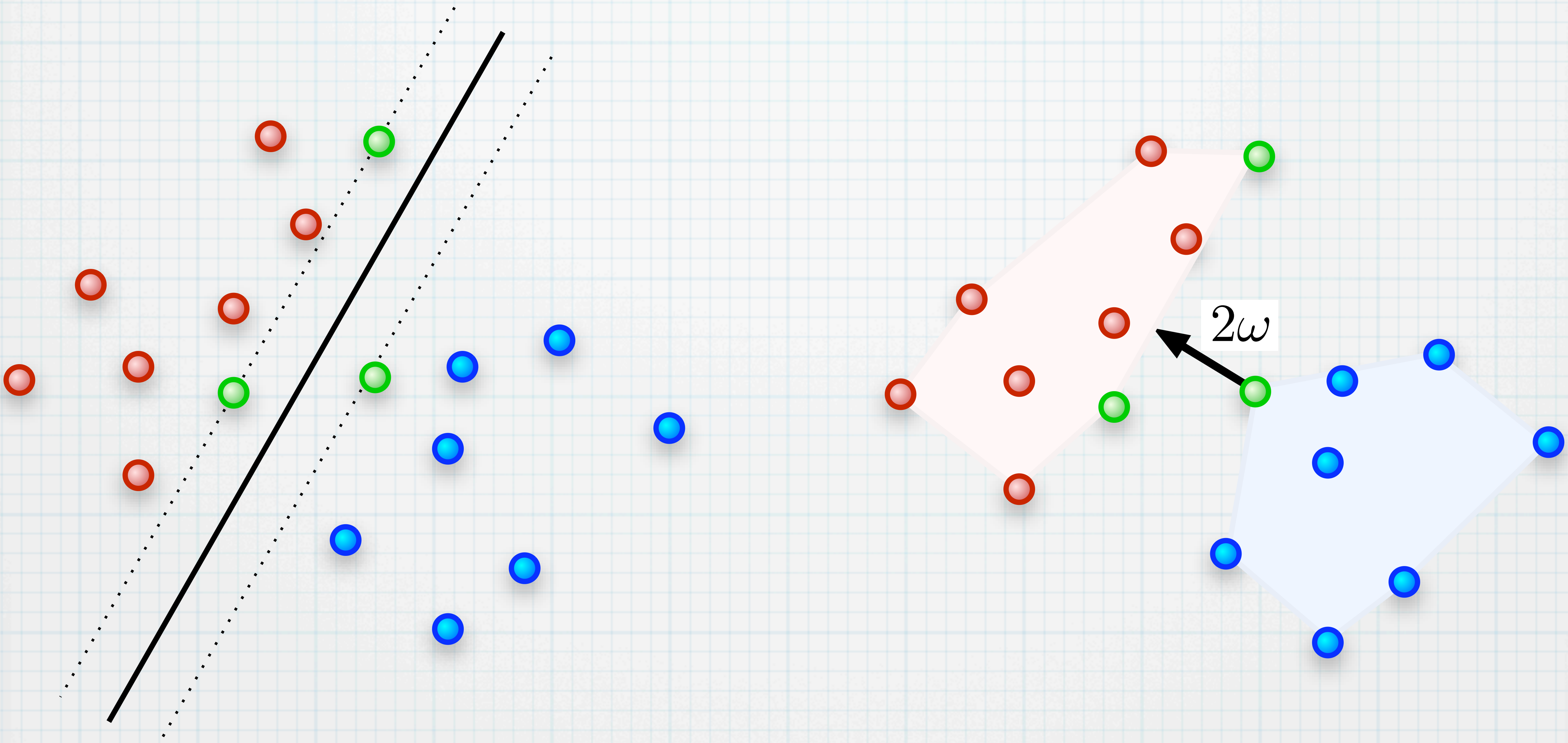
$$s.t. \qquad \alpha_i \geq 0 \qquad \forall i$$

$$\sum_{I_+} \alpha_i = 1$$

$$\sum_{I_-} \alpha_i = 1$$

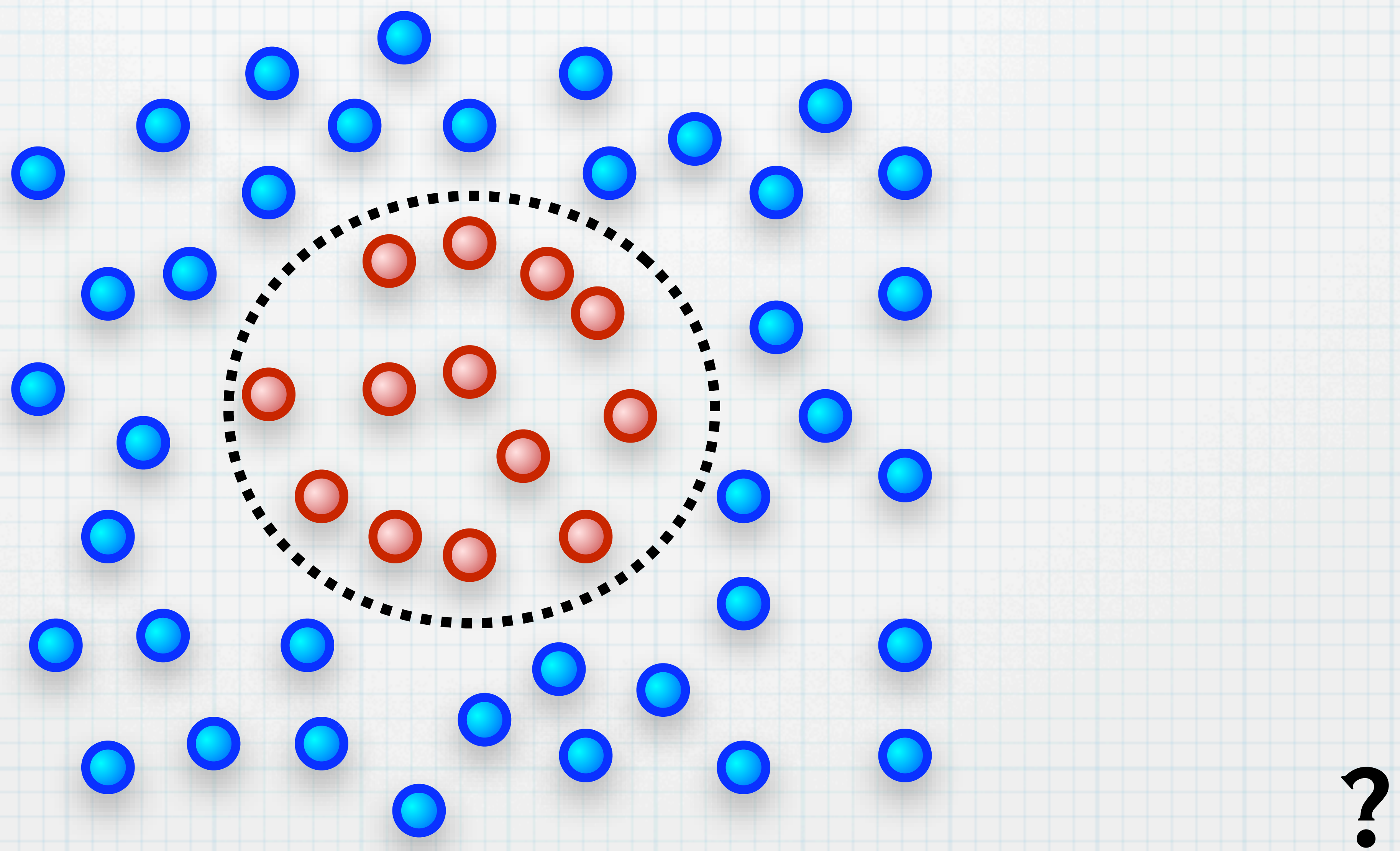
Geometric Interpretation:

$$\omega = \frac{1}{2}\sum_i \alpha_i y_i x_i$$



$$\begin{array}{lll} \min\limits_{\alpha} & \sum_{i,j} \alpha_i\alpha_j\, y_i y_j\, x_i^T x_j & \\ s.t. & \alpha_i \geq 0 & \forall i \\ & \sum_{I_+} \alpha_i = 1 & \\ & \sum_{I_-} \alpha_i = 1 & \end{array}$$

Kernels



Kernels

Idea: Before doing linear separation, map all points to some higher-dimensional space:

$$x_i \longmapsto \Phi(x_i)$$

Kernels...

Effect of the mapping $x_i \longmapsto \Phi(x_i)$

to the optimisation problem:

$$K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$$

$$\min_{\alpha} \quad \textstyle\sum_{i,j} \alpha_i \alpha_j \, y_i y_j \, K(x_i, x_j)$$

$$s.t. \quad \alpha_i \geq 0 \qquad \forall i$$

$$\textstyle\sum_{I_+} \alpha_i = 1$$

$$\textstyle\sum_{I_-} \alpha_i = 1$$

Decision function can also be evaluated by only using the kernel:

$$\omega = \tfrac{1}{2} \textstyle\sum_i \alpha_i y_i \Phi(x_i)$$

$$\omega^T \Phi(x_k) + b \quad = \tfrac{1}{2} \textstyle\sum_i \alpha_i y_i K(x_i, x_k) + b$$

Kernel Trick

What Kernel functions $K(.,.)$ are we allowed to choose?

Mercer's Theorem (1909)

If a function $K(.,.)$ is *continous*, *symmetric* and *positive semi-definite*, then there exists a corresponding mapping $\Phi(.)$ s.t.

$$K(x,y) = \langle \Phi(x), \Phi(y) \rangle \quad \forall x, y \in \mathbb{R}^n$$

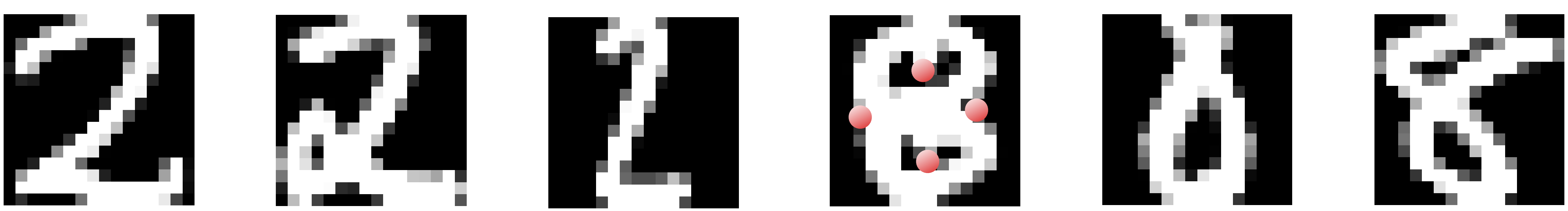
Popular Kernels

Polynomial Kernel

$$K(x,y) = \langle x, y\rangle^d, \quad d \in \mathbb{N}$$

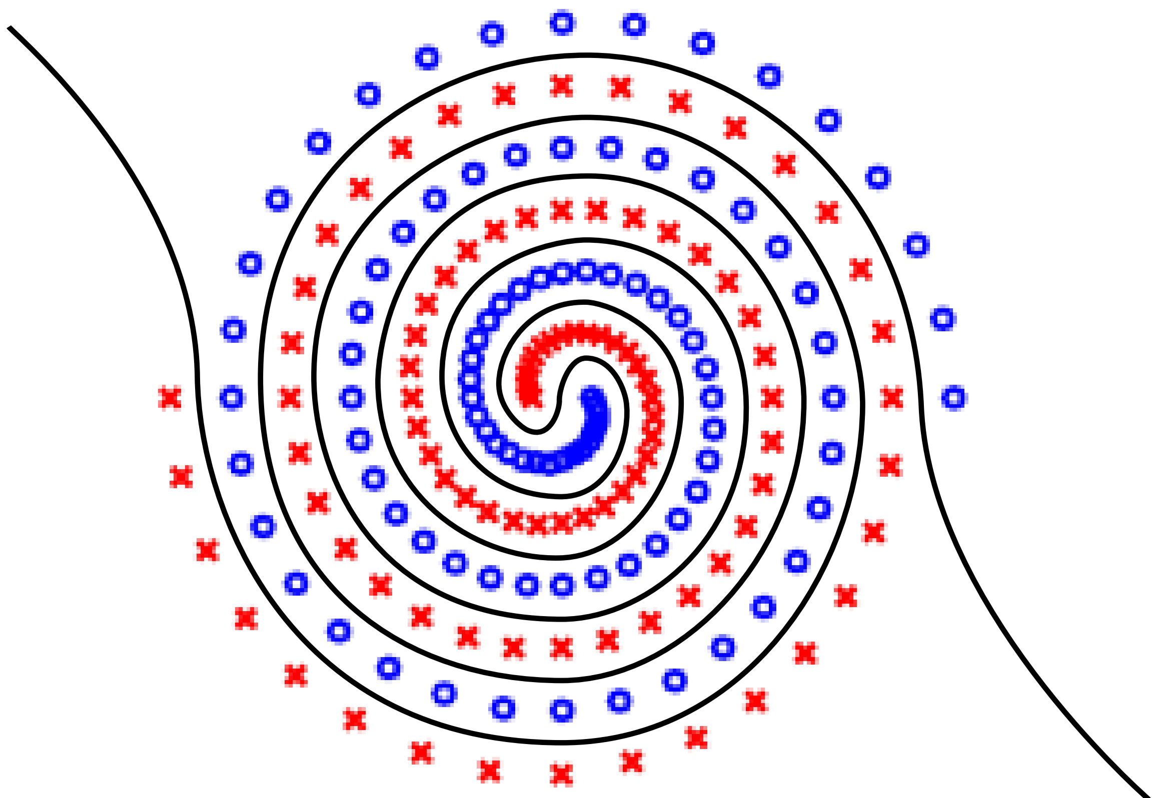
$$\langle x, y\rangle^d = \left(\sum_i x_i y_i\right)\left(\sum_i x_i y_i\right)\ldots\left(\sum_i x_i y_i\right) = \langle (x_{i_1} x_{i_2} \ldots x_{i_d}), (y_{i_1} y_{i_2} \ldots y_{i_d})\rangle$$

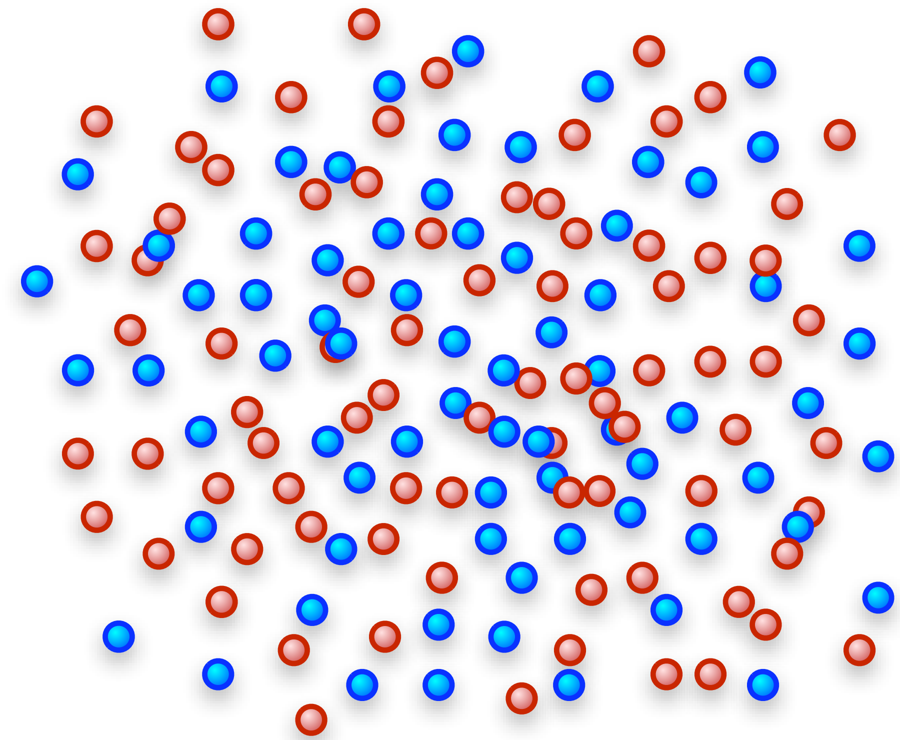
Usage in Handwritten Digit Recognition



Gauss Kernel

$$K(x, y) = e^{-\|x-y\|^2/2\sigma}$$

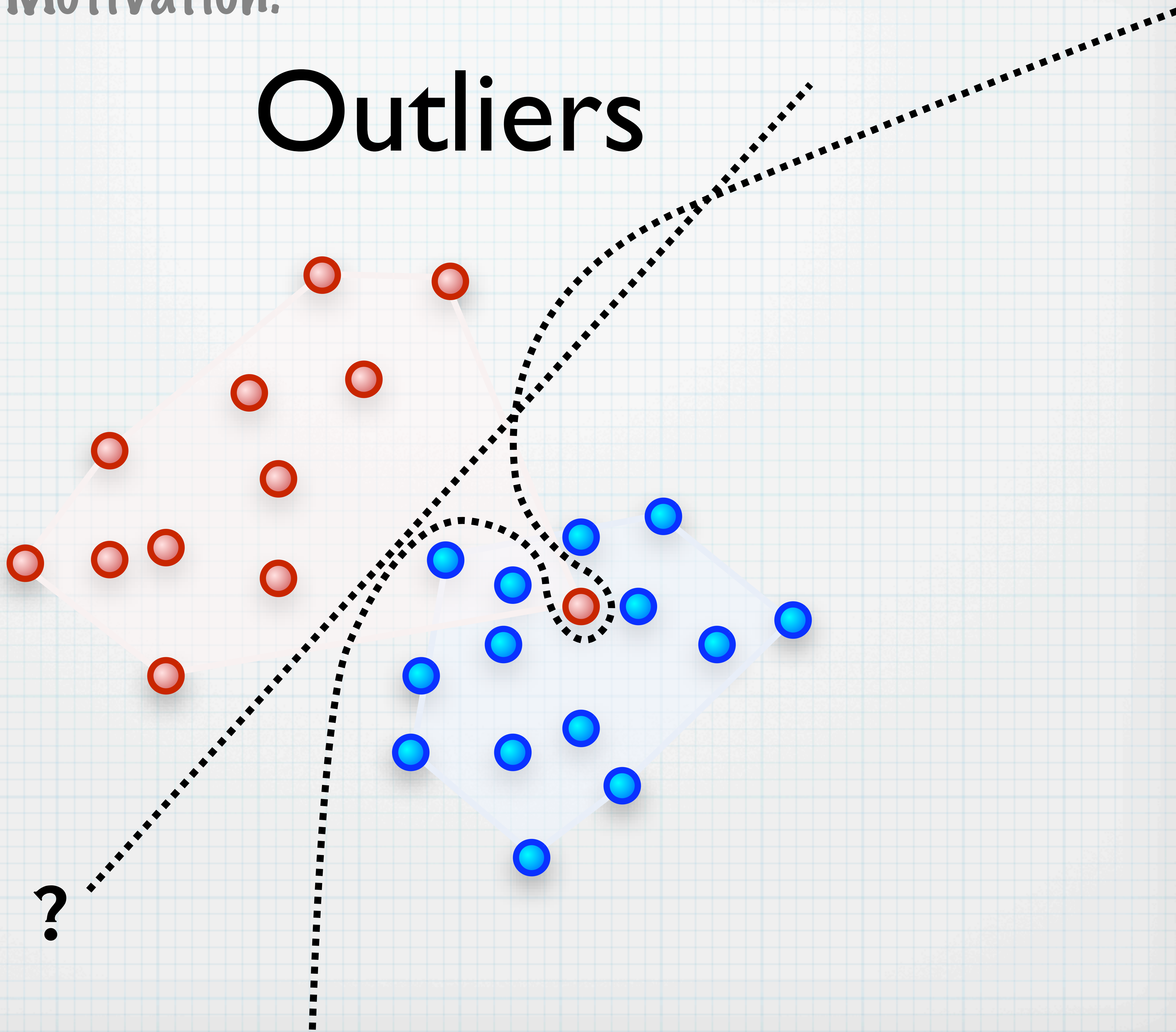




String Kernels

b l a   b l a h   a g t a a t c a t c g c t a

Outliers



Soft margin SVM

How can we allow outliers?

PRIMAL

$$\min_{\omega,b,\rho} \quad \tfrac{1}{2}\|\omega\|^2 - 2\rho$$

$$s.t. \quad y_i(\omega^T x_i + b) \geq \rho \qquad \forall i$$

DUAL

$$\min_{\alpha} \quad \|\textstyle\sum_{I_+} \alpha_i x_i - \sum_{I_-} \alpha_i x_i\|^2$$

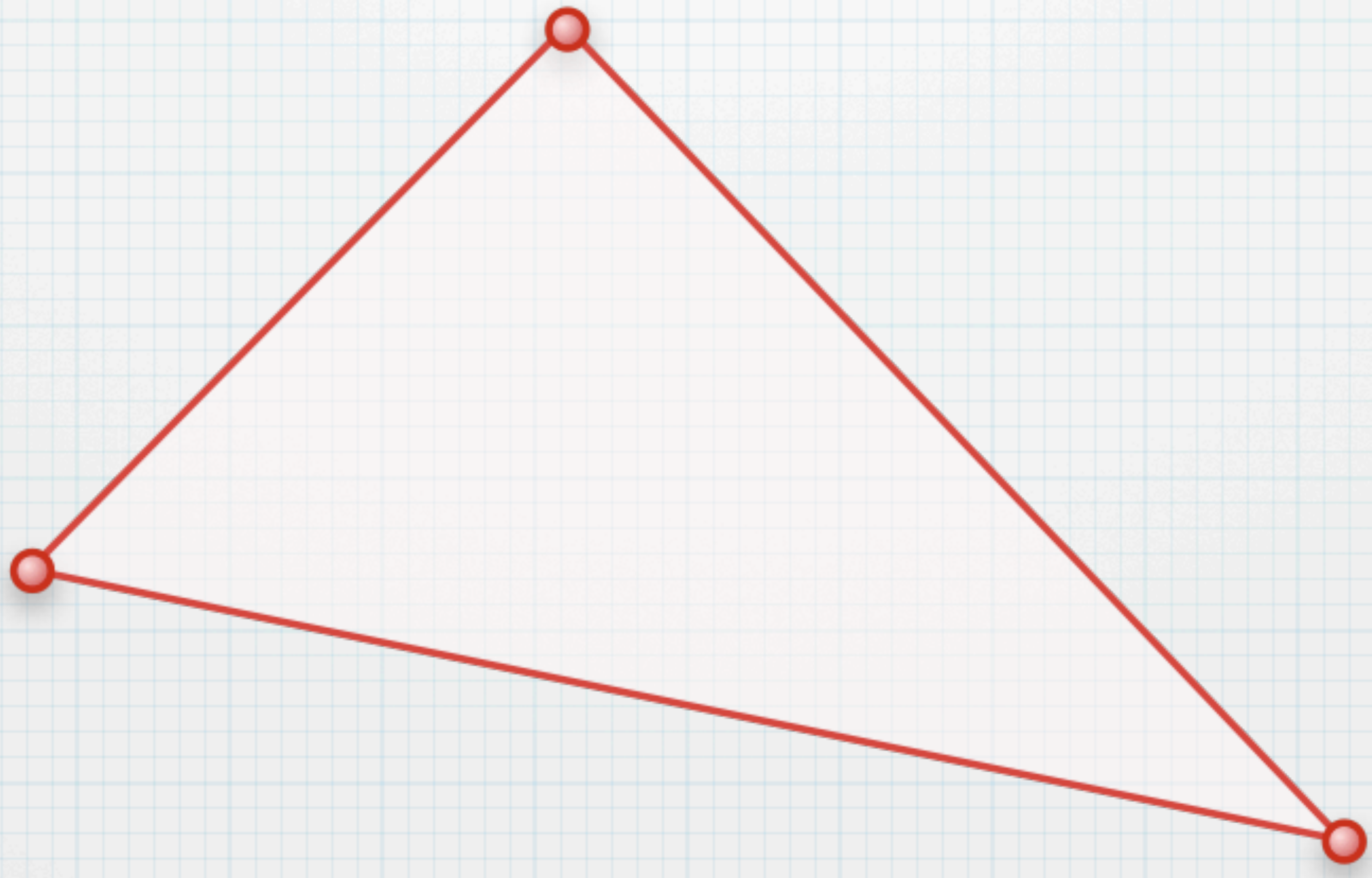
$$s.t. \quad 0 \leq \alpha_i \qquad \forall i$$

$$\textstyle\sum_{I_+} \alpha_i = 1$$

$$\textstyle\sum_{I_-} \alpha_i = 1$$

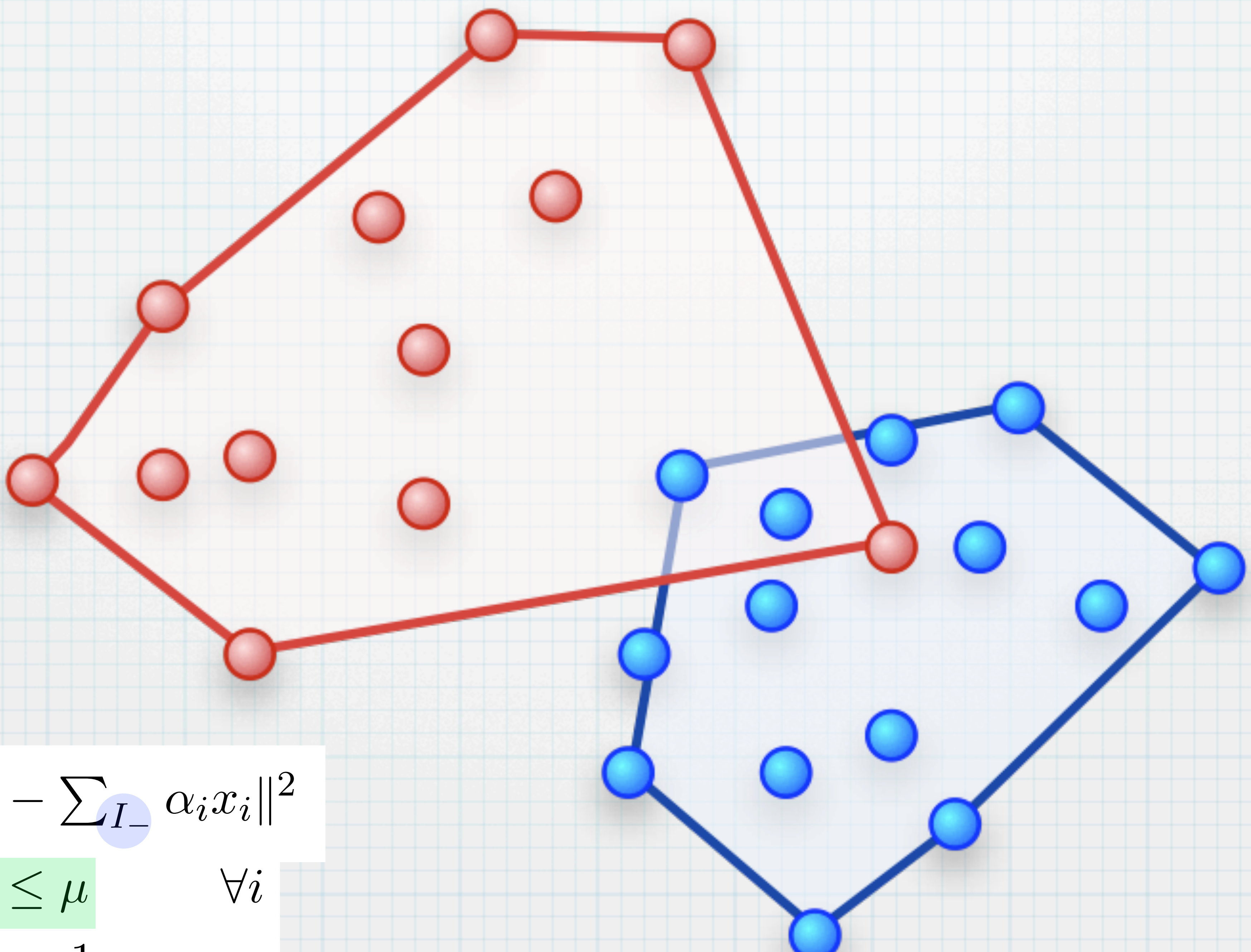
Reduced Convex Hulls

$$RCH(X,\mu) := \left\{ \sum_i \alpha_i x_i \;\middle|\; 0 \le \alpha_i \le \mu \;\; \forall i,\; \sum_i \alpha_i = 1 \right\}$$



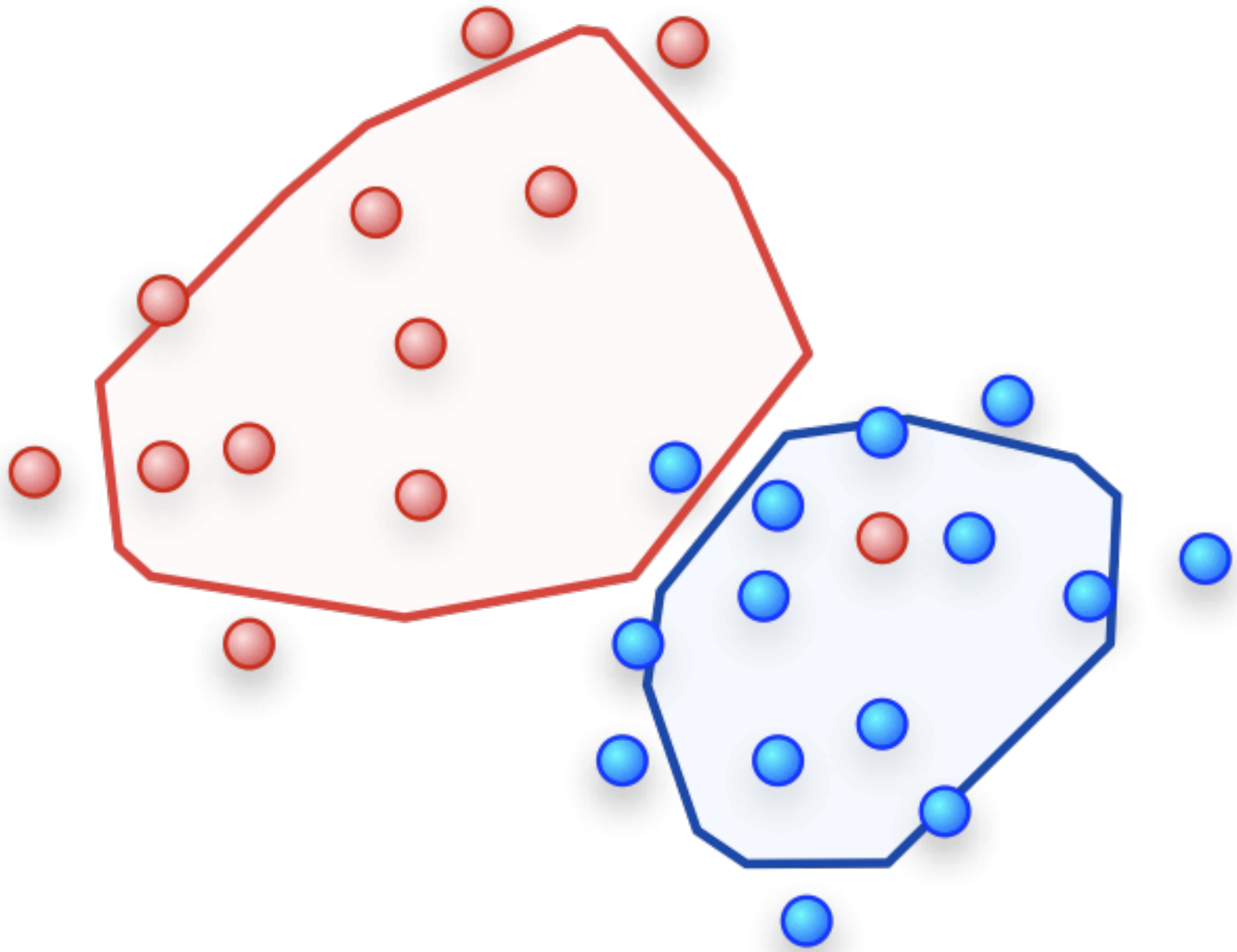
Geometric Interpretation:

Distance between reduced convex hulls



$$\begin{array}{lll} \min\limits_{\alpha} & \|\sum_{I_+} \alpha_i x_i - \sum_{I_-} \alpha_i x_i\|^2 & \\ s.t. & 0 \leq \alpha_i \leq \mu & \forall i \\ & \sum_{I_+} \alpha_i = 1 & \\ & \sum_{I_-} \alpha_i = 1 & \end{array}$$

Path of the solution?



Solution path is piece-wise linear (Hastie et al. 2004)

Algorithms

- Standard QP solvers are too slow for large problems
- Currently used algorithms are *approximation methods* trying to use the special structure of the QP (e.g. SMO by Platt, 1999)
- Very few exact bounds are known on the *speed* as well as *approximation quality*

A geometric SVM Algorithm (2006)

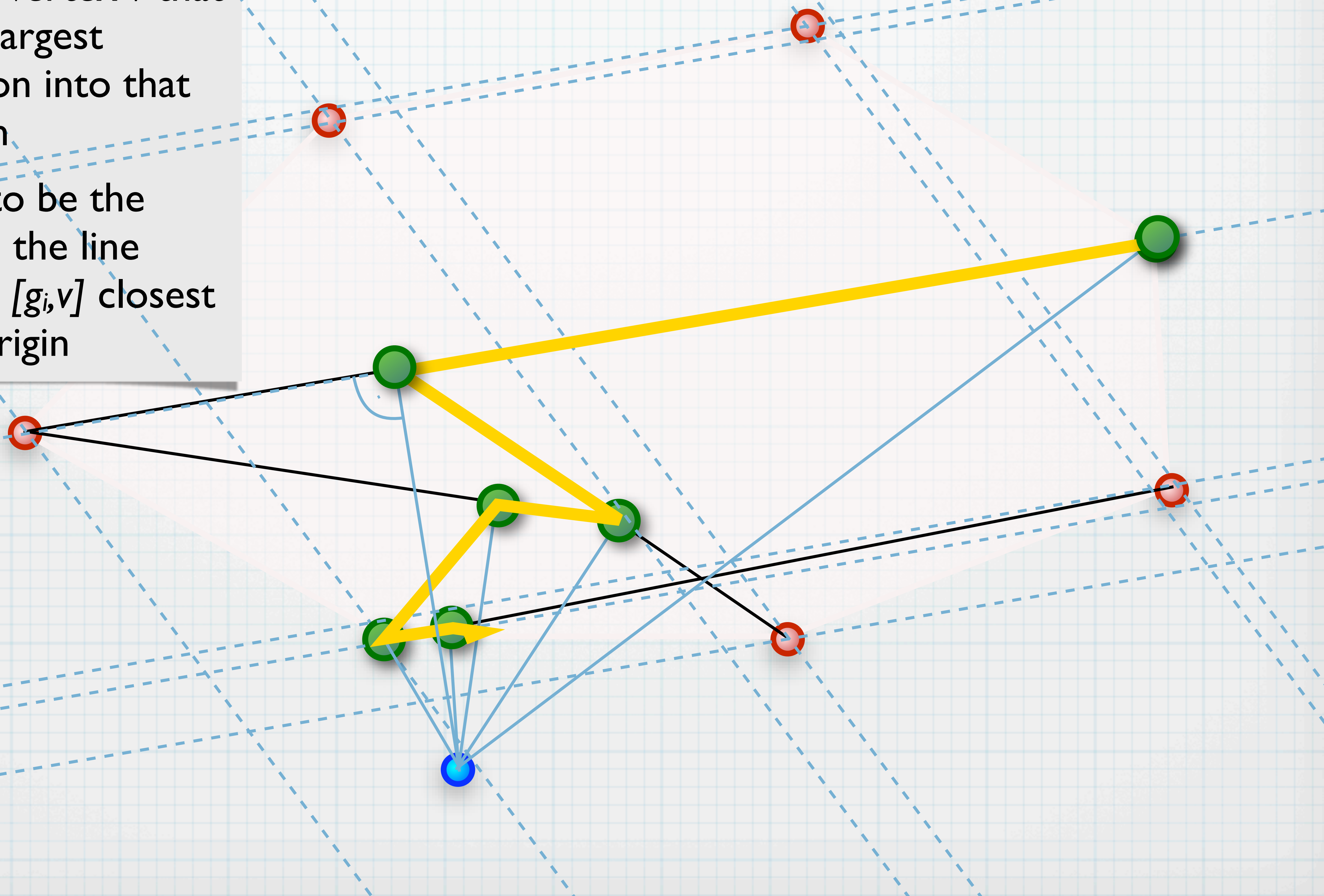
Gilbert‘s Algorithm for Polytope Distance (1966)

A specialised, more widely know variant for 3D:

Gilbert-Johnson-Keerthi (GJK) Algorithm
(used for collision detection)

Geometric Interpretation:

- At g_i: Look at the direction to the origin
- Find the vertex v that has the largest projection into that direction
- Set g_{i+1} to be the point on the line segment $[g_i,v]$ closest to the origin



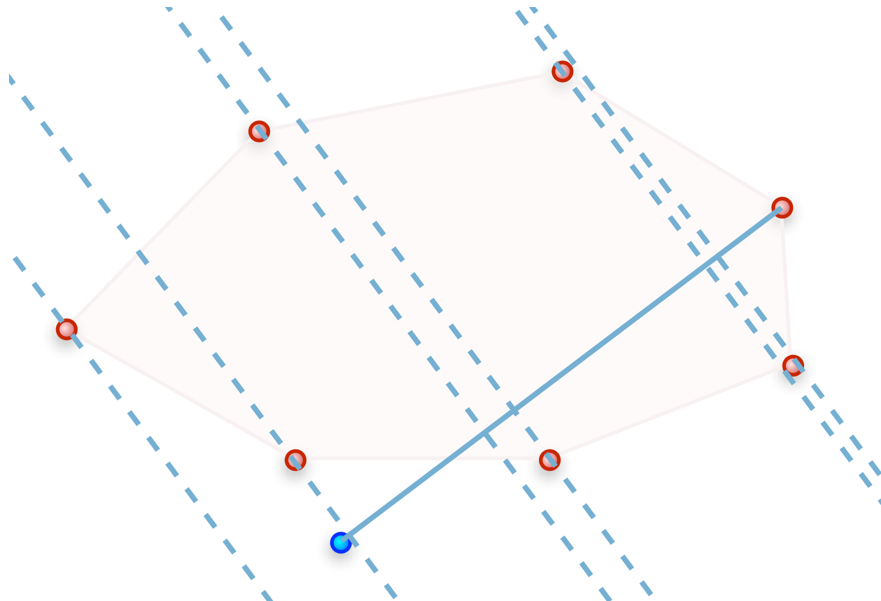
Gilbert‘s Algorithm

- Does it work for *reduced convex hulls* as well?
  - yes, the longest projection of a reduced convex hull in a given direction can be calculated fast (without having to deal with its exponentially many vertices)

If $(i_1, \ldots, i_m)$ is an decreasing ordering of the projections $\langle x_i, p \rangle$ of the points x_i along the direction p (with $\|p\| = 1$), then the extreme point of the $RCH(X, \mu)$ with largest projection in direction p is given by

$$\mu \sum_{j=1}^{m} x_{i_j} + (1 - m\mu) x_{i_{m+1}}$$

where $m = \lfloor 1/\mu \rfloor$



LP-type problems

Random sampling was successfully used to give the first provably fast randomised algorithm for SVMs.

(2001)

Disadvantage: not a good bound for high dimensions

Core Vector Machines

(2005)

Translation to the Smallest Enclosing Ball problem

$$\begin{array}{llr} \min\limits_{\alpha} & \alpha^T K \alpha - diag(K)\alpha & \\ s.t. & \alpha_i \geq 0 & \forall i \\ & \sum_i \alpha_i = 1 & \end{array}$$

Advantage: very fast *core set algorithms* exist

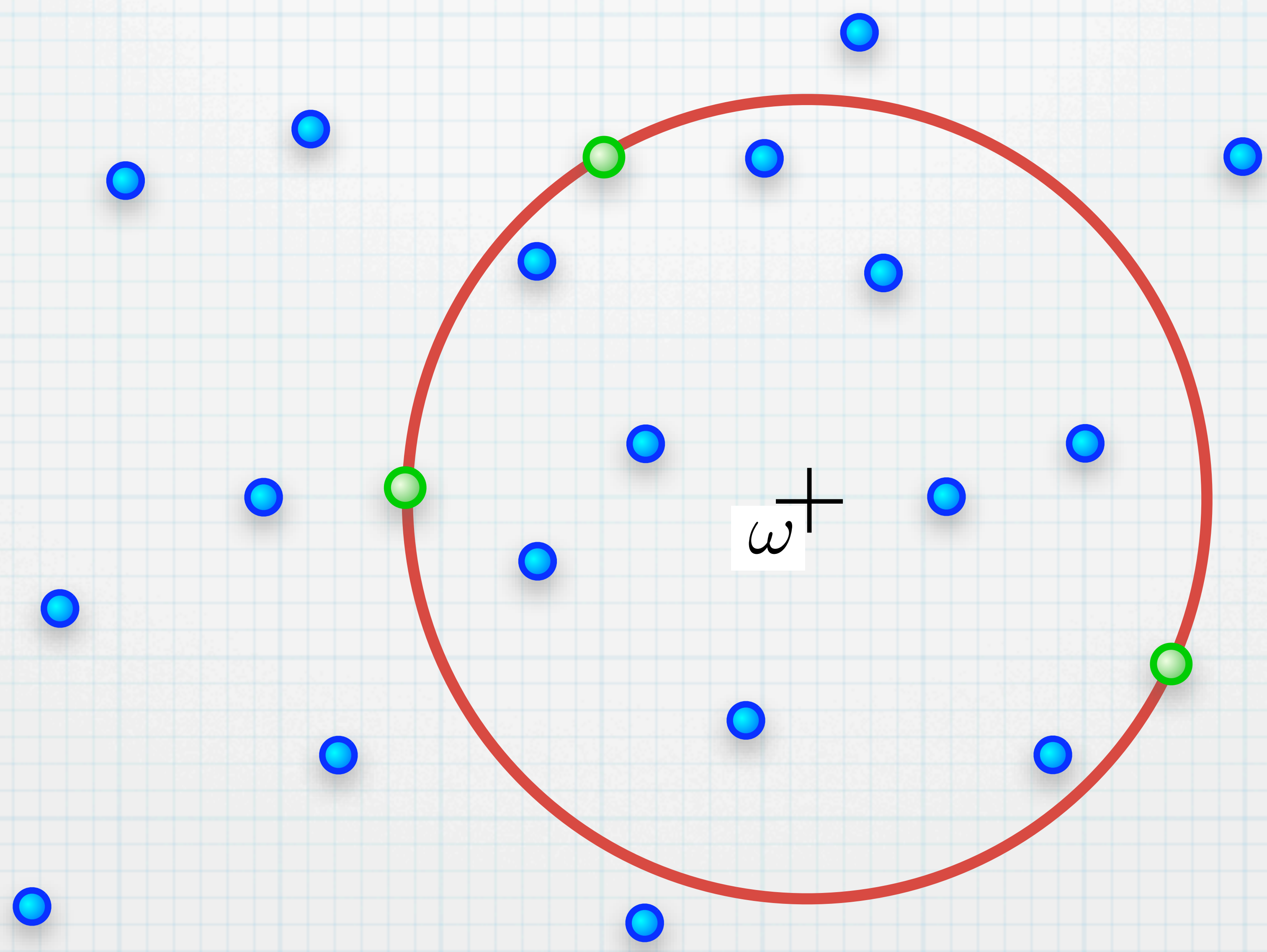
$$\begin{array}{llr} \min\limits_{\alpha} & \alpha^T K \alpha & \\ s.t. & \alpha_i \geq 0 & \forall i \\ & \sum_{I_+} \alpha_i = 1 & \\ & \sum_{I_-} \alpha_i = 1 & \end{array}$$

provable speed and

provable approximation quality

Geometric Interpretation:

Smallest Enclosing Ball



Thanks