Connections between the Lasso and Support Vector Machines

Martin Jaggi

Ecole Polytechnique

2013 / 07 / 08

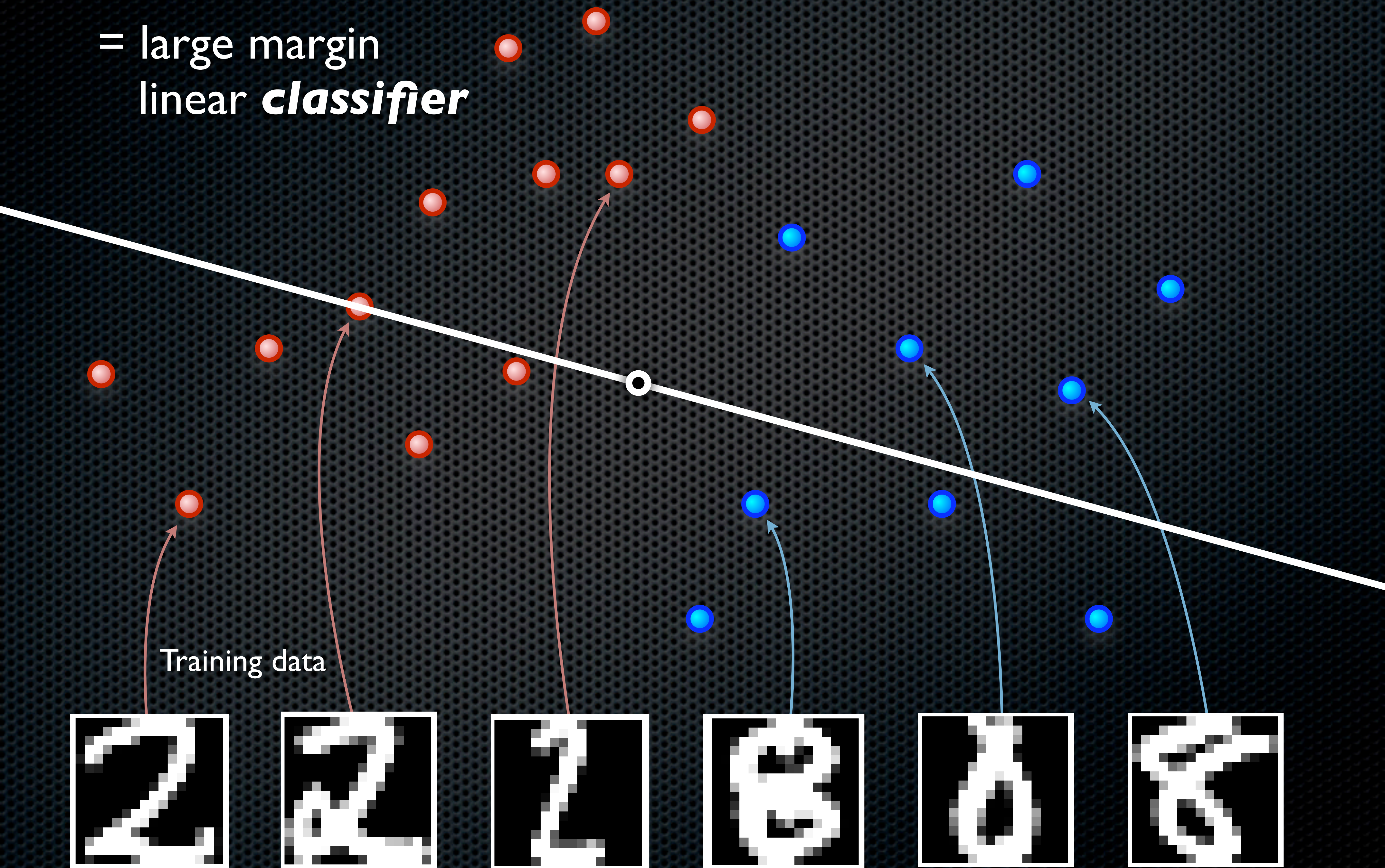
ROKS '13 - International Workshop on Advances in Regularization, Optimization, Kernel Methods and Support Vector Machines: Theory and Applications

Outline

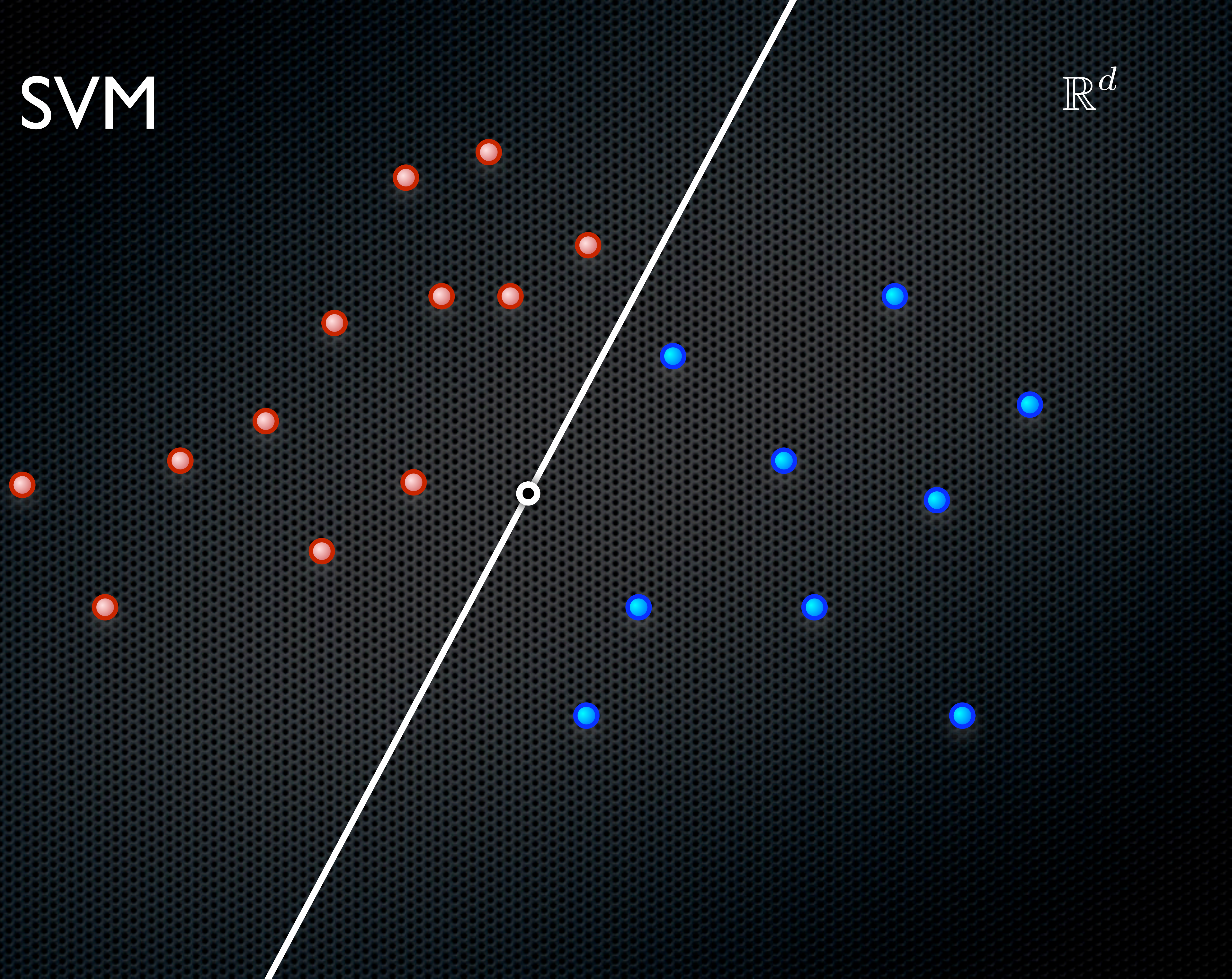
- An Equivalence between the *Lasso* and *Support Vector Machines*
  - Reduction from Lasso to SVM
  - Reduction from SVM to Lasso
  - Applications
- Greedy Algorithms (from optimization and signal processing)

SVM

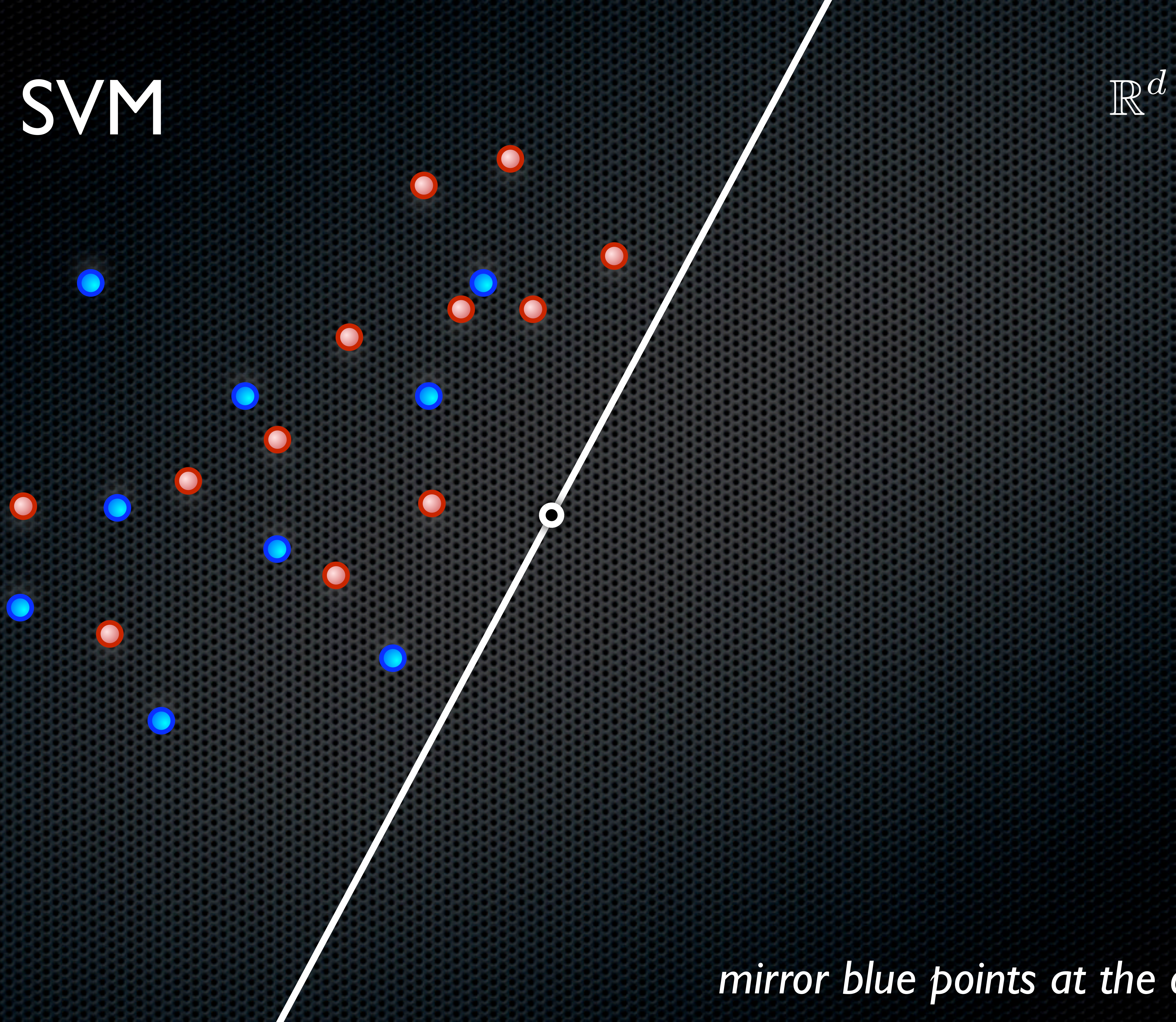
= large margin
linear ***classifier***



SVM



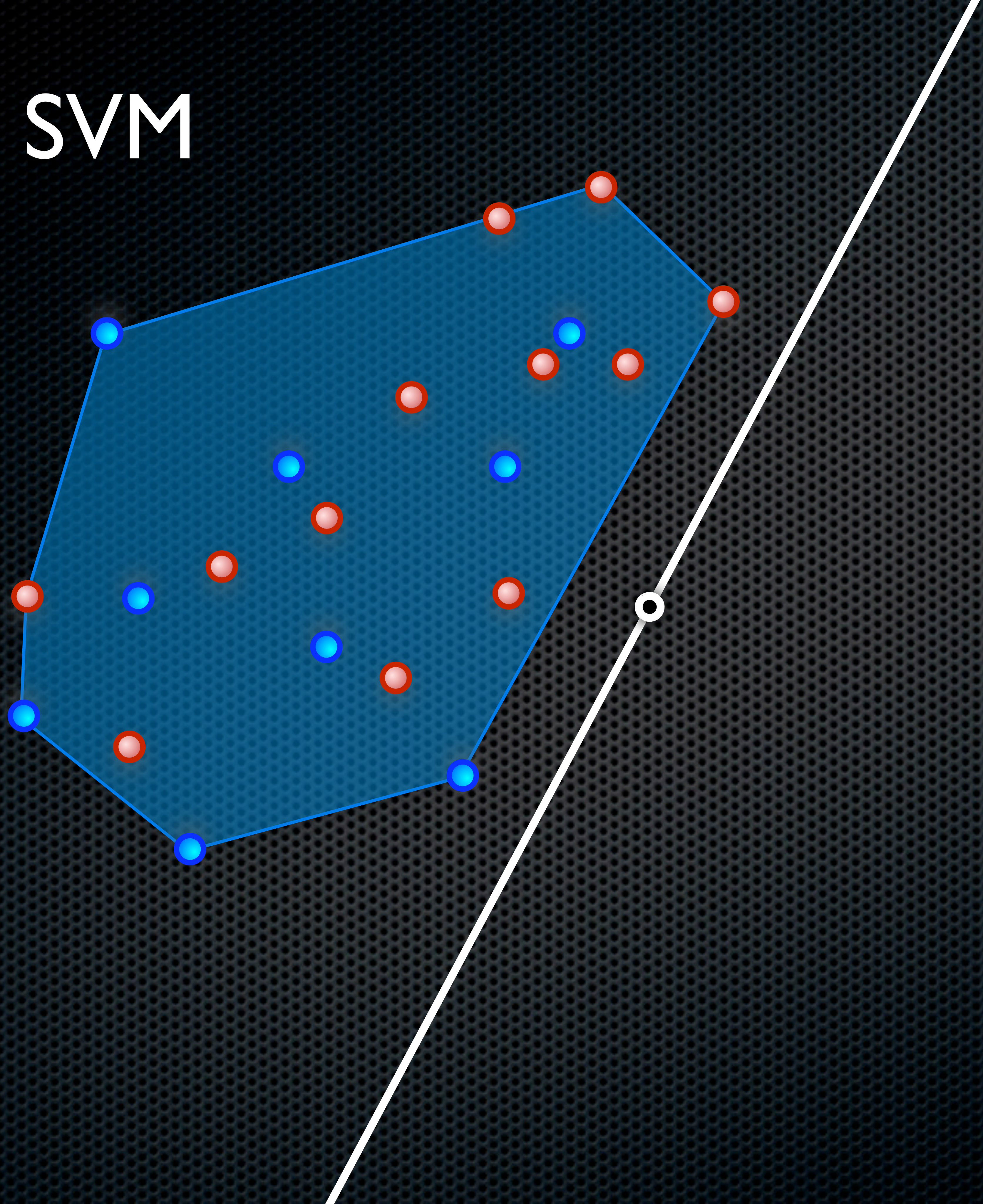
SVM



mirror blue points at the origin

SVM

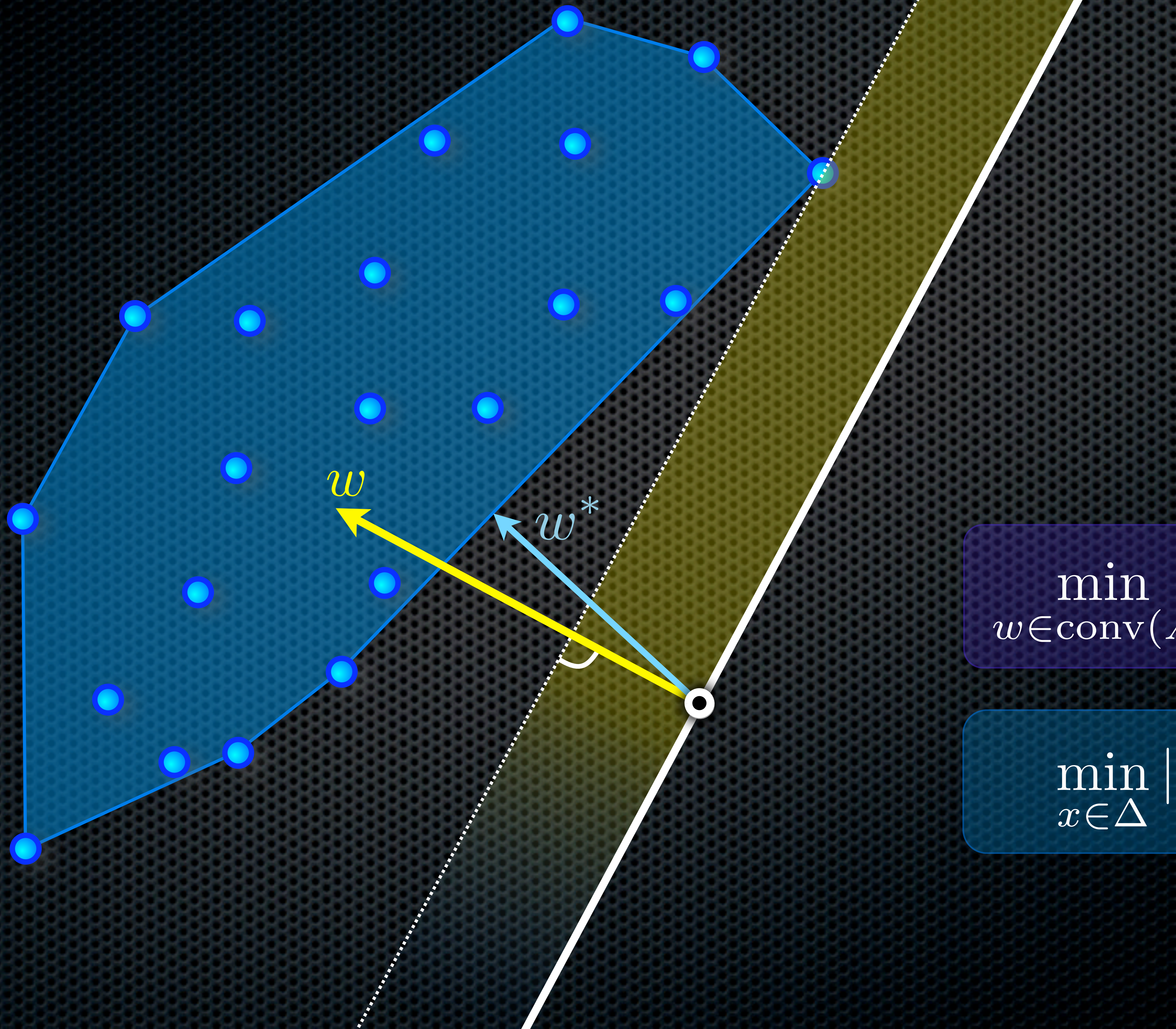
$\mathbb{R}^d$



Polytope distance

n points in $\mathbb{R}^d$

$A \in \mathbb{R}^{d \times n}$



$$\min_{w \in \text{conv}(A)} \|w\|^2$$

$$\min_{x \in \Delta} \|Ax\|^2$$

$A \in \mathbb{R}^{d \times n}$

SVM variants

whose *dual* problem is of the form $\min_{x \in \Delta} \|Ax\|^2$

	Hard margin	***Soft margin (L2-loss)***	***Soft margin (L1-loss)***
Two class *no offset/bias*	✓	✓	✗
Two class *regularized offset/bias*	✓	✓	✗
One Class	✓	✓	✗

(all with or without using ***kernels***)

$\|Ax\|^2 = x^T A^T A x$

$$\min_{\substack{\bar{w} \in \mathbb{R}^d,\ \rho \in \mathbb{R}, \\ \xi \in \mathbb{R}^n}} \quad \tfrac{1}{2}\|\bar{w}\|_2^2 - \rho + \tfrac{C}{2}\sum_i \xi_i^2$$

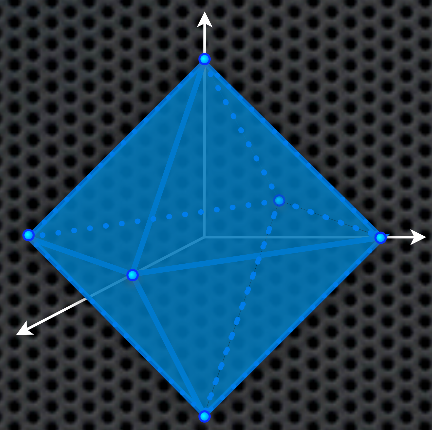
$$s.t. \quad y_i \cdot \bar{w}^T X_i \geq \rho - \xi_i \quad \forall i \in [1..n]$$

$A \in \mathbb{R}^{d \times n}$
$b \in \mathbb{R}^d$

Lasso

= ℓ_1-regularized least squares ***regression***

$$\min_{\|x\|_1 \le t} \|Ax - b\|^2$$



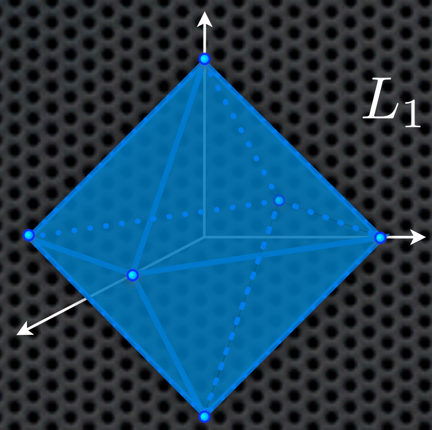
- Sparse regression
- Feature selection

$A \in \mathbb{R}^{d\times n}$
$b \in \mathbb{R}^{d}$

Lasso

= ℓ_1-regularized least squares ***regression***

$$\min_{x\in L_1} \ \|Ax-b\|^2$$



$$L_1 := \{x \in \mathbb{R}^n \mid \|x\|_1 \le 1\} = \mathrm{conv}(\{\pm \mathbf{e}_i\})$$

- Sparse regression
- Feature selection

(Lasso $\preceq$ SVM)

$A \in \mathbb{R}^{d\times n}$
$b \in \mathbb{R}^{d}$

Given a Lasso $\min_{x\in L_1} \|Ax - b\|^2$

construct an equivalent SVM instance $\min_{x'\in\Delta} \|\tilde{A}x'\|^2$

$x = \begin{pmatrix} \mathbf{I}_n & -\mathbf{I}_n \end{pmatrix} x'$

$x \in L_1 \subset \mathbb{R}^n$

$x' \in \Delta \subset \mathbb{R}^{2n}$

(***barycentric coordinates***)

$$\min_{x'\in\Delta} \|A\begin{pmatrix} \mathbf{I}_n & -\mathbf{I}_n \end{pmatrix} x' - b\|^2$$

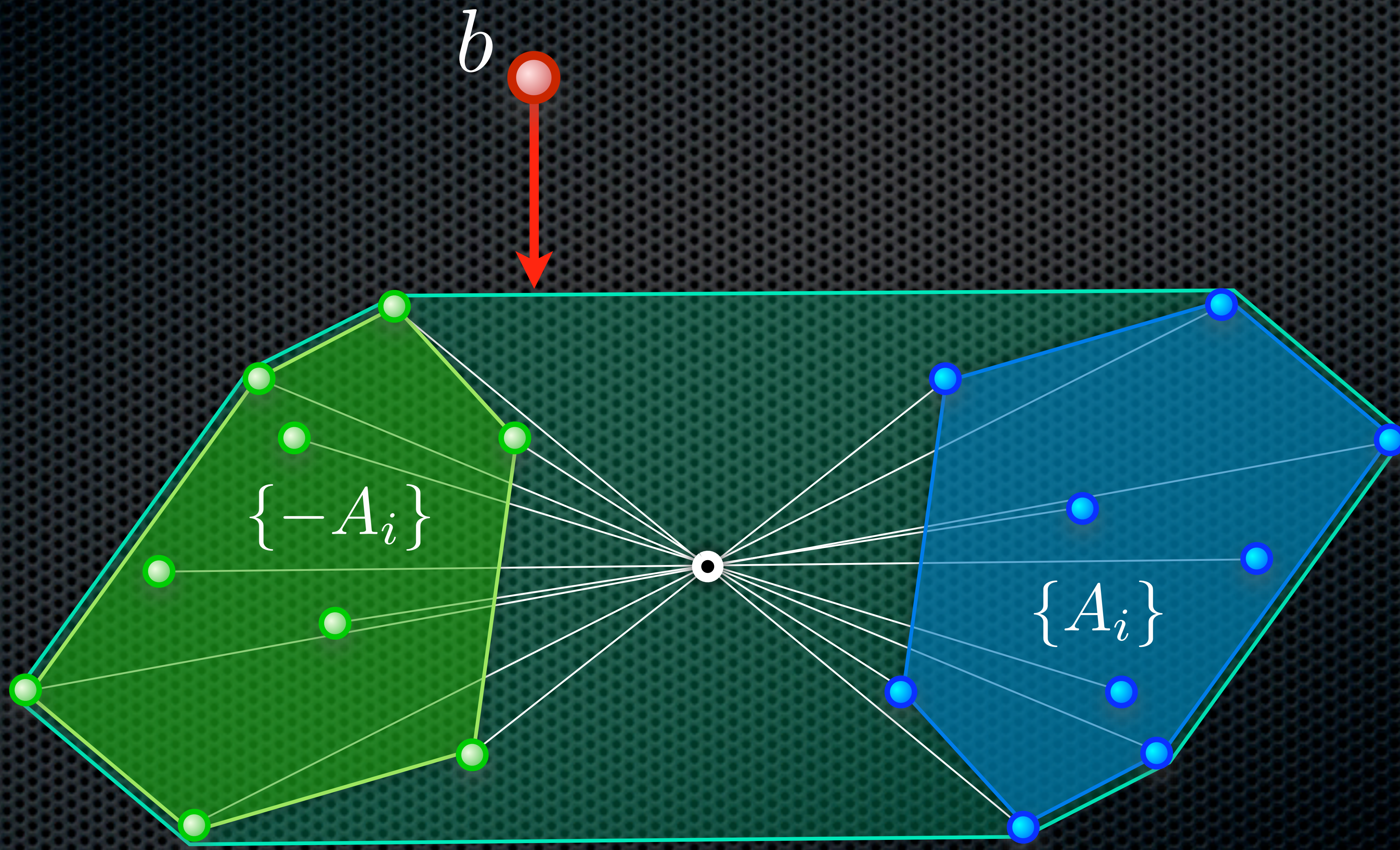
SVM:

$$\tilde{A} := \begin{pmatrix} A & -A \end{pmatrix} - b\mathbf{1}^T \in \mathbb{R}^{d\times 2n}$$

$(\text{Lasso} \preceq \text{SVM})$

Geometric interpretation:

$$\min_{x \in L_1} \|Ax - b\|^2$$



$A\,\mathrm{conv}(S) = \mathrm{conv}(AS)$

$$AL_1 = A\,\mathrm{conv}(\{\pm \mathbf{e}_i\}) = \mathrm{conv}(A\{\pm \mathbf{e}_i\}) = \mathrm{conv}(\{\pm A_i\})$$

$(\mathrm{SVM} \preceq \mathrm{Lasso})$

$A \in \mathbb{R}^{d \times n}$

Given an SVM $\min_{x \in \Delta} \|Ax\|^2$

construct an equivalent Lasso instance $\min_{x \in L_1} \|\tilde{A}x - \tilde{b}\|^2$

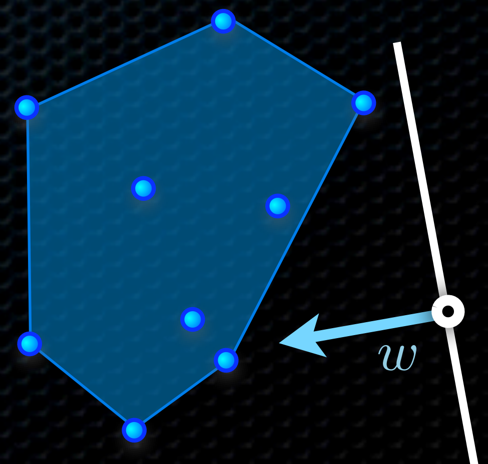
more challenging reduction!

Lasso:

$$\tilde{A} := A + \tilde{b}\mathbf{1}^T \quad \in \mathbb{R}^{d \times n}$$

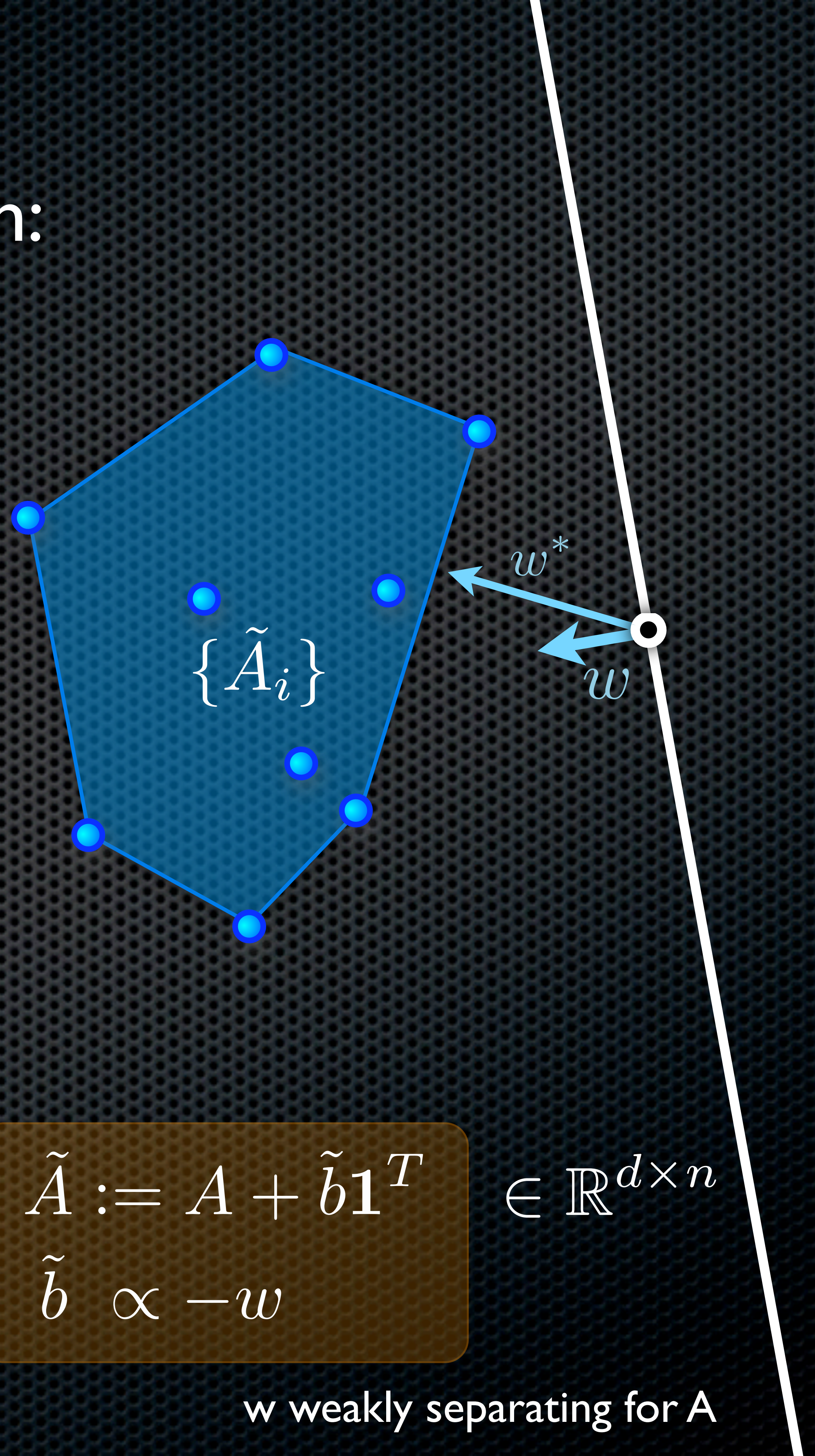
$$\tilde{b} \propto -w$$

w weakly separating for A



$(\text{SVM} \preceq \text{Lasso})$

Geometric interpretation:

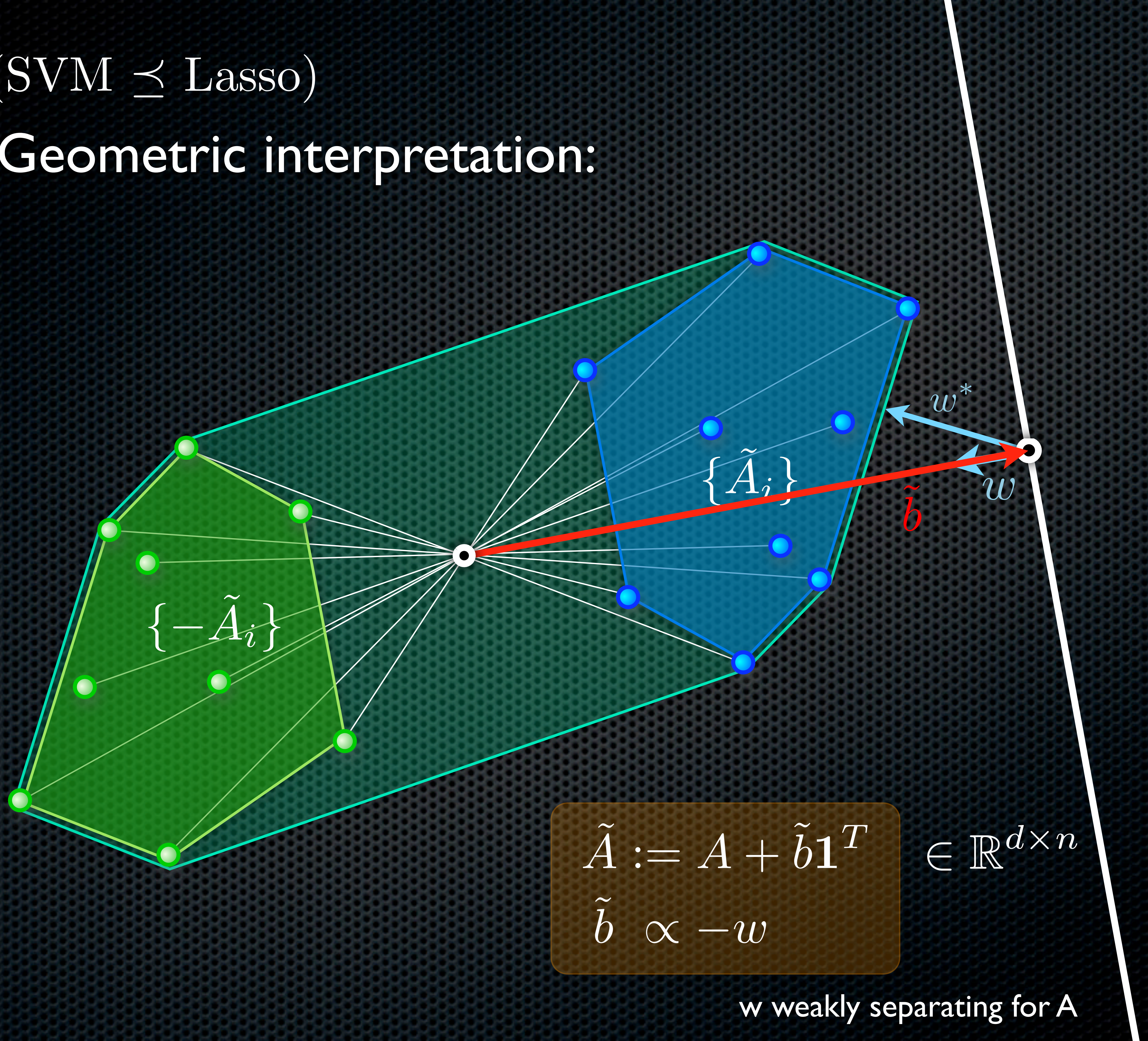


$$\tilde{A} := A + \tilde{b}\mathbf{1}^T \quad \in \mathbb{R}^{d \times n}$$
$$\tilde{b} \propto -w$$

w weakly separating for A

$(\text{SVM} \preceq \text{Lasso})$

Geometric interpretation:



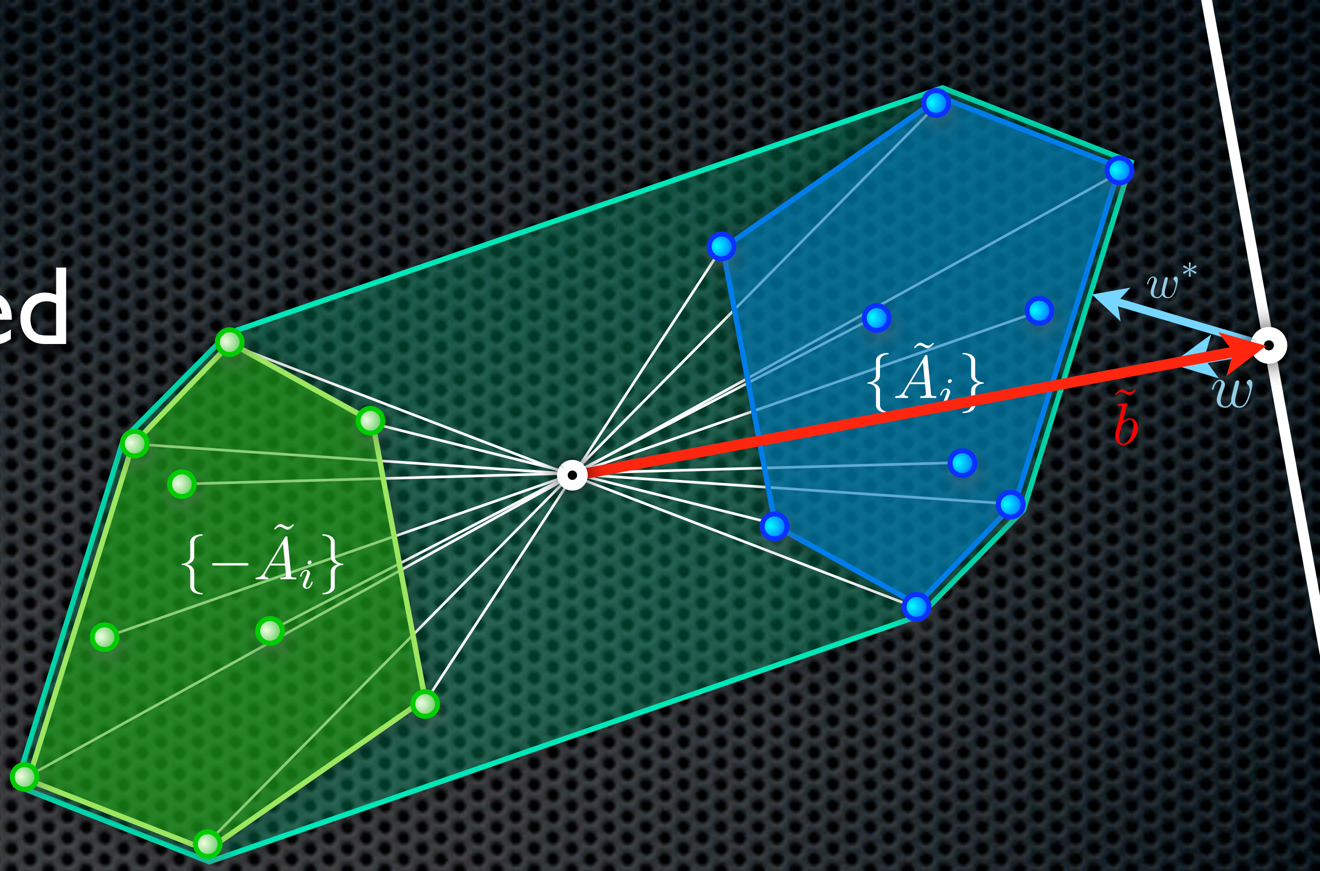
$$\tilde{A} := A + \tilde{b}\mathbf{1}^T \quad \in \mathbb{R}^{d \times n}$$
$$\tilde{b} \ \propto -w$$

w weakly separating for A

$(\text{SVM} \preceq \text{Lasso})$

Properties of the constructed Lasso instance

$$\min_{x \in L_1} \|\tilde{A}x - \tilde{b}\|^2$$



Theorem:

For any $x \in L_1$ for the Lasso, there is a vector $x' \in \Delta$, of the same or better Lasso objective.

This $x' \in \Delta$ attains the same objective in the SVM.

$$\tilde{A} := A + \tilde{b}\mathbf{1}^T \quad \in \mathbb{R}^{d \times n}$$
$$\tilde{b} \propto -w$$

w weakly separating for A

Implications:

- Algorithms apply to both problems

sublinear time algorithms $\tilde{O}(n+d)$

Implications for Lasso

- Kernelized version

$$\min_{x \in L_1} \left\| \sum_i \Psi(A_i) x_i - \Psi(b) \right\|_{\mathcal{H}}^2$$

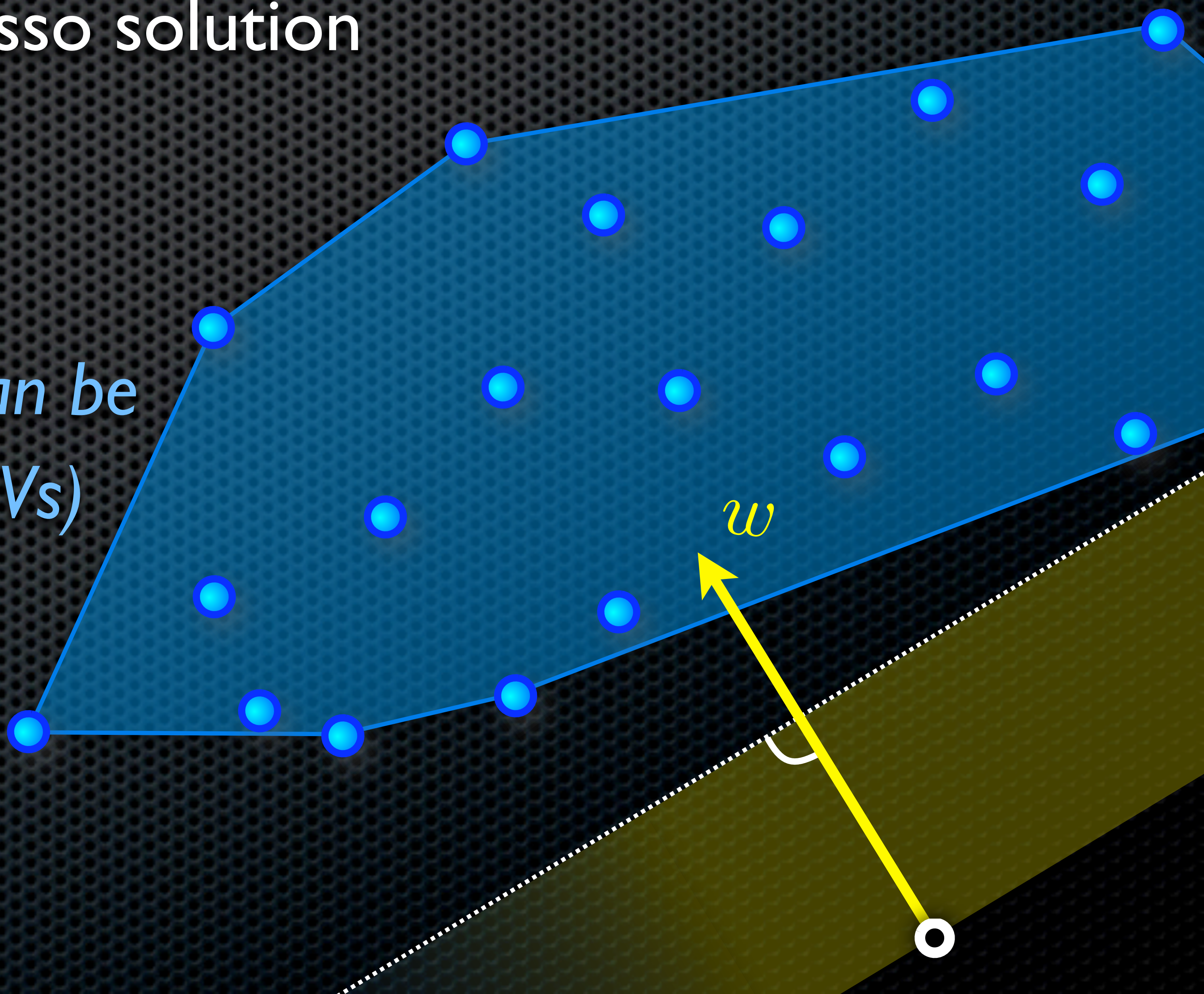
defined in terms of $\kappa(A_i, A_j),\ \kappa(A_i, b),\ \kappa(b, b)$ $\kappa(y, z) = \langle \Psi(y), \Psi(z) \rangle$

Implications for SVMs

- Support vectors
  = non-zeros in the Lasso solution
  - number of SVs

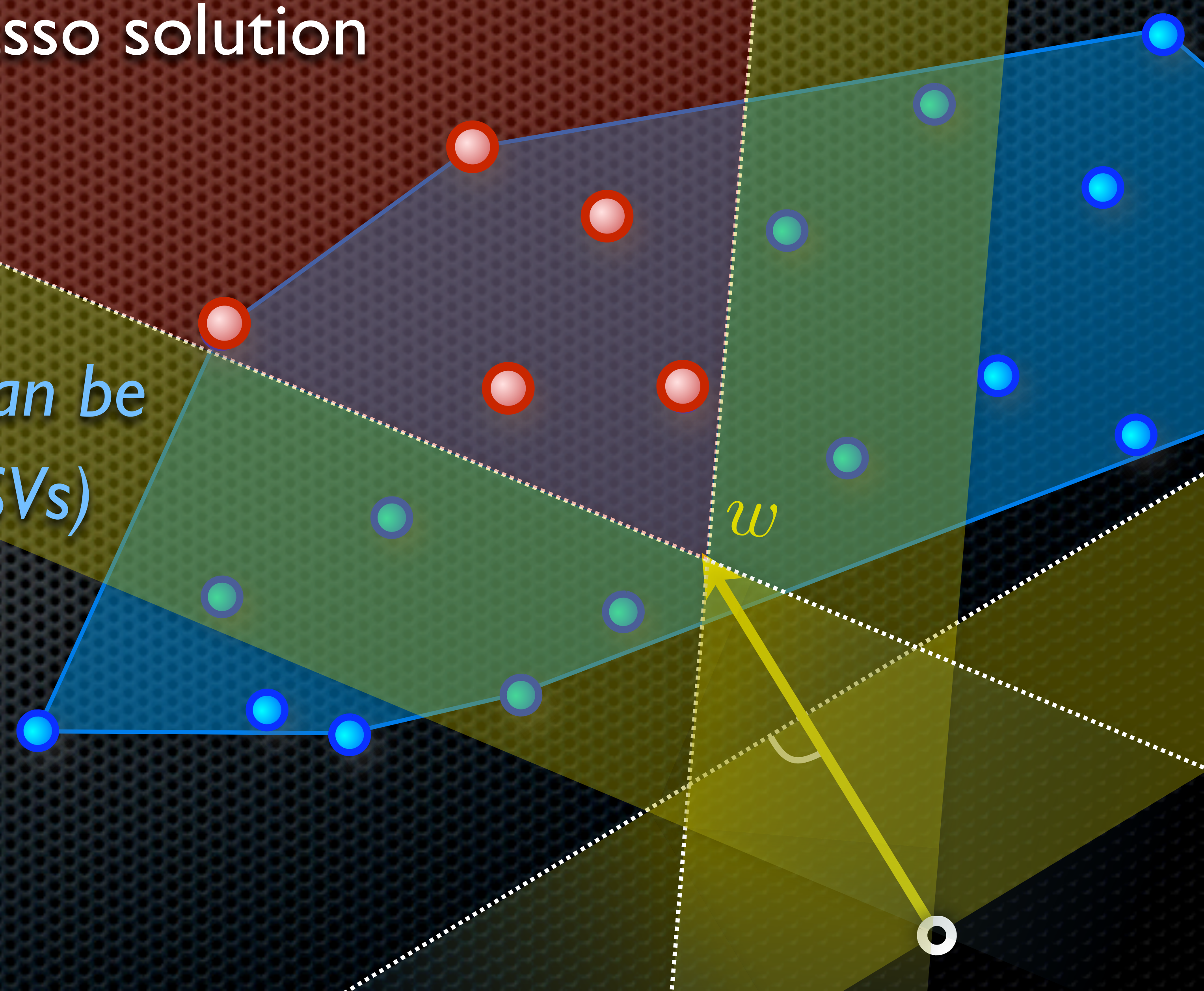
Implications for SVMs

- Support vectors
  = non-zeros in the Lasso solution
    - number of SVs
- Screening rules
  (discard points which can be guaranteed to be non-SVs)



Implications for SVMs

- Support vectors
  = non-zeros in the Lasso solution
  - number of SVs
- Screening rules
  (discard points which can be guaranteed to be non-SVs)



Greedy Algorithms

Convex optimization

methods applied to

$$\min_{x \in L_1} \|Ax - b\|^2$$

Frank-Wolfe

Signal processing

sparse recovery methods

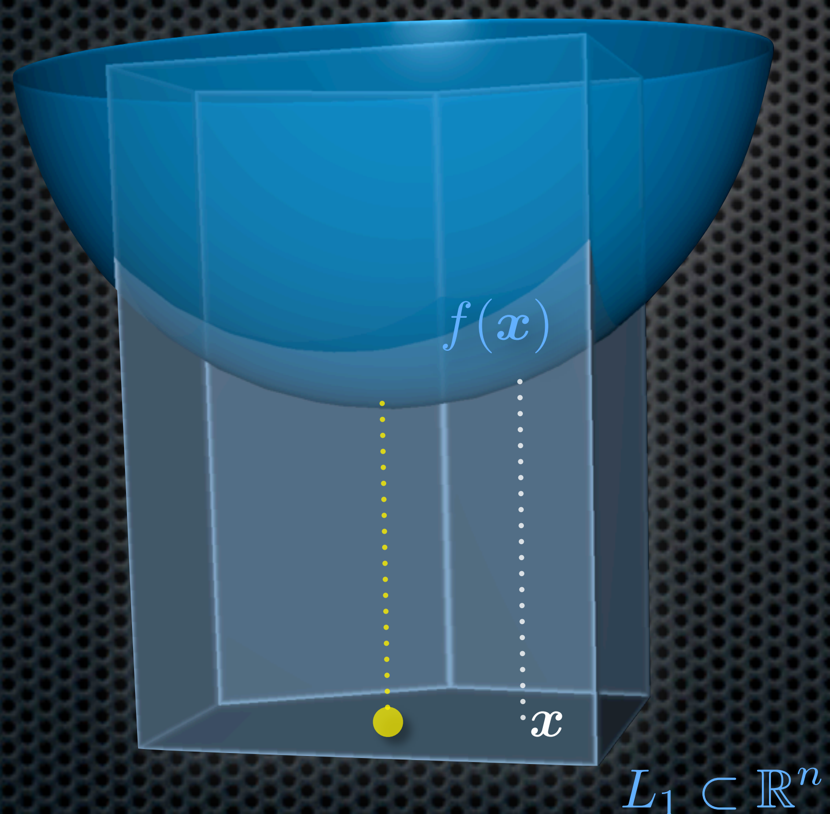
recover a sparse *x* from a noisy measurement *b* of *Ax*

Greedy Algorithms

Convex optimization

methods applied to

$$\min_{x \in L_1} \|Ax - b\|^2$$



Signal processing

sparse recovery methods

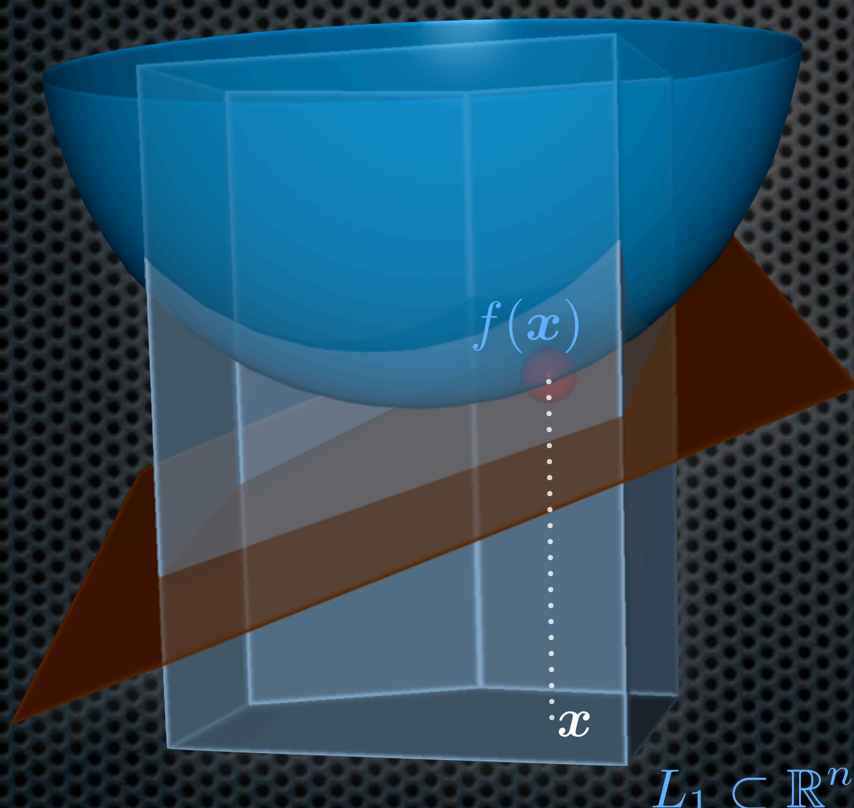
recover a sparse *x* from a noisy measurement *b* of *Ax*

Greedy Algorithms

Convex optimization

methods applied to

$$\min_{x \in L_1} \|Ax - b\|^2$$



Signal processing

sparse recovery methods

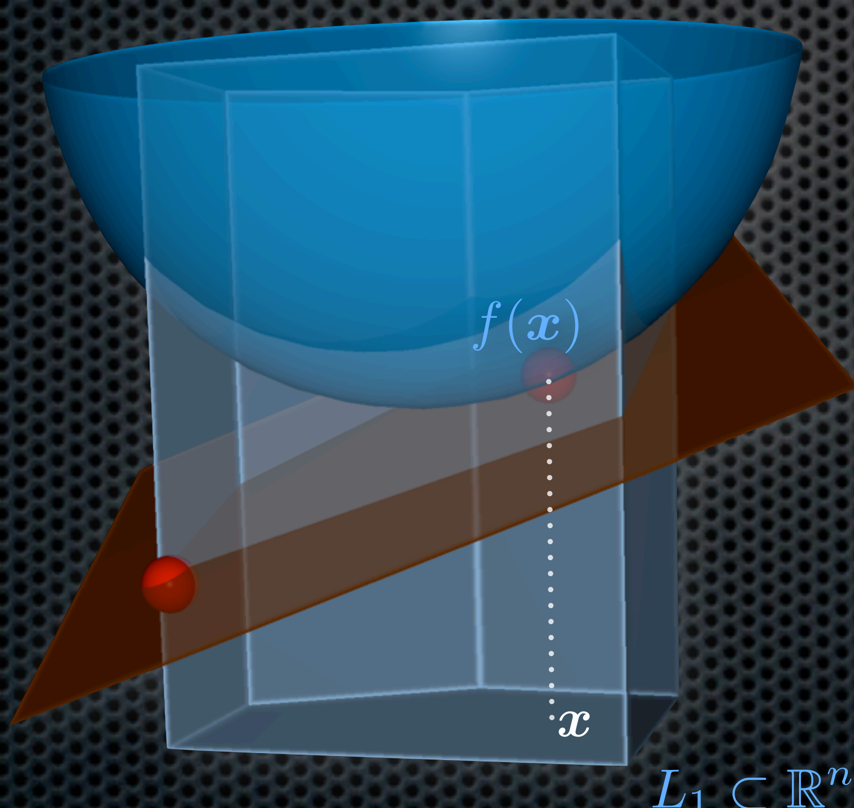
recover a sparse *x* from a noisy measurement *b* of *Ax*

Greedy Algorithms

Convex optimization

methods applied to

$$\min_{x \in L_1} \|Ax - b\|^2$$



Signal processing

sparse recovery methods

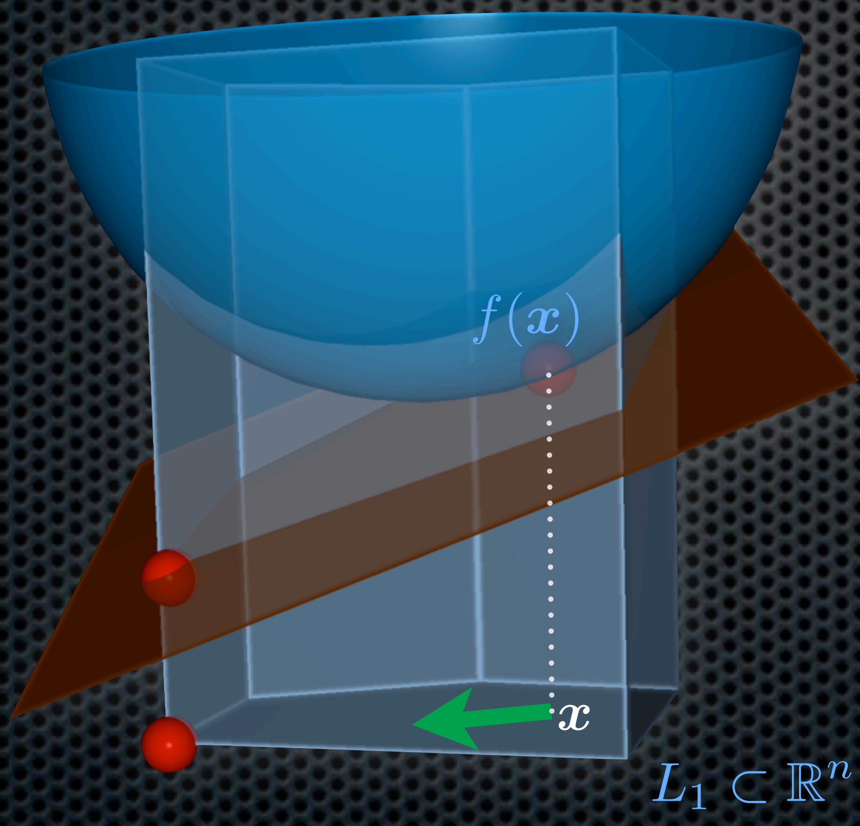
recover a sparse *x* from a noisy measurement *b* of *Ax*

Greedy Algorithms

Convex optimization

methods applied to

$$\min_{x \in L_1} \|Ax - b\|^2$$



Signal processing

sparse recovery methods

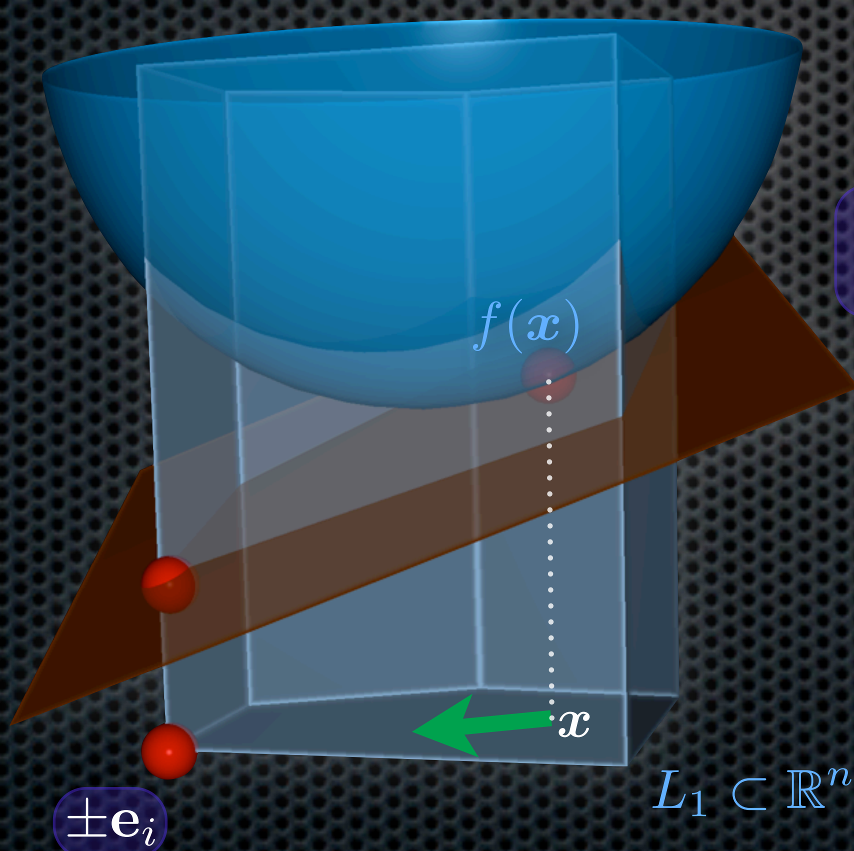
recover a sparse *x* from a noisy measurement *b* of *Ax*

Greedy Algorithms

Convex optimization

methods applied to

$$\min_{x \in L_1} \|Ax - b\|^2$$



$$i := \arg\max_i |\nabla f(x)_i|$$

Signal processing

sparse recovery methods

recover a sparse *x* from a noisy measurement *b* of *Ax*

Greedy Algorithms

Convex optimization

methods applied to

$$\min_{x \in L_1} \|Ax - b\|^2$$

Signal processing

sparse recovery methods

recover a sparse *x* from a noisy measurement *b* of *Ax*

Frank-Wolfe $\xleftrightarrow[\textit{atom per step}]{\textit{selects the same}}$ matching pursuit

$$i := \arg\max_i |\nabla f(x)_i|$$

fully corrective Frank-Wolfe $\xleftrightarrow{\textit{equivalent to}}$ OMP

Thanks