Sparse Convex Optimization Methods for Machine Learning

PhD Defense Talk

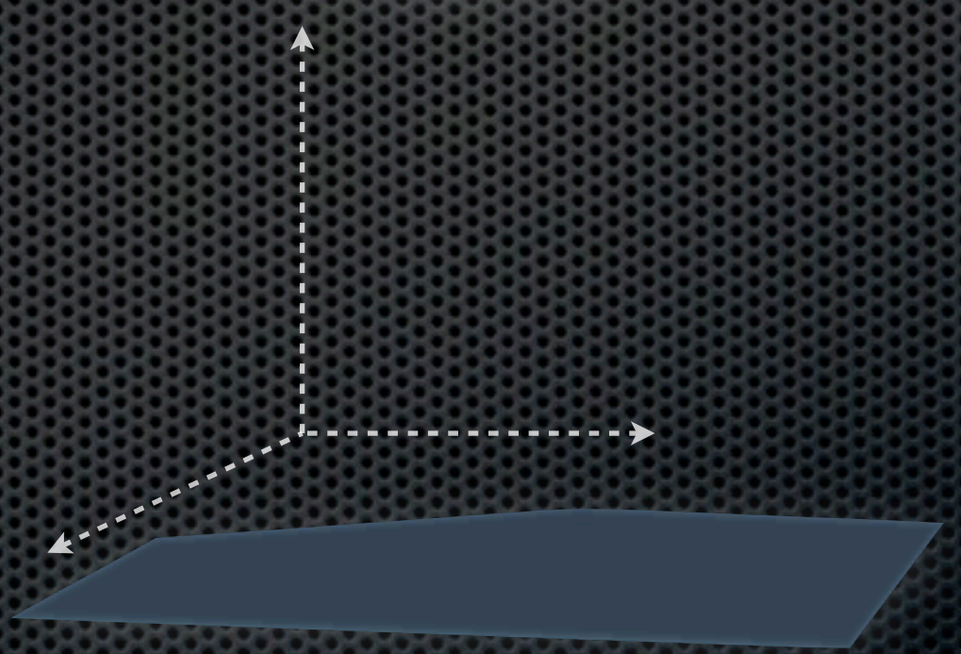
2011 / 10 / 04

Martin Jaggi

Examiner:
Emo Welzl

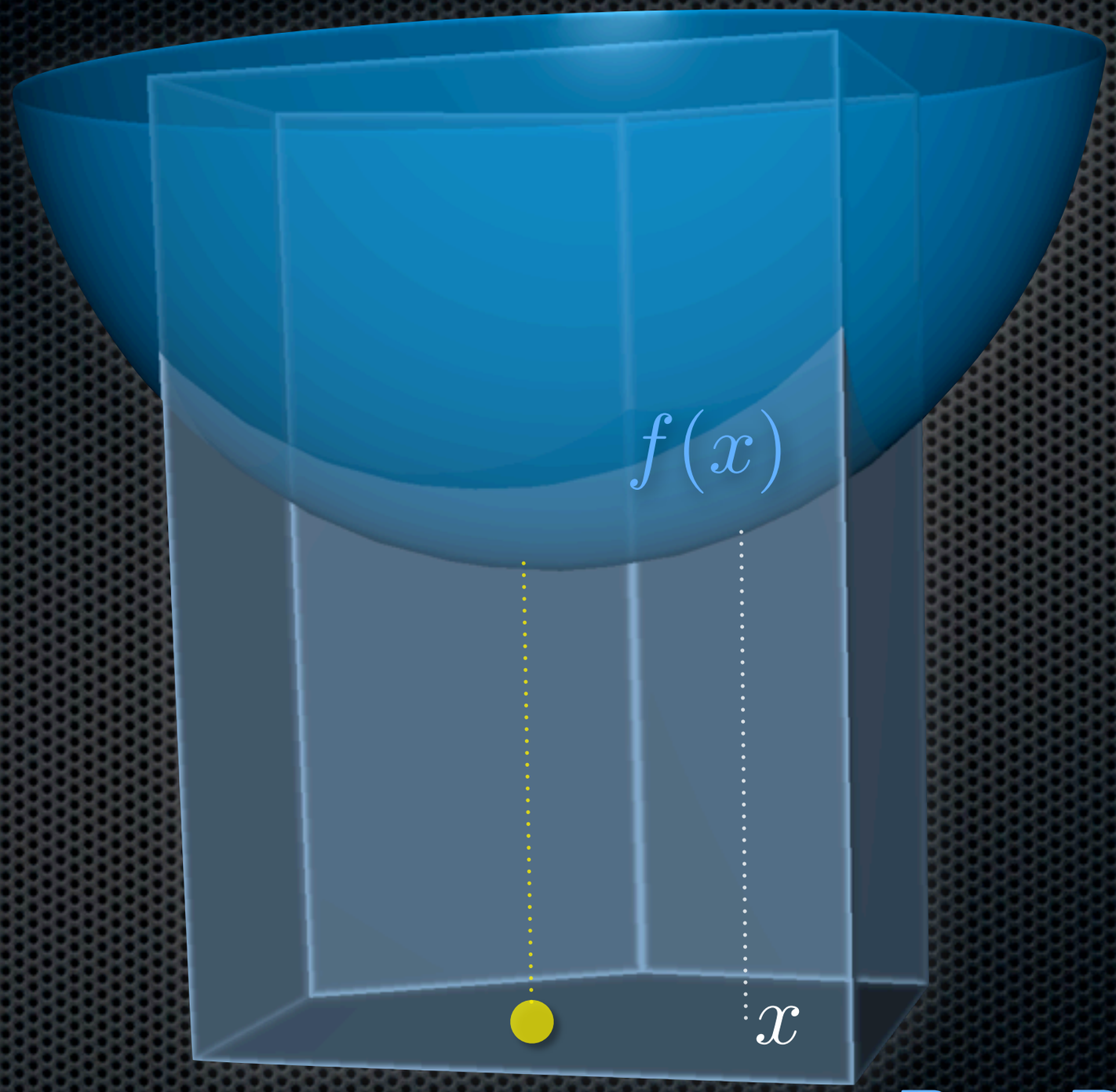
Co-Examiners:
Bernd Gärtner, Elad Hazan, Joachim Giesen, Joachim Buhmann

Convex Optimization



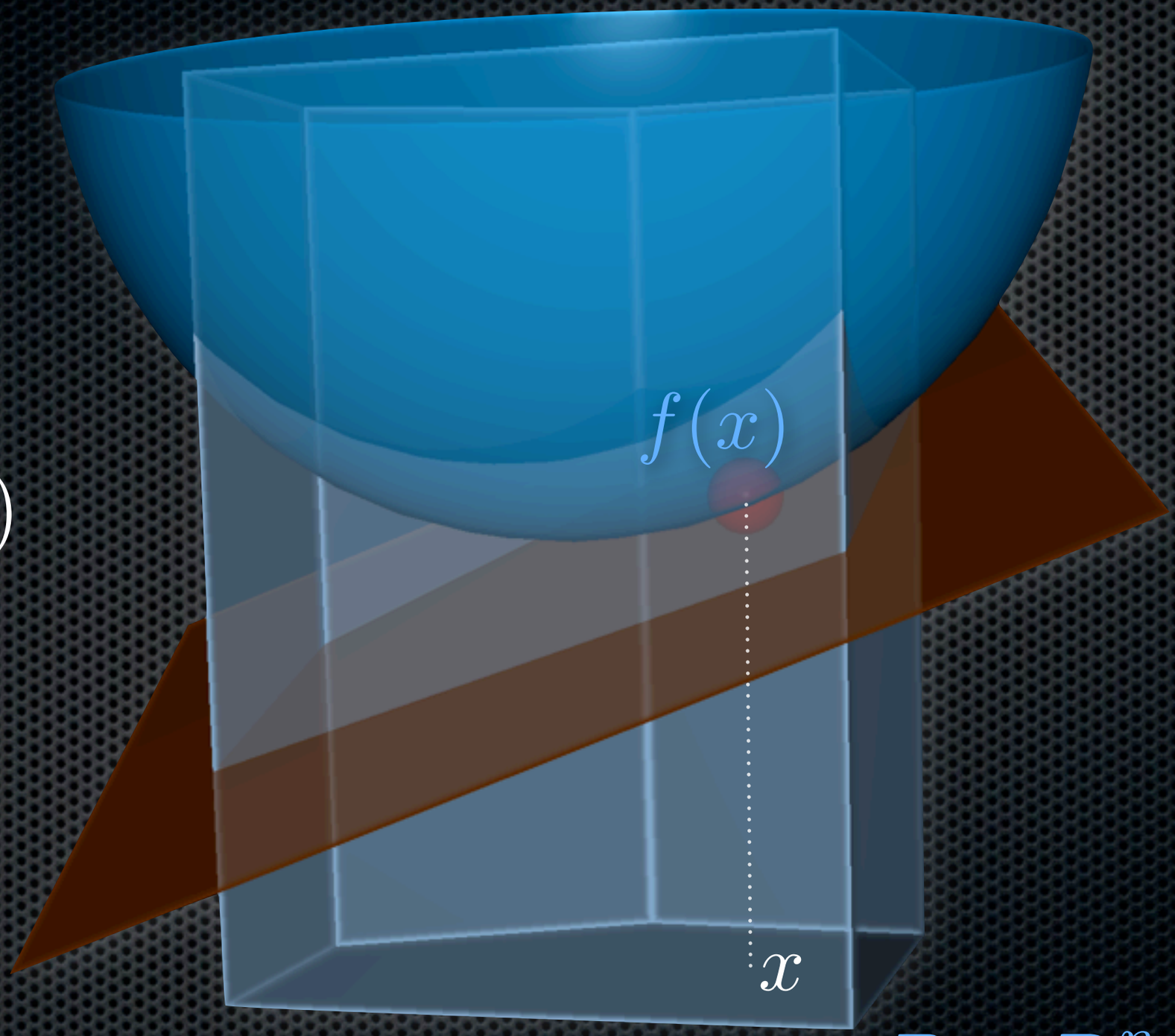
$D \subset \mathbb{R}^n$

$$\min_{x \in D} f(x)$$

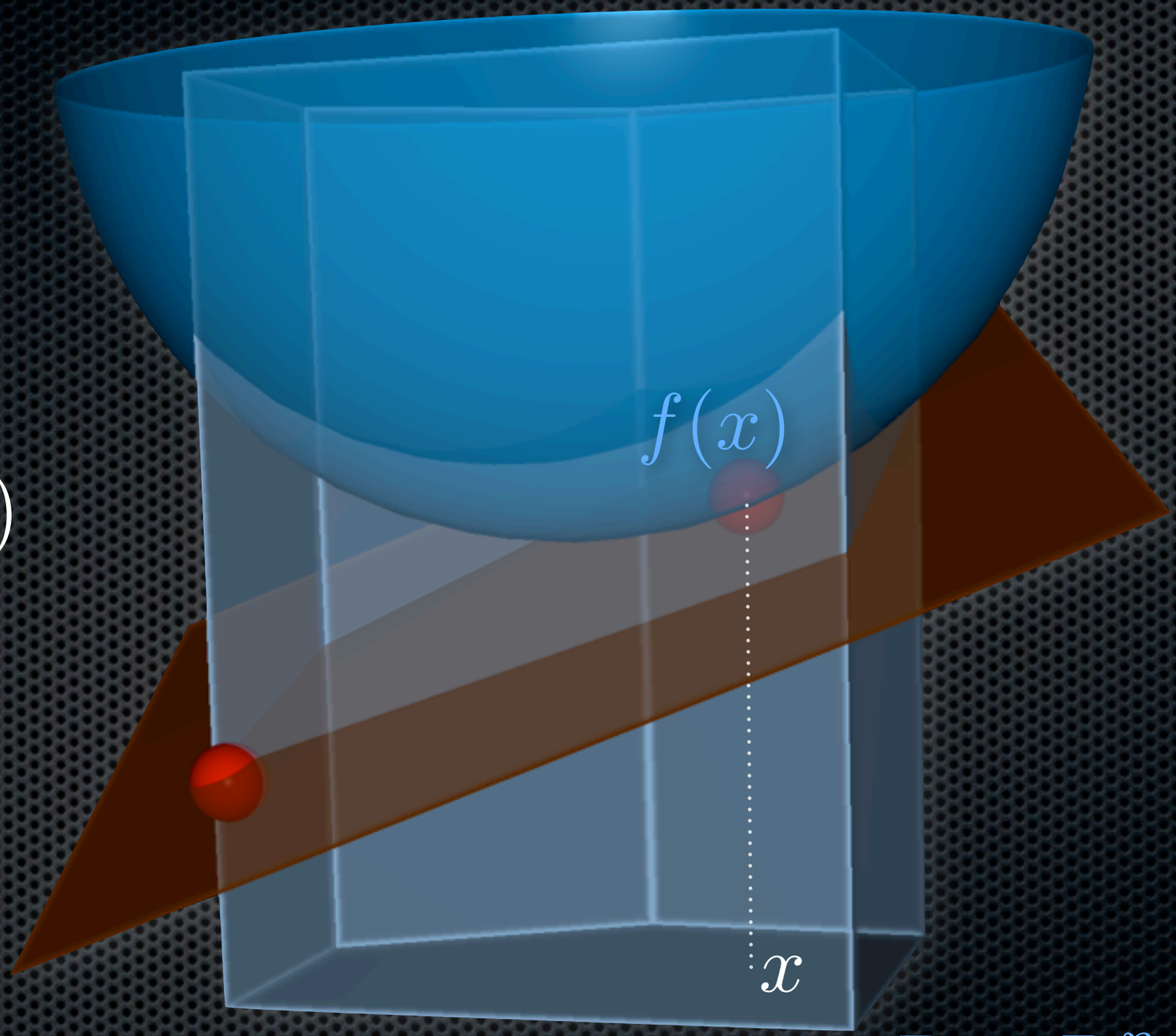


$D \subset \mathbb{R}^n$

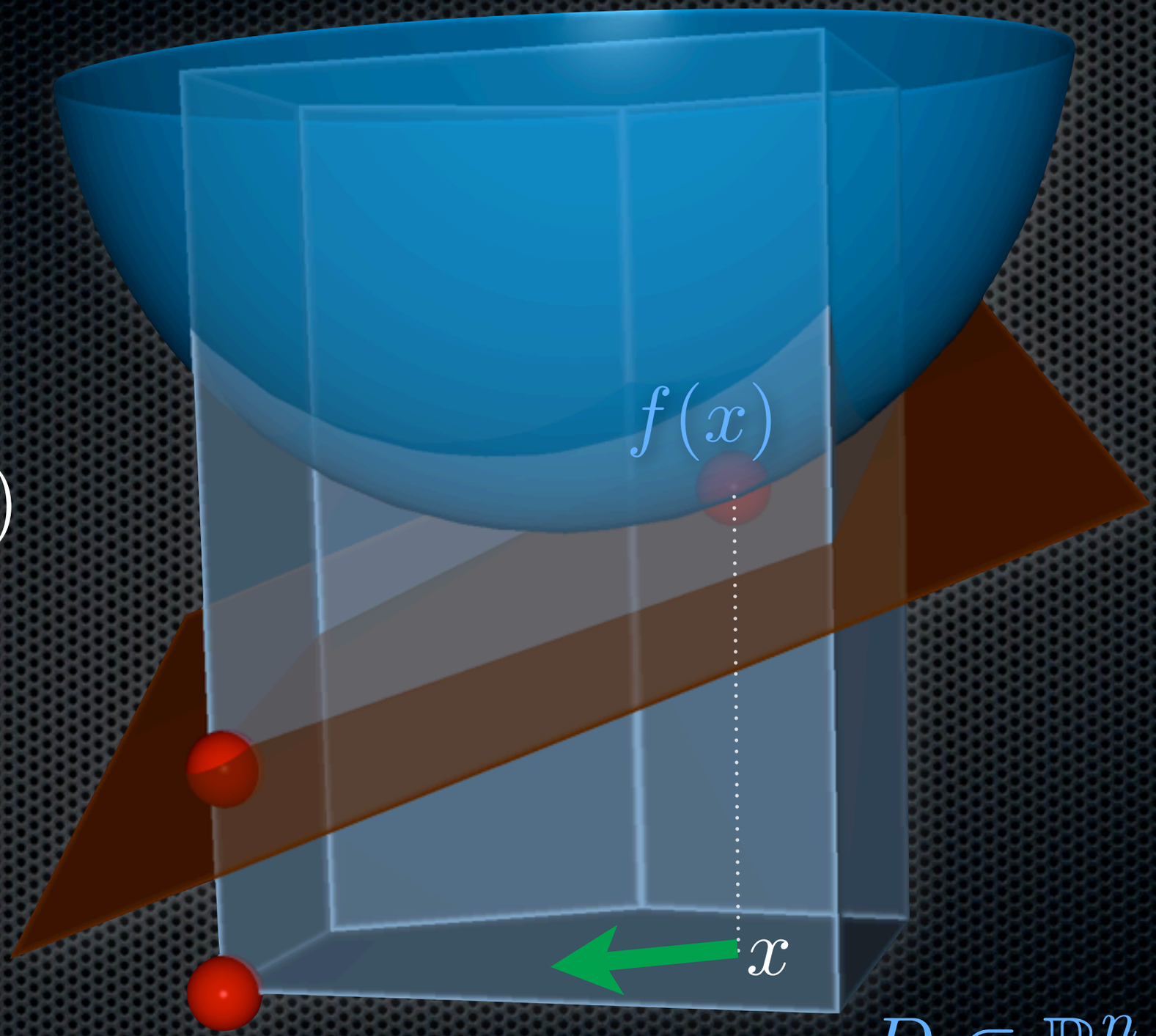
$$\min_{x \in D} f(x)$$



$$\min_{x \in D} f(x)$$

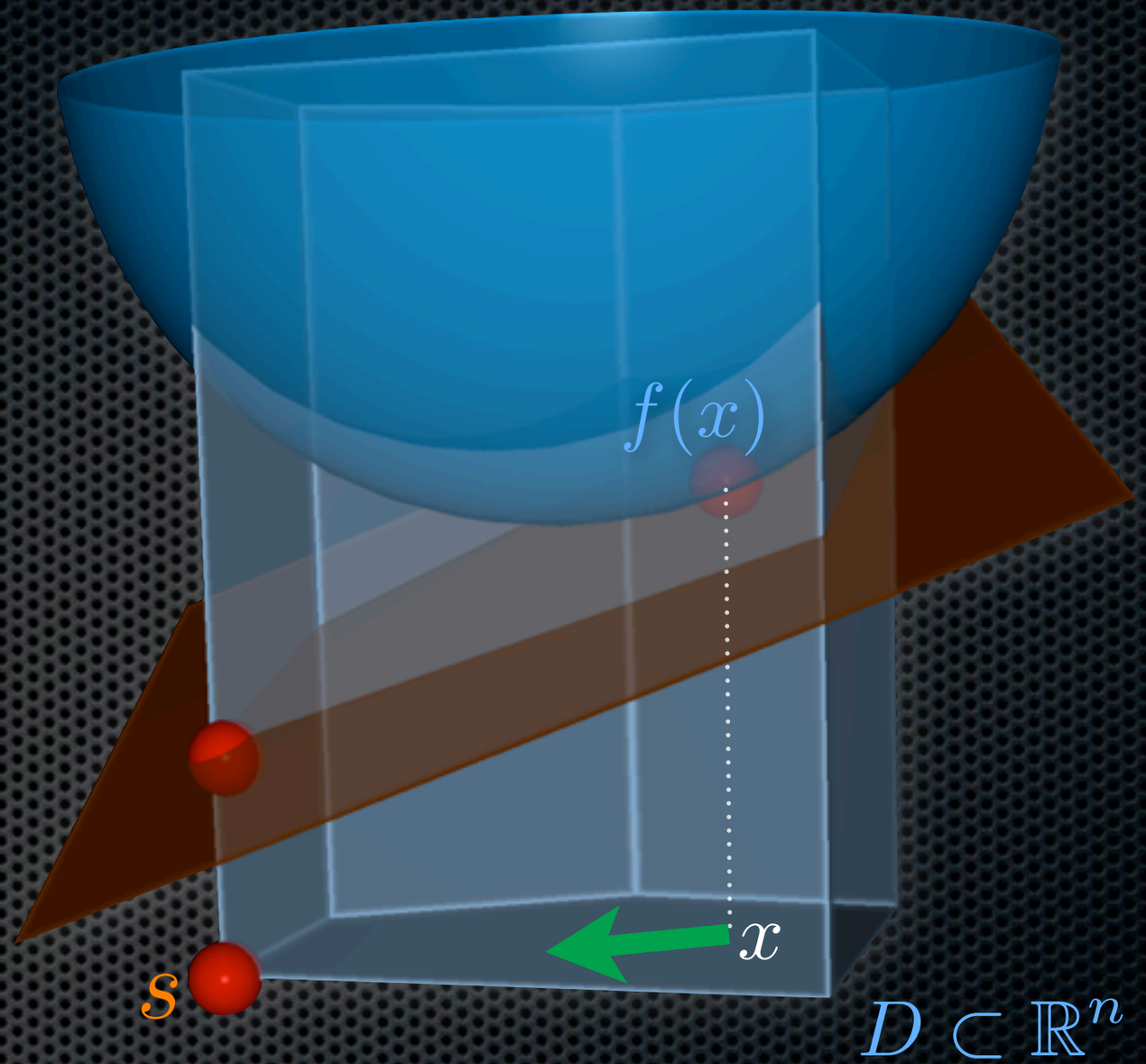


$$\min_{x \in D} f(x)$$



The Linearized Problem

$$\min_{y \in D} f(x) + \langle y - x, d_x \rangle$$



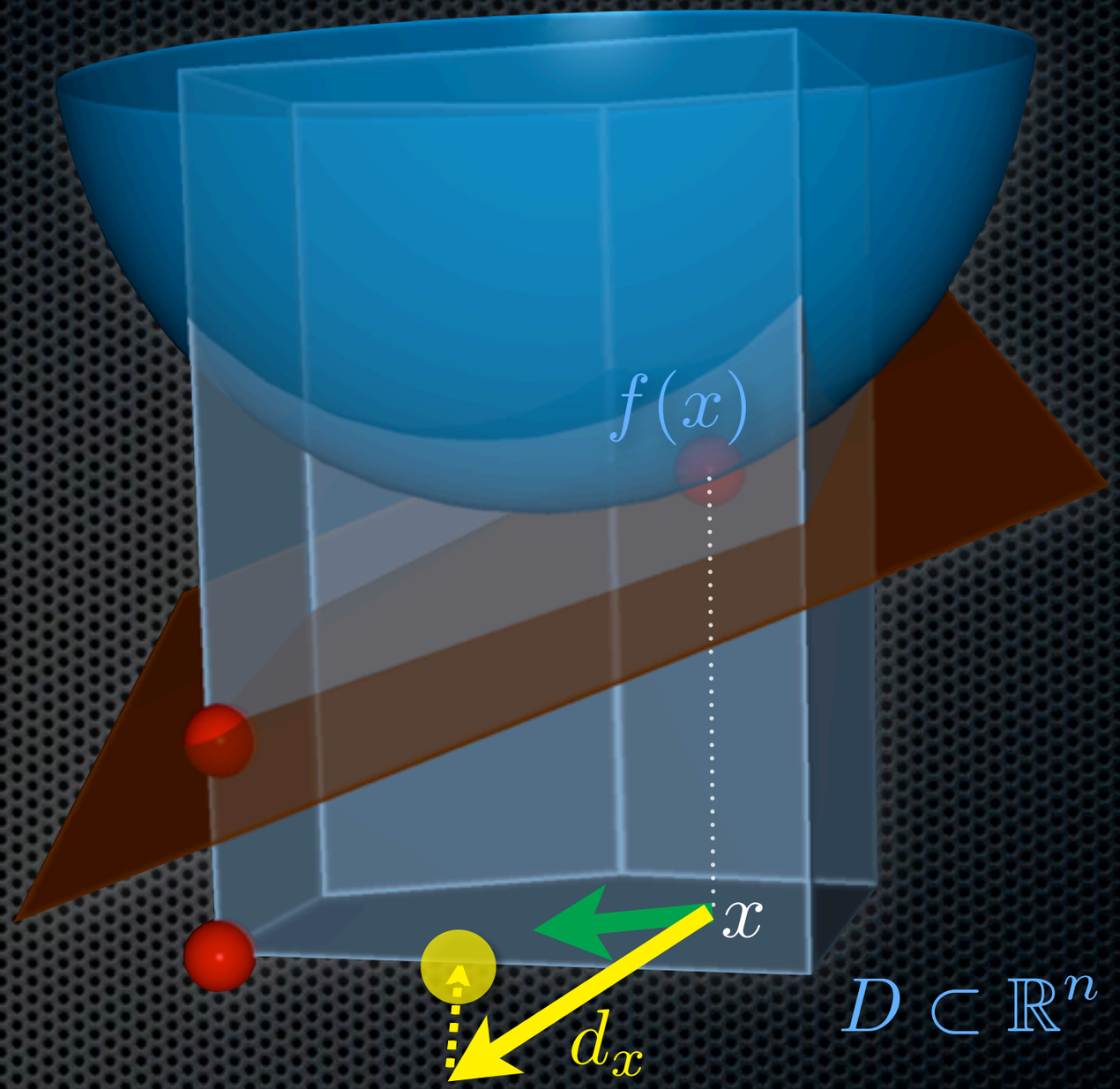
Algorithm 1 Greedy on a Compact Convex Set

Pick an arbitrary starting point $x^{(0)} \in D$
for $k = 0 \ldots \infty$ **do**
  Let $d_x \in \partial f(x^{(k)})$ be a subgradient to f at $x^{(k)}$
  Compute $s := \text{approx} \arg\min_{y \in D} \langle y, d_x \rangle$
  Let $\alpha := \frac{2}{k+2}$
  Update $x^{(k+1)} := x^{(k)} + \alpha(s - x^{(k)})$
end for

Theorem:
Algorithm obtains **accuracy** $O\left(\frac{1}{k}\right)$ after k **steps**.

The Linearized Problem

$$\min_{y \in D} \; f(x) + \langle y - x, d_x \rangle$$



	Our Method	***Gradient Descent***
Cost per step	*Approx.* solve linearized problem on D	Projection back to D
Convergence	$1/k$	$1/k$
Sparse / Low Rank Solutions	✓ (depending on the domain)	✗

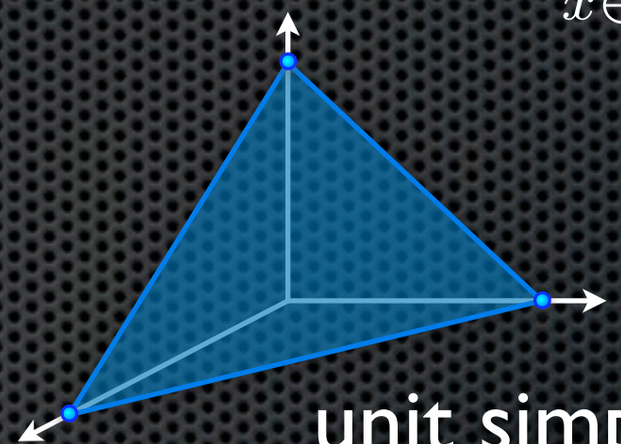
History & Related Work

	Domain	*Known Stepsize*	*Approx. Subproblem*	*Primal-Dual Guarantee*
Frank & Wolfe 1956	linear inequality constraints	✗	✗	✗
Dunn 1978, 1980	general bounded convex domain	✗	✓	✗
Zhang 2003	convex hulls	✗	✓	✗
Clarkson 2008, 2010	unit simplex	✓	✗	✓
Hazan 2008	semidefinite matrices of bounded trace	✓	✓	✓
J. PhD Thesis	general bounded convex domain	✓	✓	✓

Sparse Approximation

$$\min_{x \in \Delta_n} f(x)$$

$$D := conv(\{\mathbf{e}_i \mid i \in [n]\})$$



unit simplex

for $k = 0 \ldots \infty$ **do**
  Let $d_x \in \partial f(x^{(k)})$ be a subgradient to f at $x^{(k)}$
  Compute $i := \arg\min_i (d_x)_i$
  Let $\alpha := \frac{2}{k+2}$
  Update $x^{(k+1)} := x^{(k)} + \alpha(\mathbf{e}_i - x^{(k)})$
end for

Corollary:
Algorithm gives an ε-approximate solution of **sparsity** $O\left(\frac{1}{\varepsilon}\right)$.

[Clarkson SODA '08]

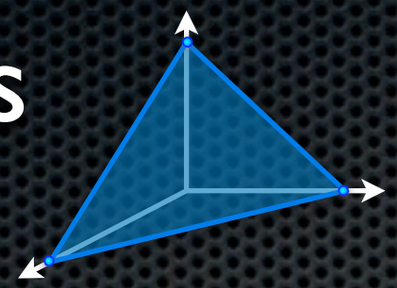
"Coresets"

Sparsity as a function of the approximation quality

lower bound:

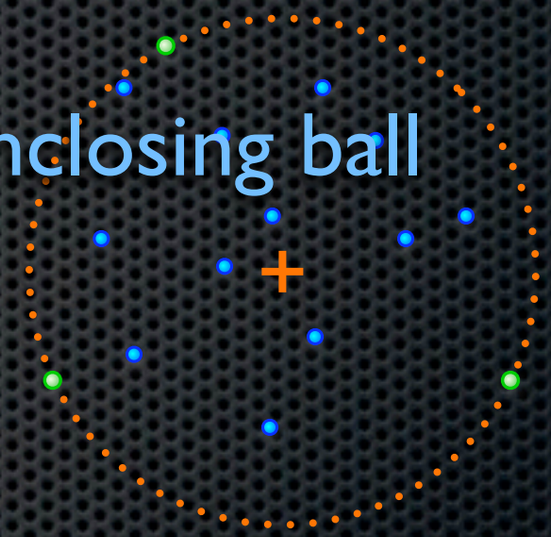
$$\Omega\left(\frac{1}{\varepsilon}\right)$$

Applications

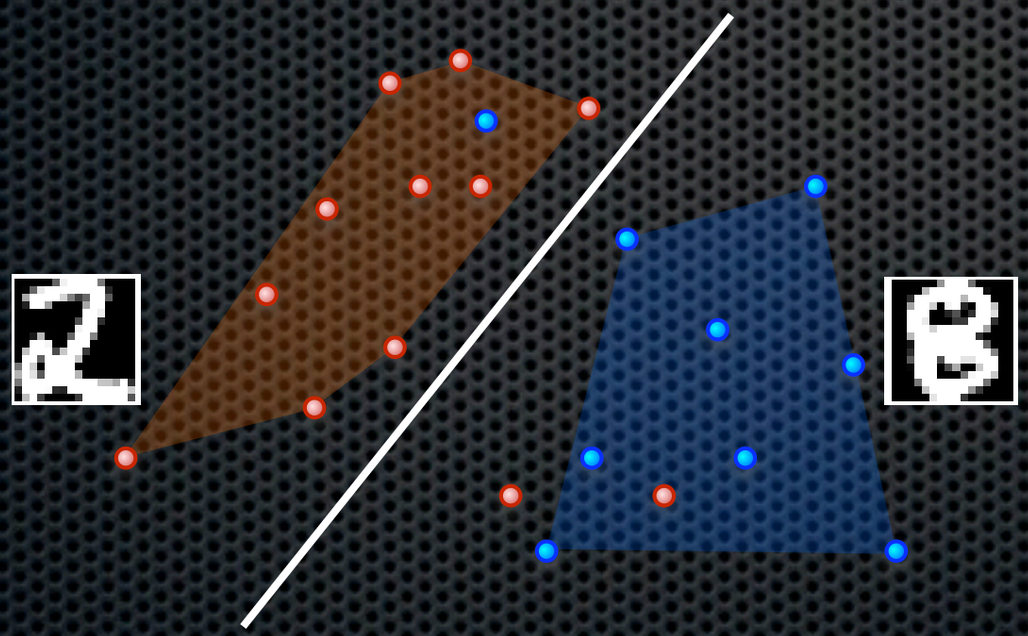


- Linear Classifiers
  (such as Support Vector Machines, ℓ_2-loss)

$$\min_{x \in \Delta_n} x^T (K + t\mathbf{1}) x$$



- Smallest enclosing ball



- Model Predictive Control

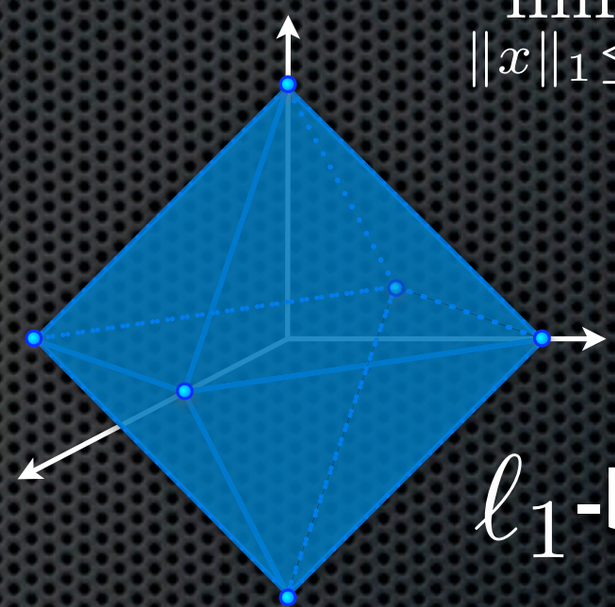
- Mean Variance
  Portfolio Optimization

$$\min_{x \in \Delta_n} x^T C x - t \cdot b^T x$$

Sparse Approximation

$$\min_{\|x\|_1 \le 1} f(x)$$

$$D := conv(\{\pm \mathbf{e}_i \,|\, i \in [n]\})$$



ℓ_1-ball

for $k = 0 \dots \infty$ **do**
  Let $d_x \in \partial f(x^{(k)})$ be a subgradient to f at $x^{(k)}$
  Compute $i := \arg\max_i |(d_x)_i|$,
  and let $s := \mathbf{e}_i \cdot \operatorname{sign}\left((-d_x)_i\right)$
  Let $\alpha := \frac{2}{k+2}$
  Update $x^{(k+1)} := x^{(k)} + \alpha(s - x^{(k)})$
end for

Corollary:
Algorithm gives an ε-approximate solution of **sparsity** $O\left(\frac{1}{\varepsilon}\right)$.

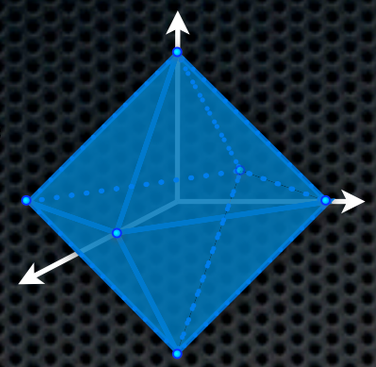
"Coresets"

Sparsity as a function of the approximation quality

lower bound:

$$\Omega\left(\frac{1}{\varepsilon}\right)$$

Applications



- ℓ_1-regularized regression

$$\min_{\|x\|_1 \leq t} \|Ax - b\|_2^2$$

Sparse Recovery

Low Rank Approximation

$$\min_{x \in D} f(x)$$

$$D := conv(\{vv^T \mid \substack{v \in \mathbb{R}^n, \\ \|v\|_2 = 1}\}) \quad = \quad \begin{array}{l} X \in Sym^{n \times n} \\ X \succeq 0 \\ Tr(X) = 1 \end{array}$$

spectahedron

for $k = 0 \ldots \infty$ **do**
  Let $D_X \in \partial f(X^{(k)})$ be a subgradient to f at $X^{(k)}$
  Let $\alpha := \frac{2}{k+2}$
  Compute $v := v^{(k)} =$ ApproxEV $(D_X, \alpha C_f)$
  Update $X^{(k+1)} := X^{(k)} + \alpha(vv^T - X^{(k)})$
end for

Corollary:
Algorithm gives an ε-approximate solution of **rank** $O\left(\frac{1}{\varepsilon}\right)$.

[Hazan LATIN '08]

lower bound:

$$\Omega\left(\frac{1}{\varepsilon}\right)$$

Applications

- Trace norm regularized problems

$$\min_{\|X\|_* \leq t} \quad f(X)$$

Low-Rank Matrix Recovery

- Max norm regularized problems

Matrix Factorizations

for recommender systems

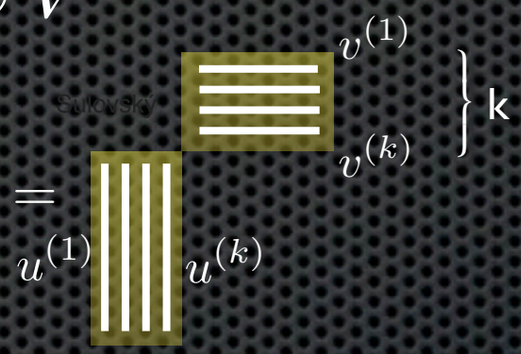
$Y =$

Movies

Customers

1	3			
	4			
1				
2		3		
5				3
	2			
2		1	3	

$\approx \; UV^T$



The Netflix challenge:

17k Movies

500k Customers

100M Observed Entries

(≈ 1%)

$$\min_{U,V} \sum_{(i,j)\in\Omega} (Y_{ij} - (UV^T)_{ij})^2$$

$$s.t. \; \|U\|_{Fro}^2 + \|V\|_{Fro}^2 = t$$

Is equivalent to:

$$\min_{X \succeq 0} f(X)$$

$$s.t. \; Tr(X) = t$$

$$\begin{pmatrix} UU^T & UV^T \\ VU^T & VV^T \end{pmatrix} =: X$$

[J, Sulovský ICML 2010]

A Simple Alternative Optimization Duality

The Problem

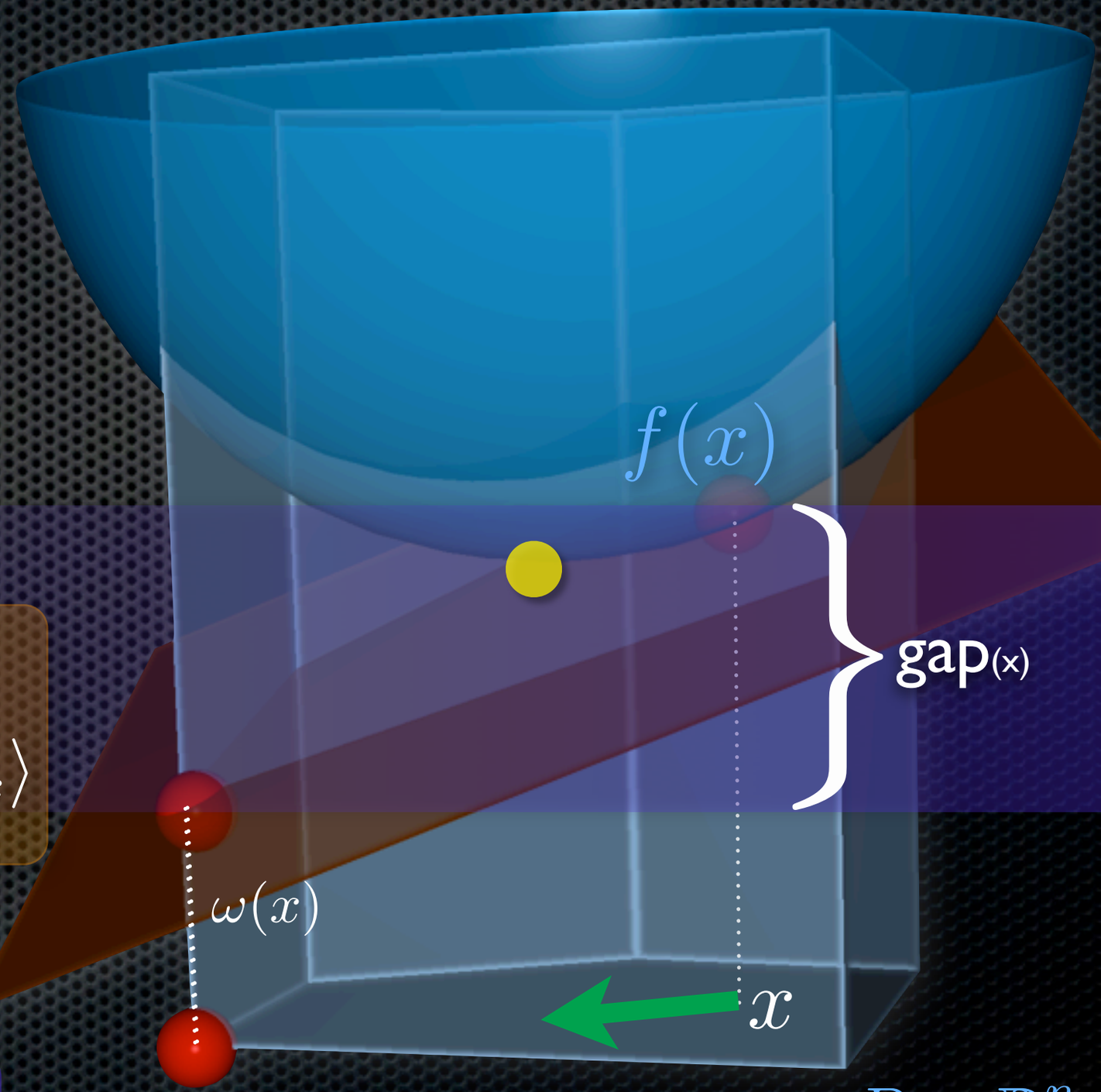
$$\min_{x \in D} f(x)$$

The Dual

$$\omega(x) := \min_{y \in D} f(x) + \langle y - x, d_x \rangle$$

Weak Duality

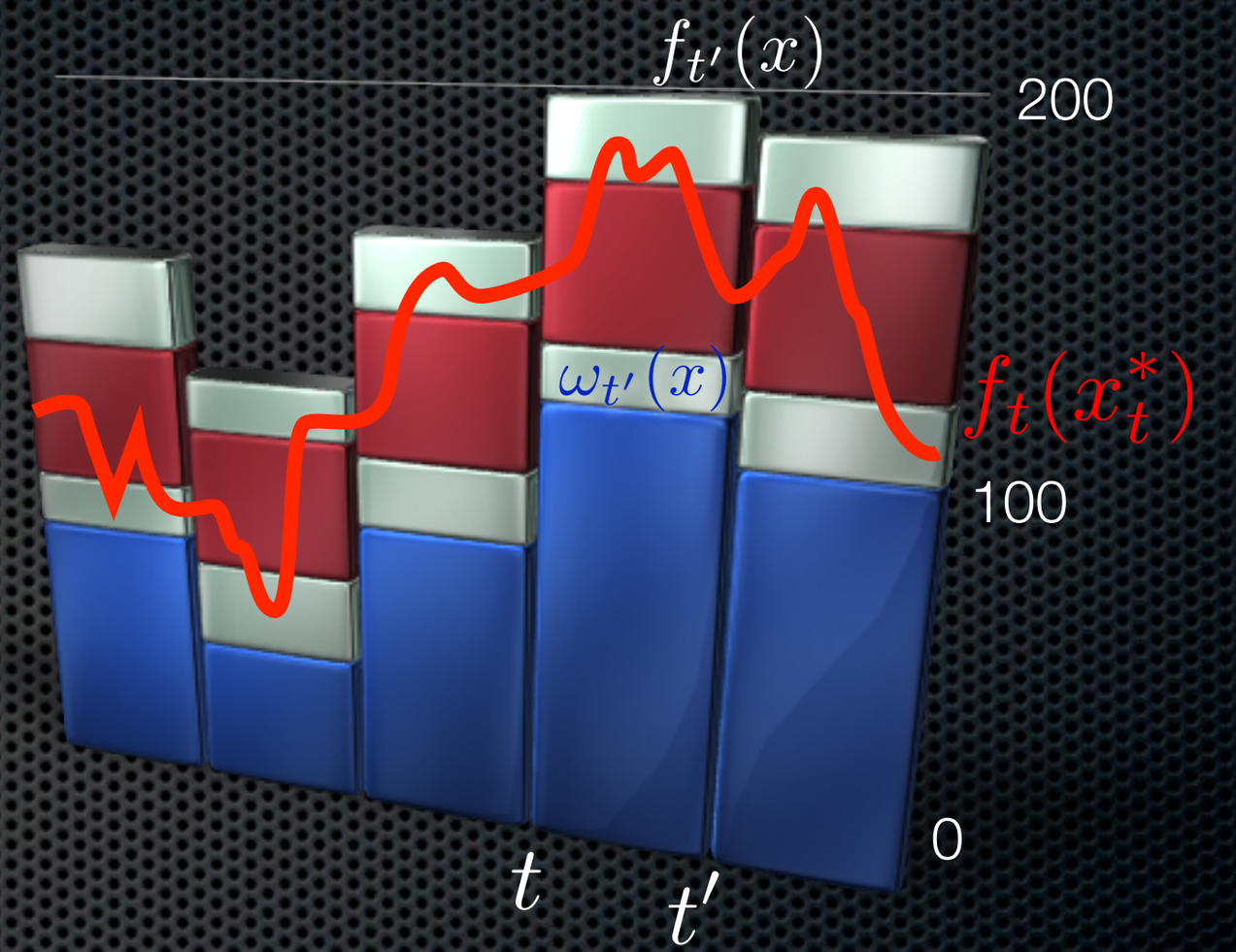
$$\omega(x) \leq f(x^*) \leq f(x')$$



Pathwise Optimization

The Parameterized Problem

$$\min_{x \in D} f_t(x)$$



"Better than necessary"

$$g_t(x) \leq \tfrac{\varepsilon}{2}$$

The difference

$$g_{t'}(x) - g_t(x) \leq \varepsilon\left(1 - \tfrac{1}{2}\right) \quad \Leftarrow \quad |t' - t| \leq \varepsilon \cdot P_f$$

"Continuity in the parameter"

"Still good enough"

$$g_{t'}(x) \leq \varepsilon$$

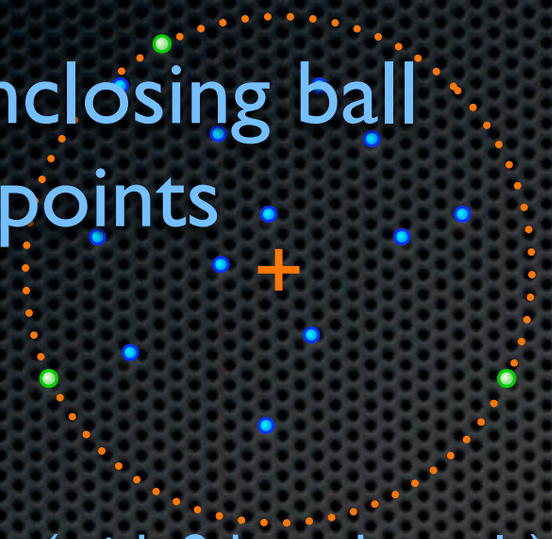
Theorem:

There are $O\left(\frac{1}{\varepsilon}\right)$ many intervals of piecewise constant ε-approx. solutions.

[Giesen, J, Laue ESA 2010]

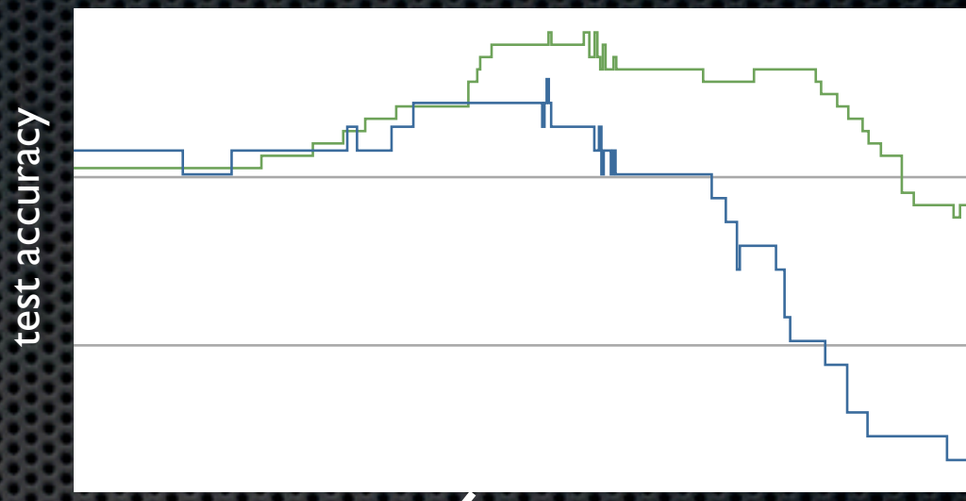
Applications

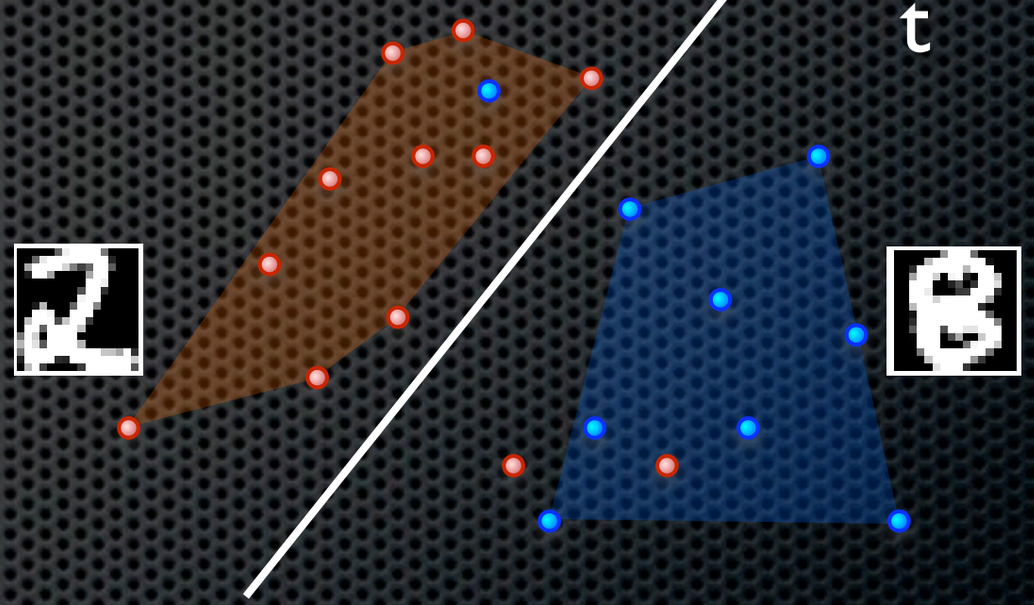
- Smallest enclosing ball of moving points



- SVMs, MKL (with 2 base kernels)

$$\min_{x \in \Delta_n} x^T (K + t\mathbf{1}) x$$





- Model Predictive Control

- Mean Variance Portfolio Optimization

$$\min_{x \in \Delta_n} x^T C x - t \cdot b^T x$$

- robust PCA

- Recommender Systems

Thanks

co-authors:

Bernd Gärtner

Joachim Giesen

Soeren Laue

Marek Sulovský

3D visualization:

Robert Carnecky