A fresh look at the Frank-Wolfe algorithm

with applications to sparse convex optimization

Martin Jaggi

Ecole Polytechnique, Paris

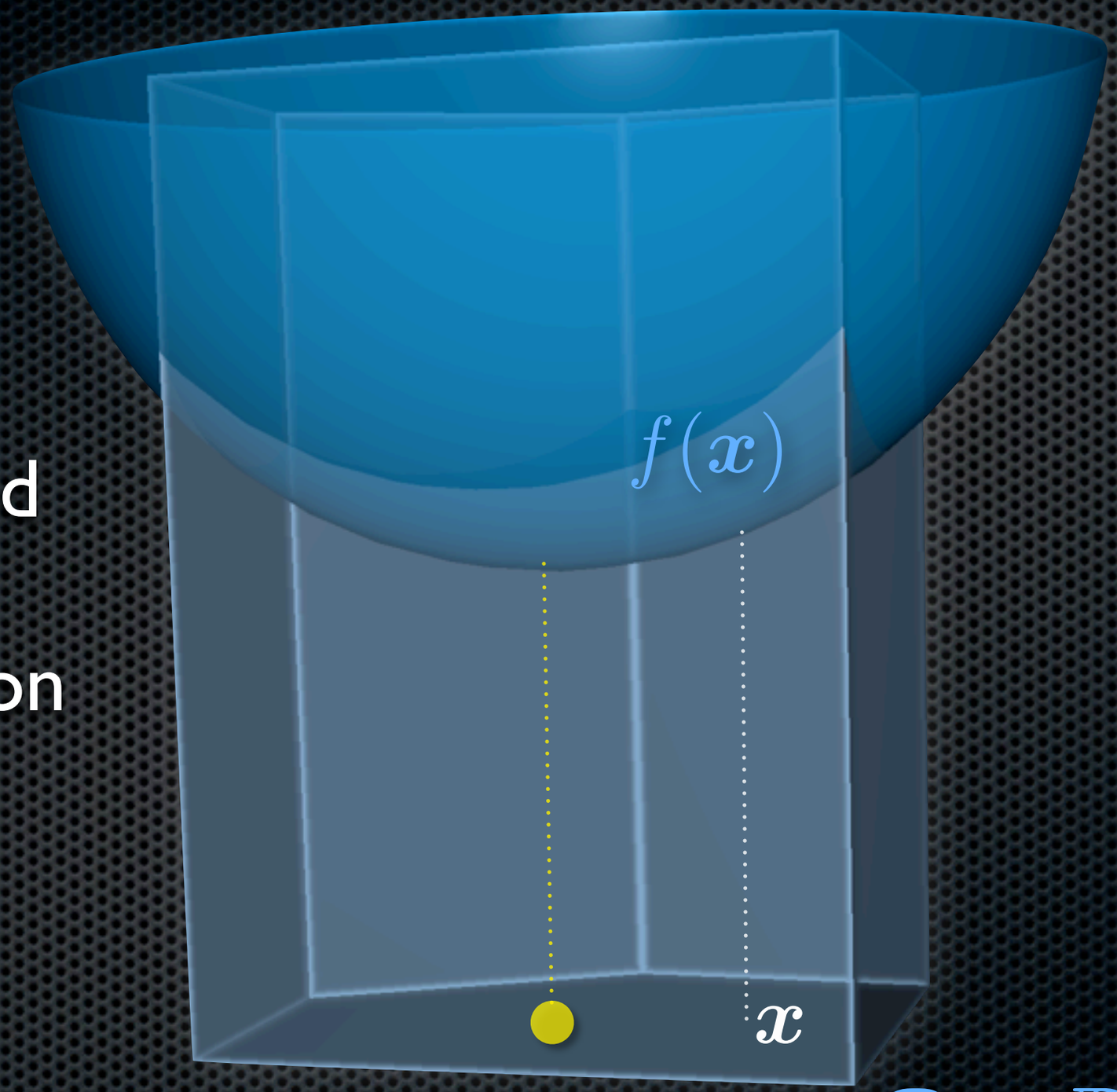
ICCOPT 2013 / 08 / 01

[J. 2011 PhD Thesis,
summary: J. 2013 ICML Paper]

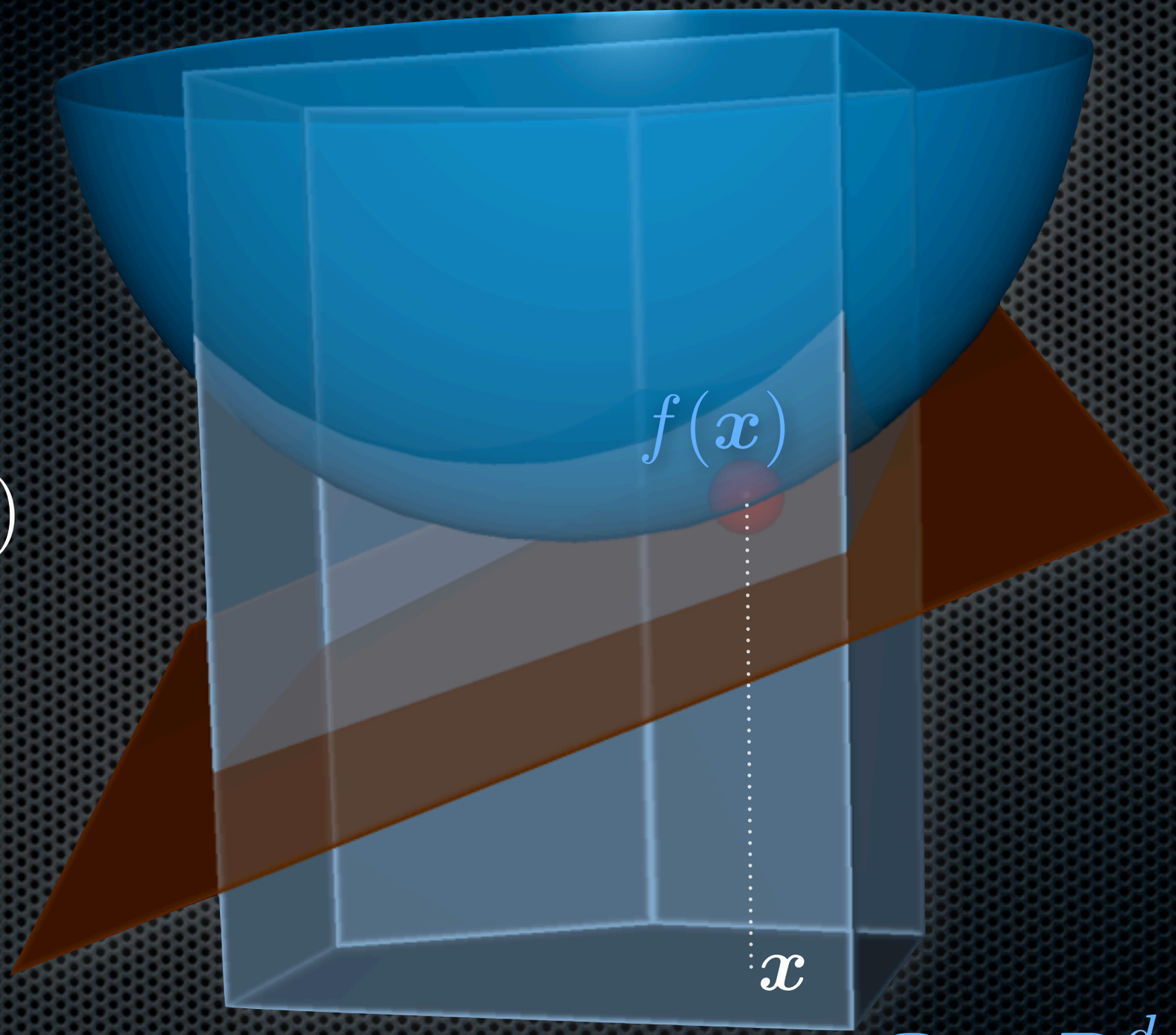
Constrained Convex Optimization

Constrained Convex Optimization

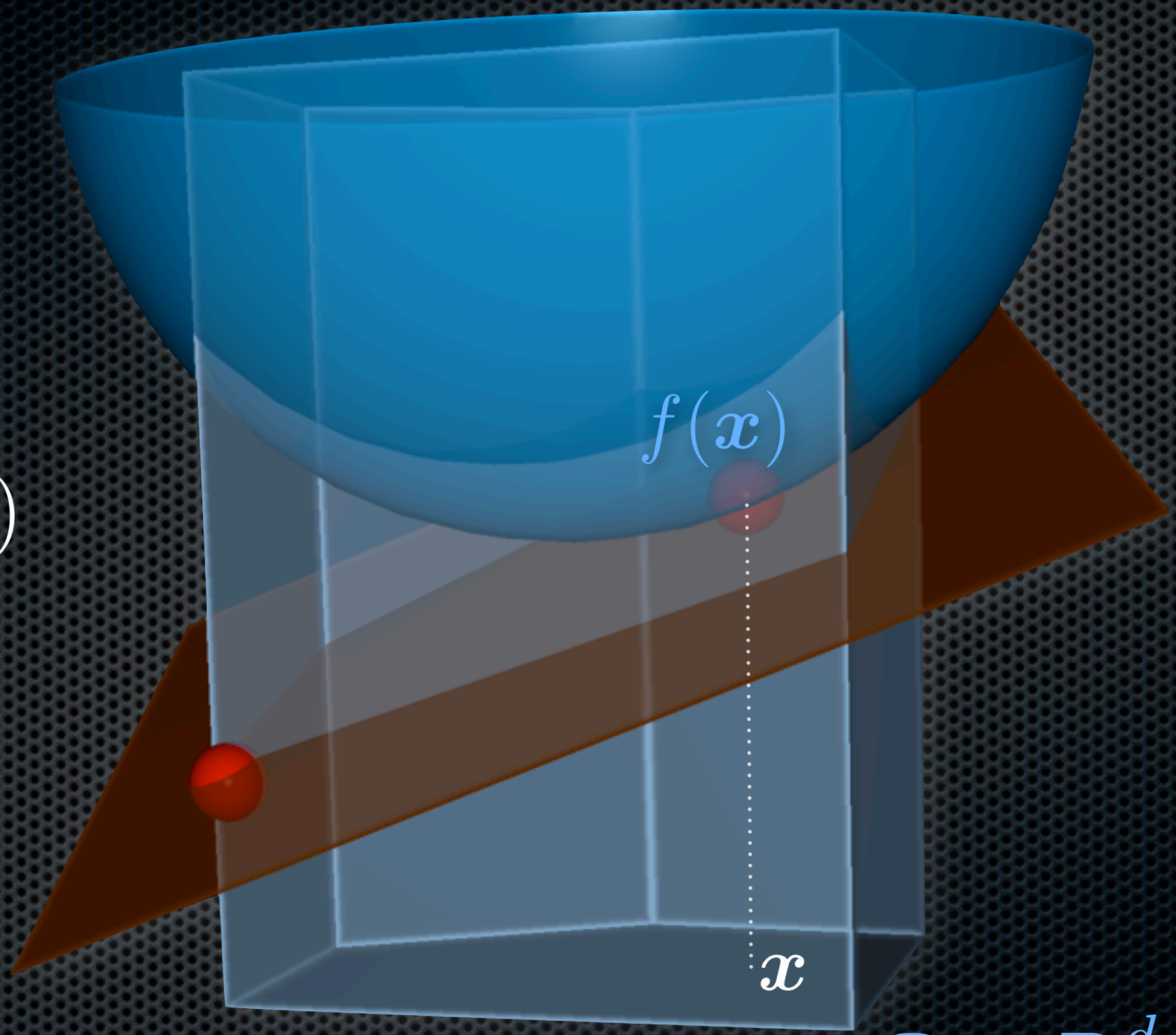
$$\min_{\boldsymbol{x} \in \mathcal{D}} f(\boldsymbol{x})$$



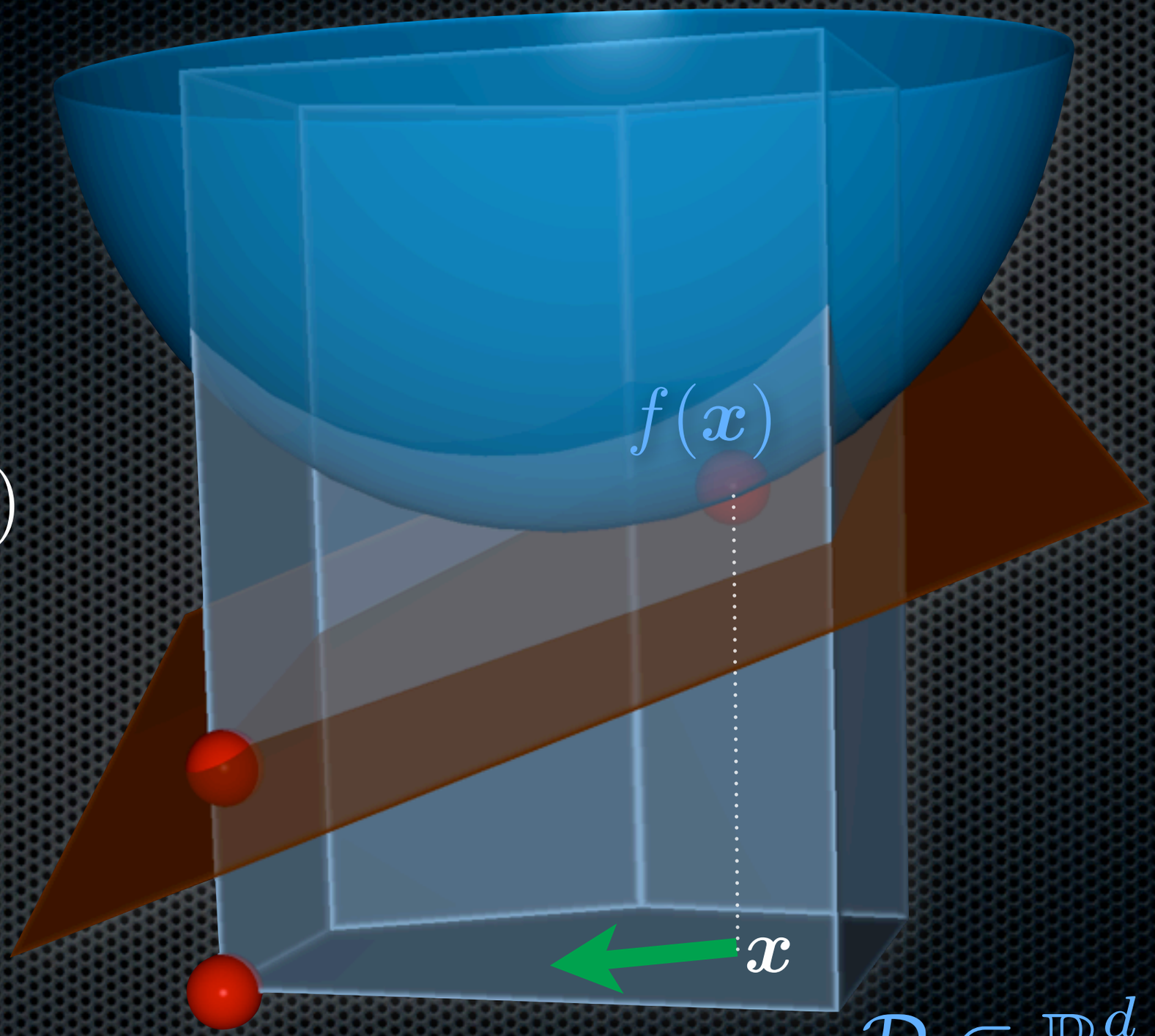
$$\min_{\boldsymbol{x} \in \mathcal{D}} f(\boldsymbol{x})$$



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

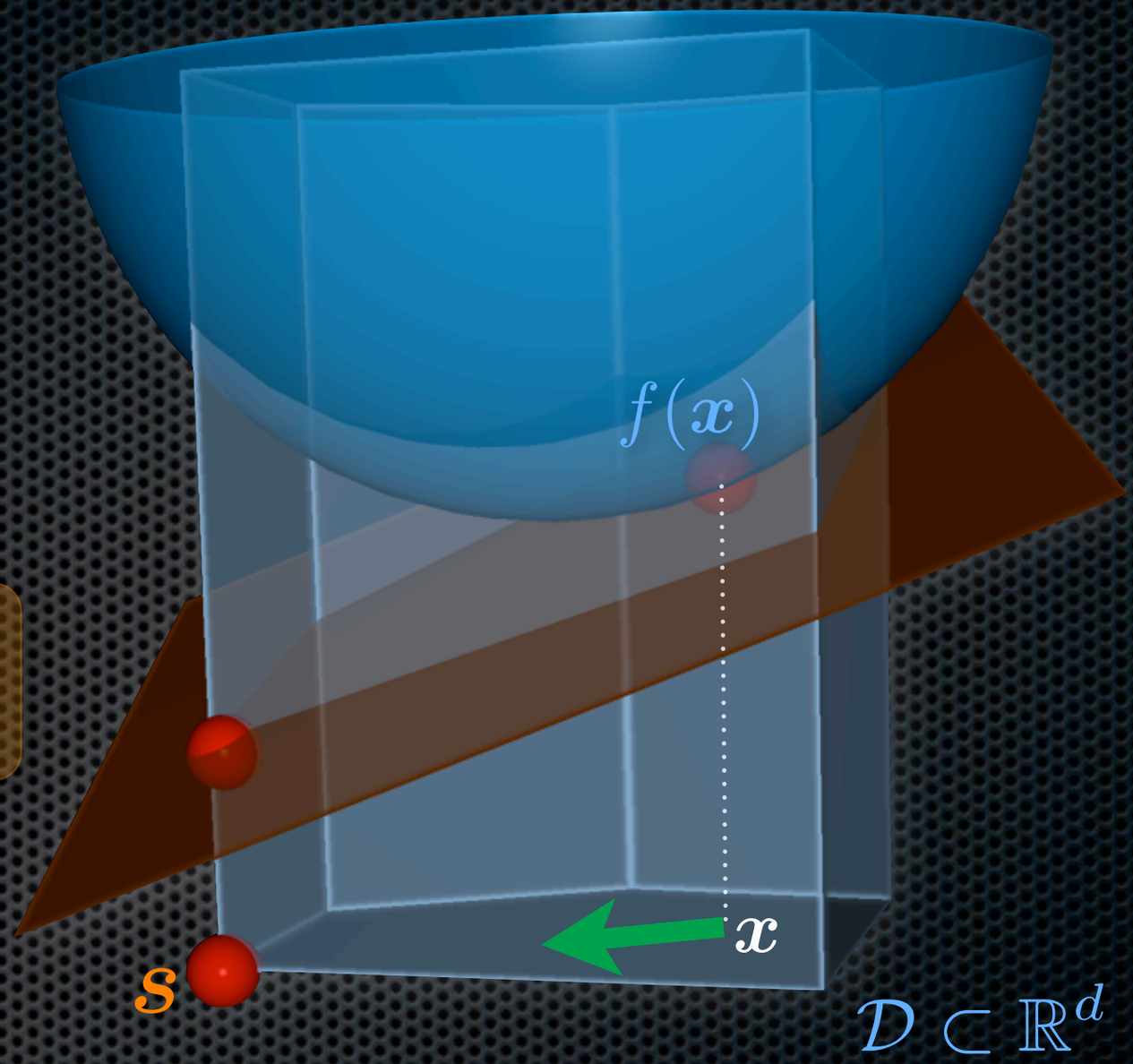


$$\min_{\boldsymbol{x}\in\mathcal{D}} f(\boldsymbol{x})$$



The Linearized Problem

$$\min_{\boldsymbol{s}' \in \mathcal{D}} f(\boldsymbol{x}) + \left\langle \boldsymbol{s}' - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \right\rangle$$

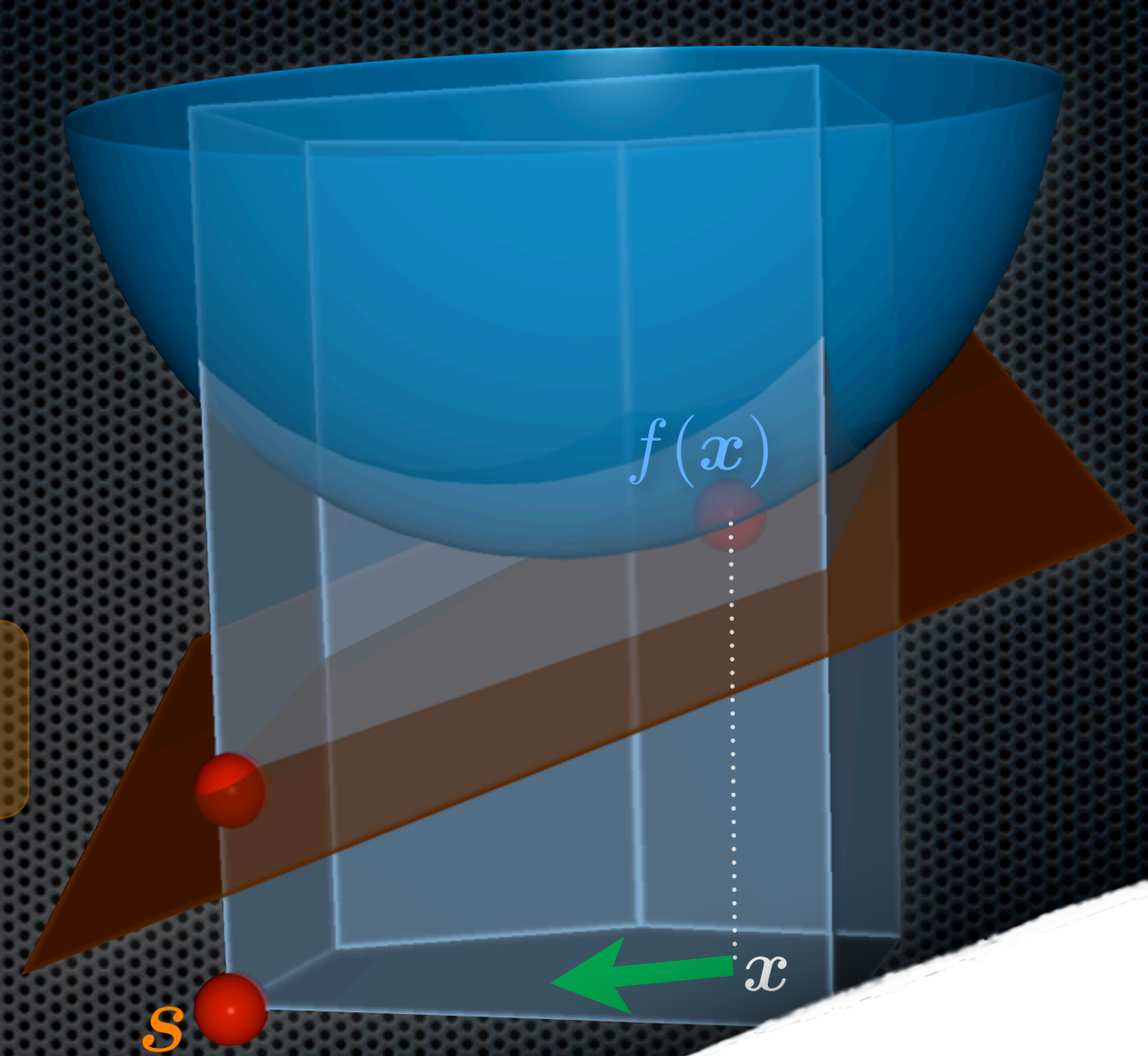


Algorithm 1 Frank-Wolfe

Let $\boldsymbol{x}^{(0)} \in \mathcal{D}$
for $k = 0 \dots K$ **do**
  Compute $\boldsymbol{s} := \arg\min_{\boldsymbol{s}' \in \mathcal{D}} \left\langle \boldsymbol{s}', \nabla f(\boldsymbol{x}^{(k)}) \right\rangle$
  Let $\gamma := \frac{2}{k+2}$
  Update $\boldsymbol{x}^{(k+1)} := (1-\gamma)\boldsymbol{x}^{(k)} + \gamma \boldsymbol{s}$
end for

The Linearized Problem

$$\min_{\boldsymbol{s}' \in \mathcal{D}} f(\boldsymbol{x}) + \left\langle \boldsymbol{s}' - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \right\rangle$$



Algorithm 1 Frank-Wolfe

Let $\boldsymbol{x}^{(0)} \in \mathcal{D}$
for $k = 0 \dots K$ **do**
  Compute $\boldsymbol{s} := \arg\min_{\boldsymbol{s}' \in \mathcal{D}} \left\langle \boldsymbol{s}', \nabla f \right.$
  Let $\gamma := \frac{2}{k+2}$
  Update $\boldsymbol{x}^{(k+1)} := (1 -$
end for

AN ALGORITHM FOR QUADRATIC PROGRAMMING

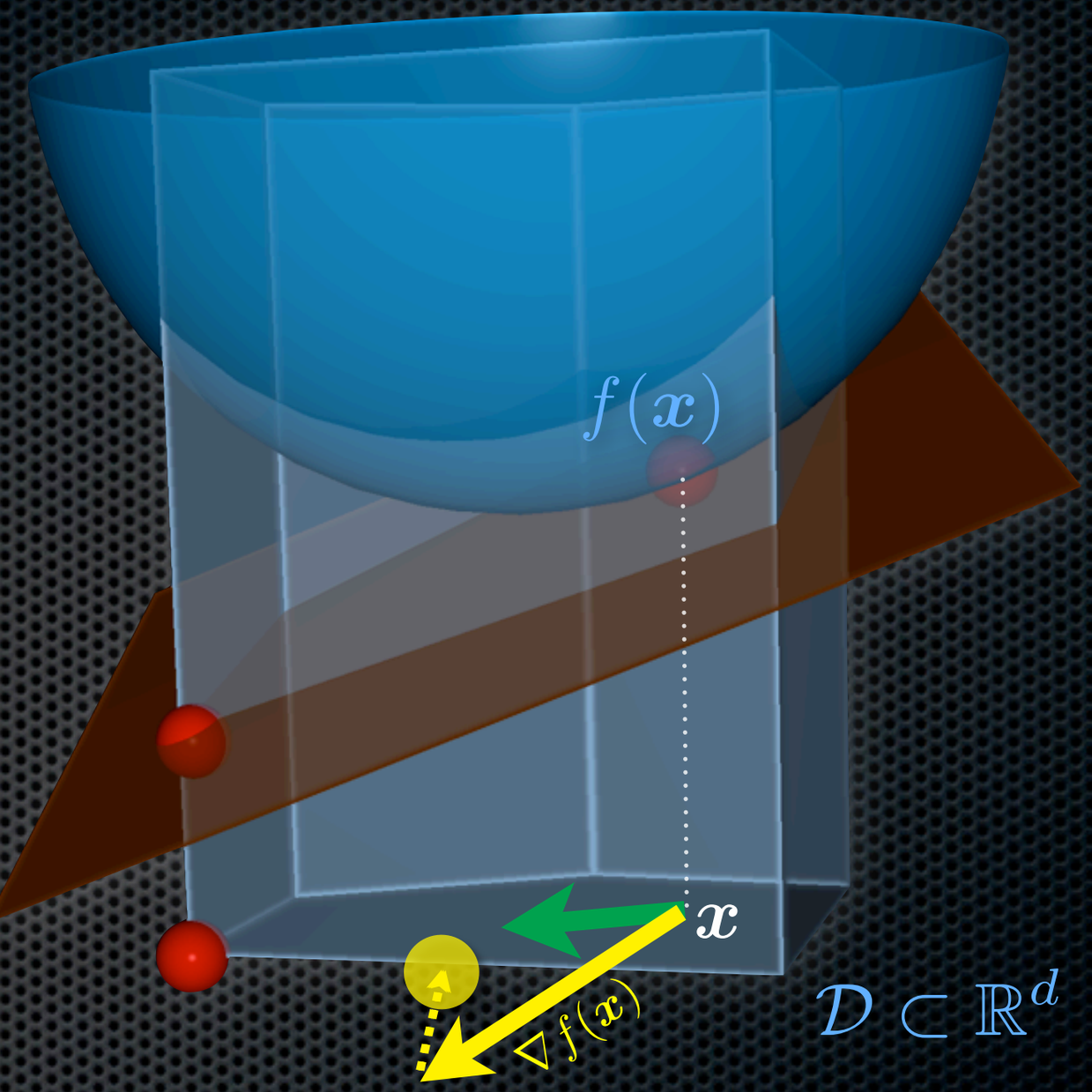
Marguerite Frank and Philip Wolfe[1]
Princeton University

1956

A finite iteration method for calculating the solution of quad
ogramming problems is described. Extensions to more genera
roblems are suggested.

atic function whose
recent stu

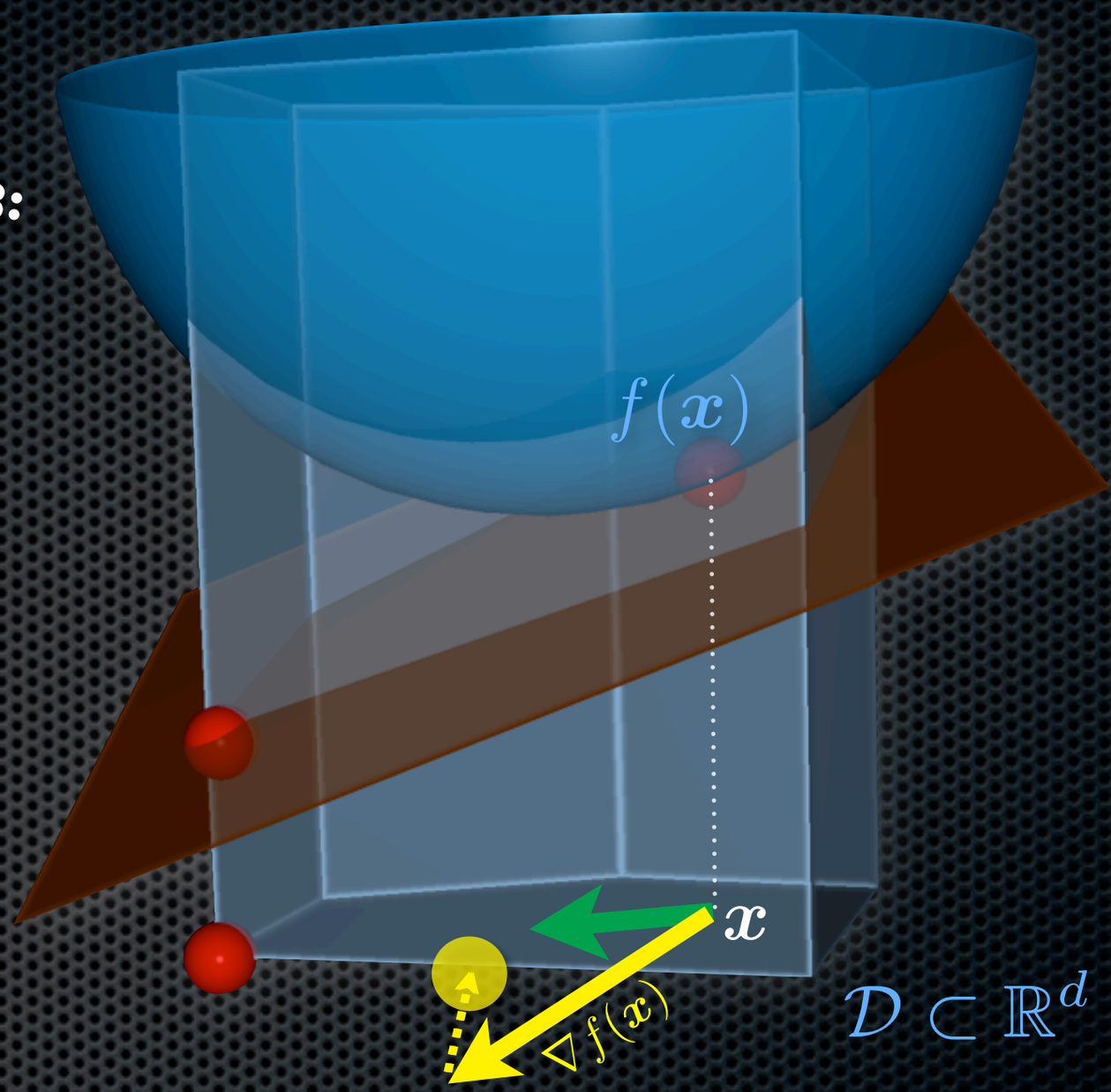
Two kinds of first-order methods



	Frank-Wolfe	***Gradient Descent***
Iteration cost	*(approx.)* solve ***linearized problem*** on D	***projection*** back to D
Iterates	***sparse*** ✓ (in terms of used vertices)	***dense*** ✗

other Frank-Wolfe talks at ICCOPT '13:

Pradeep Ravikumar (Mon),
Federico Pierucci (Tue),
Guanghui Lan (Wed),
Zaid Harchaoui (Wed),
Simon Lacoste-Julien (Wed),
Robert M. Freund (Wed),
Paul Grigas (Wed),
Ronny Luss (Wed),
Dan Garber (Wed),
(and *summer school* by Steve Wright)



	Frank-Wolfe	***Gradient Descent***
Iteration cost	*(approx.)* solve ***linearized problem*** on D	***projection*** back to D
Iterates	***sparse*** ✓ (in terms of used vertices)	***dense*** ✗

Algorithm Variants

Predefined Stepsize

Algorithm 1 Frank-Wolfe

for $k = 0 \dots K$ **do**

Compute $\boldsymbol{s} := \arg\min_{\boldsymbol{s}' \in \mathcal{D}} \left\langle \boldsymbol{s}', \nabla f(\boldsymbol{x}^{(k)}) \right\rangle$

Let $\gamma := \frac{2}{k+2}$

Update $\boldsymbol{x}^{(k+1)} := (1-\gamma)\boldsymbol{x}^{(k)} + \gamma \boldsymbol{s}$

end for

Line-Search

Algorithm 2 Frank-Wolfe

for $k = 0 \dots K$ **do**

Compute $\boldsymbol{s} := \arg\min_{\boldsymbol{s}' \in \mathcal{D}} \left\langle \boldsymbol{s}', \nabla f(\boldsymbol{x}^{(k)}) \right\rangle$

Optimize γ by line-search

Update $\boldsymbol{x}^{(k+1)} := (1-\gamma)\boldsymbol{x}^{(k)} + \gamma \boldsymbol{s}$

end for

Fully Corrective

Algorithm 3 Frank-Wolfe

for $k = 0 \dots K$ **do**

Compute $\boldsymbol{s} := \arg\min_{\boldsymbol{s}' \in \mathcal{D}} \left\langle \boldsymbol{s}', \nabla f(\boldsymbol{x}^{(k)}) \right\rangle$

Update $\boldsymbol{x}^{(k+1)} := \arg\min_{\boldsymbol{x} \in \mathrm{conv}(\boldsymbol{s}^{(0)}, \dots, \boldsymbol{s}^{(k+1)})} f(\boldsymbol{x})$

end for

- Approximate Subproblems

  [Dunn et al. 1978]

- Away-Steps

  [Guélat et al. 1986]

What's new?

[J. 2011 PhD Thesis,
summary: J. 2013 ICML Paper]

- Primal-Dual Analysis

  with ***certificates*** for approximation quality

- Approximate Subproblems

  and ***inexact gradients*** (and domains)

- Affine Invariance

- Lower Bound

  and ***optimality*** in terms of ***sparsity***

- More Applications

Convergence Analysis

Primal Rate

$$f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^*) \leq \frac{2C_f}{k+2}$$

[Frank & Wolfe 1956, Dunn et al. 1978, 1980 : $O\left(\frac{1}{k}\right)$

Primal-Dual

$$g(\boldsymbol{x}^{(\hat{k})}) \leq \frac{7C_f}{k+2}$$

efficient ***certificates*** for approximation quality

[Simplex domain: Clarkson 2008]

Here: F***or all algorithm variants***, and ***affine invariant***

Curvature constant

$$C_f := \sup_{\substack{\boldsymbol{x},\boldsymbol{s}\in\mathcal{D},\\ \gamma\in[0,1],\\ \boldsymbol{y}=\boldsymbol{x}+\gamma(\boldsymbol{s}-\boldsymbol{x})}} \frac{2}{\gamma^2}\big(f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \boldsymbol{y}-\boldsymbol{x}, \nabla f(\boldsymbol{x})\rangle\big)$$

$$C_f \leq \operatorname{diam}_{\|.\|}(\mathcal{D})^2 L$$

A Simple Duality Gap

Original Problem

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

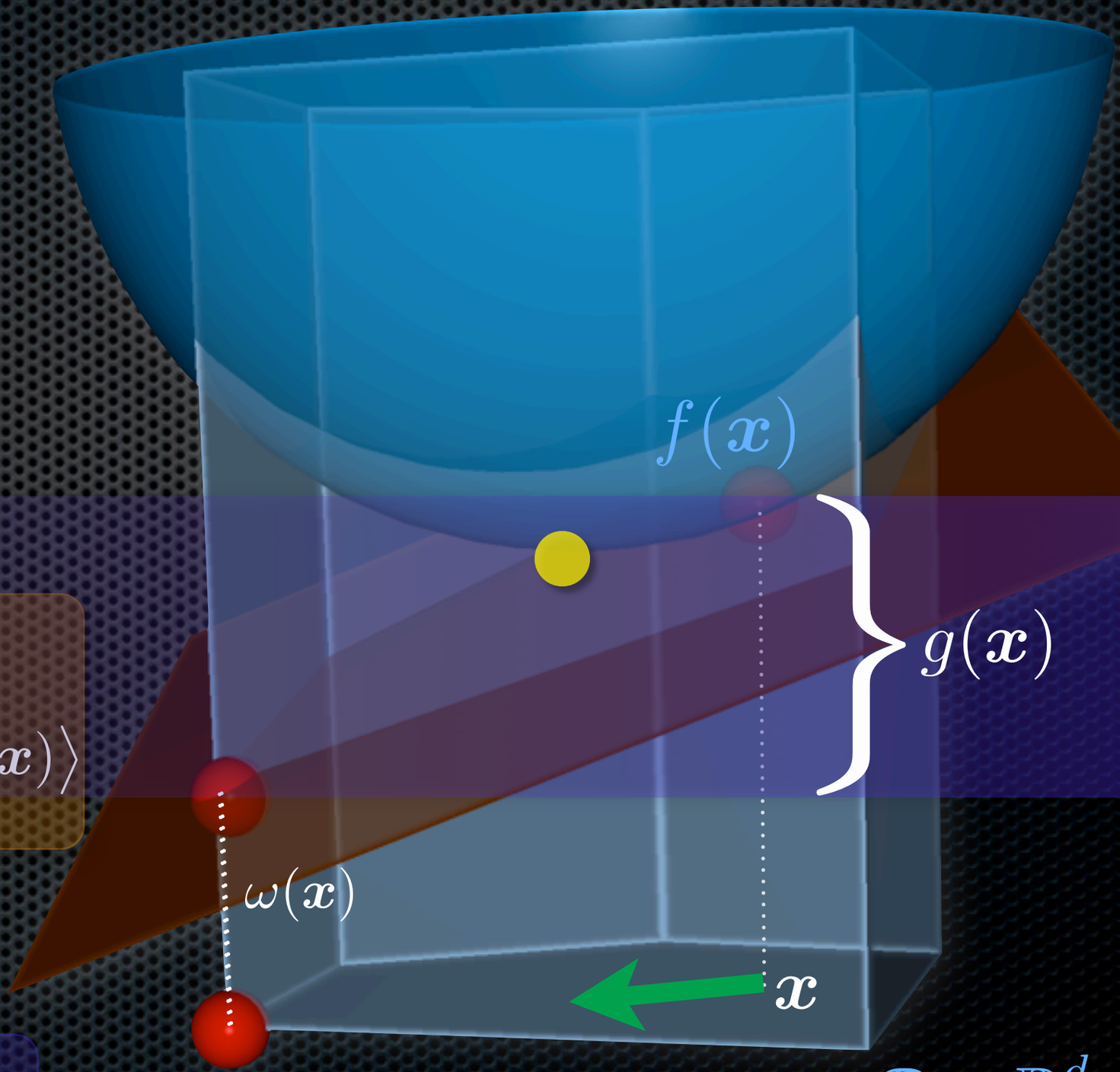
The Dual Value

$$\omega(\boldsymbol{x}) :=$$

$$\min_{\boldsymbol{s}' \in \mathcal{D}} f(\boldsymbol{x}) + \langle \boldsymbol{s}' - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle$$

Weak Duality

$$\omega(\boldsymbol{x}) \leq f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}')$$



Affine Invariance

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

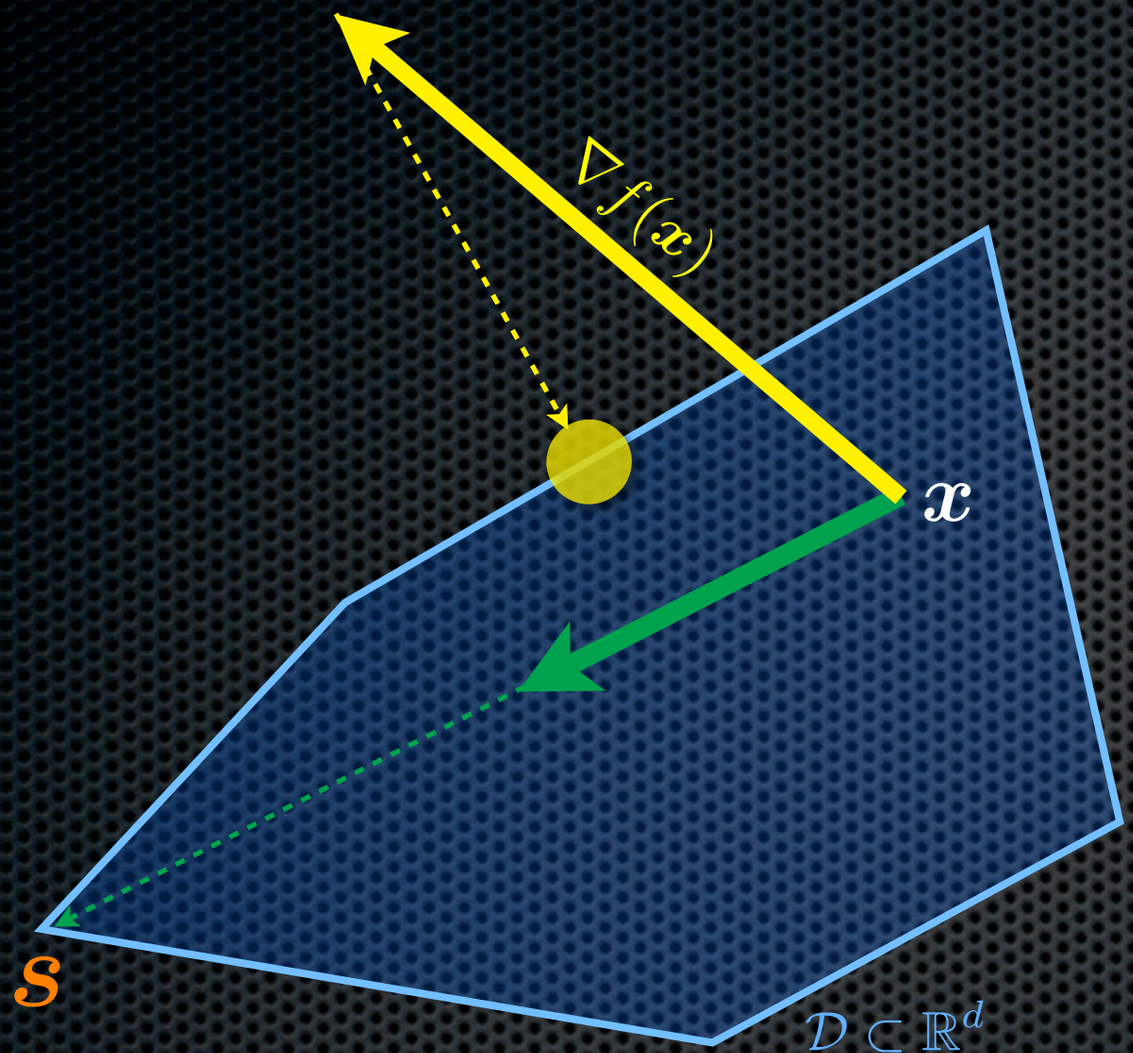
Convergence

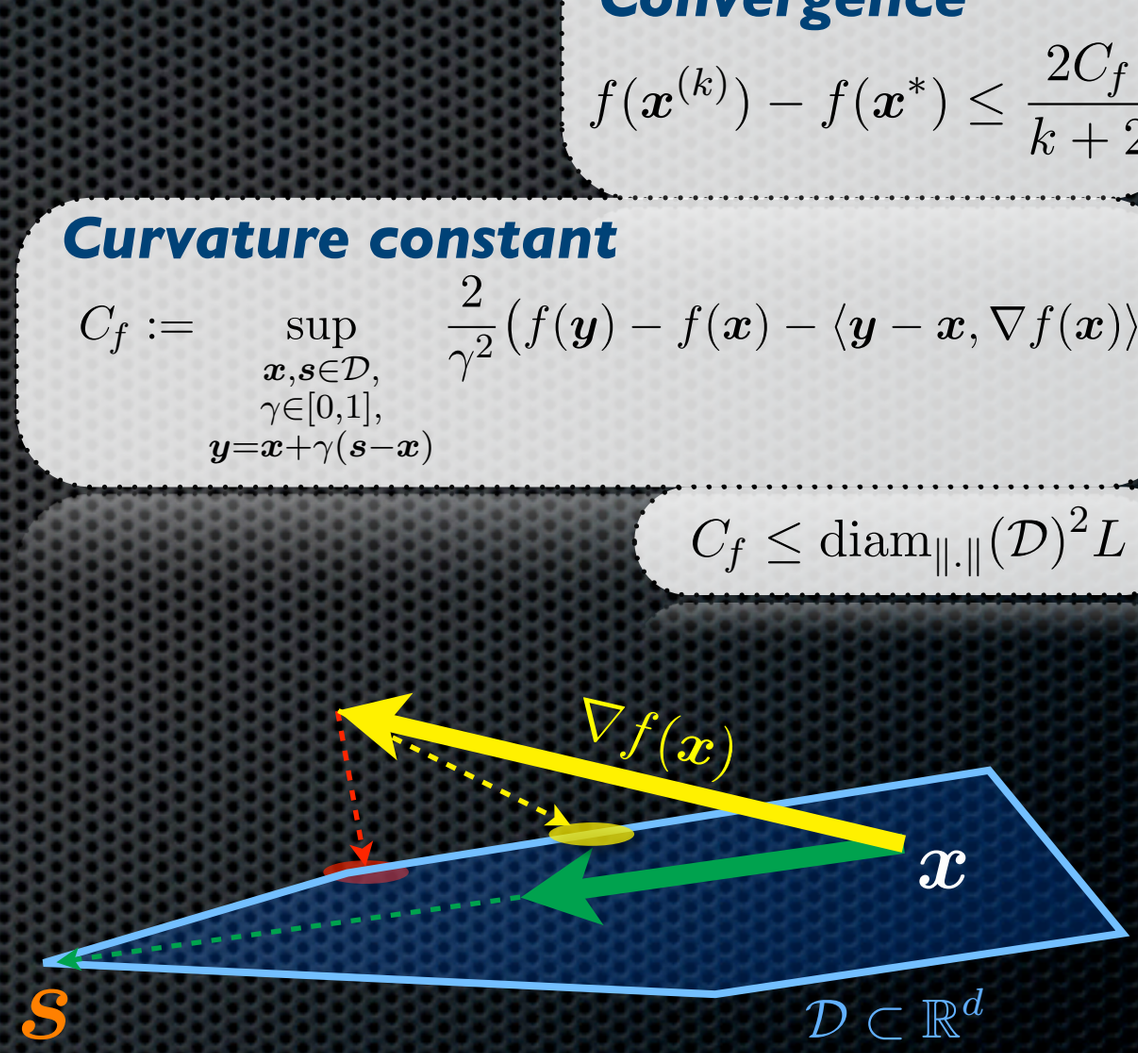
$$f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^*) \leq \frac{2C_f}{k+2}$$

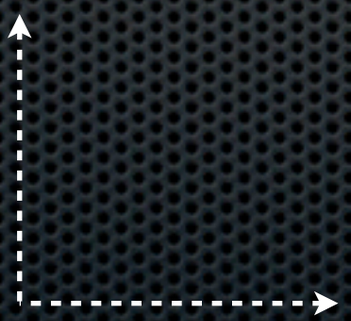
Curvature constant

$$C_f := \sup_{\substack{\boldsymbol{x},\boldsymbol{s}\in\mathcal{D},\\ \gamma\in[0,1],\\ \boldsymbol{y}=\boldsymbol{x}+\gamma(\boldsymbol{s}-\boldsymbol{x})}} \frac{2}{\gamma^2}\big(f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \boldsymbol{y}-\boldsymbol{x}, \nabla f(\boldsymbol{x})\rangle\big)$$

$$C_f \leq \mathrm{diam}_{\|.\|}(\mathcal{D})^2 L$$







Lower Bound

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

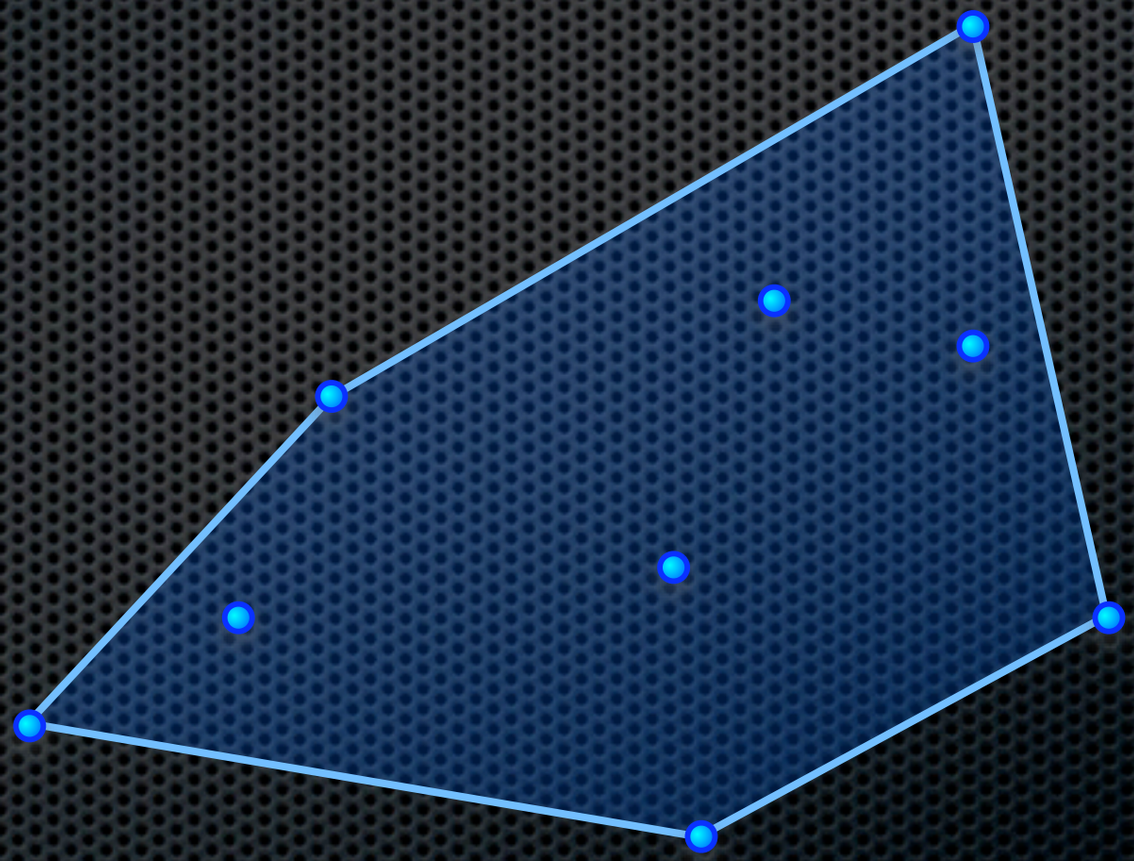
For an approximation ***quality of 1/k***,
any *x* must have ***at least k non-zeros***.
(use ***at least k corners***)

Trade-Off: ***Approximation quality*** ***vs*** ***sparsity***

Optimization over Atomic Sets

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

convex hull of ***things*** $\mathcal{A}$



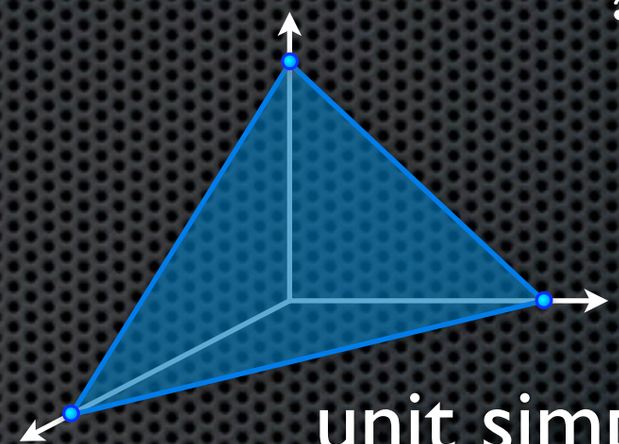
$$\mathcal{D} := \operatorname{conv}(\mathcal{A})$$

[Chandrasekaran et al. 2012]

Sparse Approximation

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

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



unit simplex

Corollary:
Obtain $O\left(\frac{1}{k}\right)$ -approximate solution of **sparsity** k .

[Clarkson 2008]

Trade-Off:
Approximation quality vs **sparsity**

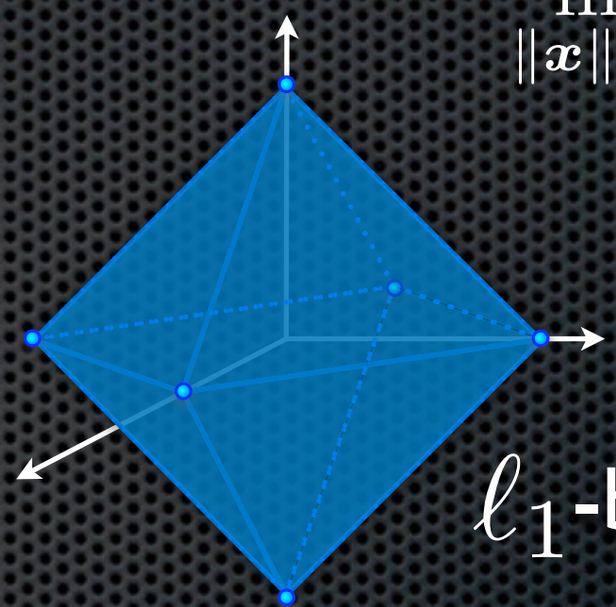
lower bound:
$\Omega\left(\frac{1}{k}\right)$

$f(\boldsymbol{x}) := \|\boldsymbol{x}\|_2^2$

Sparse Approximation

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

$$\mathcal{D} := \operatorname{conv}\left(\{\pm \mathbf{e}_i \mid i \in [n]\}\right)$$



ℓ_1-ball

Corollary:
Obtain $O\left(\frac{1}{k}\right)$-approximate solution of **sparsity** k.

Trade-Off:
Approximation quality vs **sparsity**

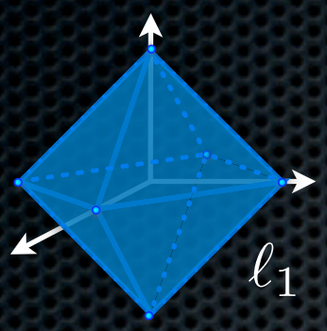
lower bound:
$\Omega\left(\frac{1}{k}\right)$

Greedy meets Frank-Wolfe

Convex optimization

methods applied to

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



Frank-Wolfe

fully corrective Frank-Wolfe

Signal processing

sparse recovery methods

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

selects the same atom per step

matching pursuit

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

equivalent to

OMP

Low Rank Approximation

$$\min_{\|X\|_* \leq 1} f(X)$$

$$\mathcal{D} := \operatorname{conv}\left(\left\{ \boldsymbol{u}\boldsymbol{v}^T \;\middle|\; \begin{array}{l} \boldsymbol{u}\in\mathbb{R}^n,\ \|\boldsymbol{u}\|_2=1 \\ \boldsymbol{v}\in\mathbb{R}^m,\ \|\boldsymbol{v}\|_2=1 \end{array} \right\}\right)$$

trace-norm-ball

Corollary:
Obtain $O\left(\frac{1}{k}\right)$ -approximate solution of **rank** k .

lower bound:
$\Omega\left(\frac{1}{k}\right)$

Trade-Off:
Approximation quality vs **rank**

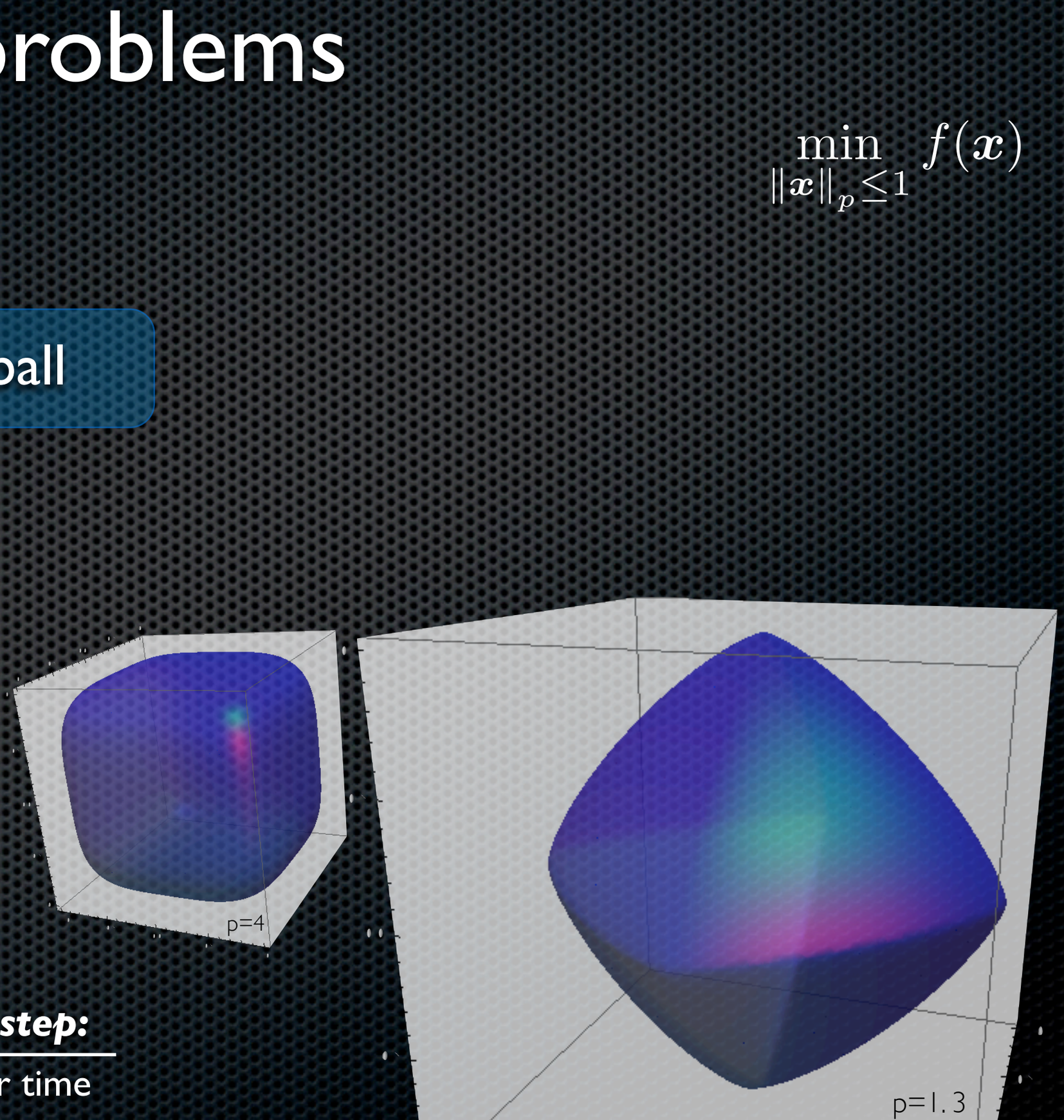
Projection:	***FW-step:***
Requires full SVD	approx. top singular vector

[J. & Sulovský 2010]

ℓ_p-norm problems

$$\min_{\|\boldsymbol{x}\|_p \leq 1} f(\boldsymbol{x})$$

$\mathcal{D} :=$ ℓ_p-ball



Projection:	***FW-step:***
unknown?	linear time

Examples of *Atomic Domains* Suitable for Frank-Wolfe

$\mathcal{X}$	**Optimization Domain** Atoms $\mathcal{A}$	$\mathcal{D} = \text{conv}(\mathcal{A})$	**Complexity of one Frank-Wolfe Iteration** $\sup_{\boldsymbol{s}\in\mathcal{D}}\langle \boldsymbol{s}, \boldsymbol{y}\rangle$	Complexity
$\mathbb{R}^n$	Sparse Vectors	$\|.\|_1$-ball	$\|\boldsymbol{y}\|_\infty$	$O(n)$
$\mathbb{R}^n$	Sign-Vectors	$\|.\|_\infty$-ball	$\|\boldsymbol{y}\|_1$	$O(n)$
$\mathbb{R}^n$	ℓ_p-Sphere	$\|.\|_p$-ball	$\|\boldsymbol{y}\|_q$	$O(n)$
$\mathbb{R}^n$	Sparse Non-neg. Vectors	Simplex Δ_n	$\max_i\{\boldsymbol{y}_i\}$	$O(n)$
$\mathbb{R}^n$	Latent Group Sparse Vec.	$\|.\|_\mathcal{G}$-ball	$\max_{g\in\mathcal{G}} \left\|\boldsymbol{y}_{(g)}\right\|_g^*$	$\sum_{g\in\mathcal{G}} \|g\|$
$\mathbb{R}^{m\times n}$	Matrix Trace Norm	$\|.\|_{tr}$-ball	$\|\boldsymbol{y}\|_{op} = \sigma_1(\boldsymbol{y})$	$\tilde{O}\big(N_f/\sqrt{\varepsilon'}\big)$ (Lanczos)
$\mathbb{R}^{m\times n}$	Matrix Operator Norm	$\|.\|_{op}$-ball	$\|\boldsymbol{y}\|_{tr} = \|(\sigma_i(\boldsymbol{y}))\|_1$	SVD
$\mathbb{R}^{m\times n}$	Schatten Matrix Norms	$\|(\sigma_i(.))\|_p$-ball	$\|(\sigma_i(\boldsymbol{y}))\|_q$	SVD
$\mathbb{R}^{m\times n}$	Matrix Max-Norm	$\|.\|_{\max}$-ball		$\tilde{O}\big(N_f(n+m)^{1.5}/\varepsilon'^{2.5}\big)$
$\mathbb{R}^{n\times n}$	Permutation Matrices	Birkhoff polytope		$O(n^3)$
$\mathbb{R}^{n\times n}$	Rotation Matrices			SVD (Procrustes prob.)
$\mathbb{S}^{n\times n}$	Rank-1 PSD matrices of unit trace	$\{\boldsymbol{x}\succeq 0,\ \text{Tr}(\boldsymbol{x})=1\}$	$\lambda_{\max}(\boldsymbol{y})$	$\tilde{O}\big(N_f/\sqrt{\varepsilon'}\big)$ (Lanczos)
$\mathbb{S}^{n\times n}$	PSD matrices of bounded diagonal	$\{\boldsymbol{x}\succeq 0,\ \boldsymbol{x}_{ii}\leq 1\}$		$\tilde{O}\big(N_f\, n^{1.5}/\varepsilon'^{2.5}\big)$

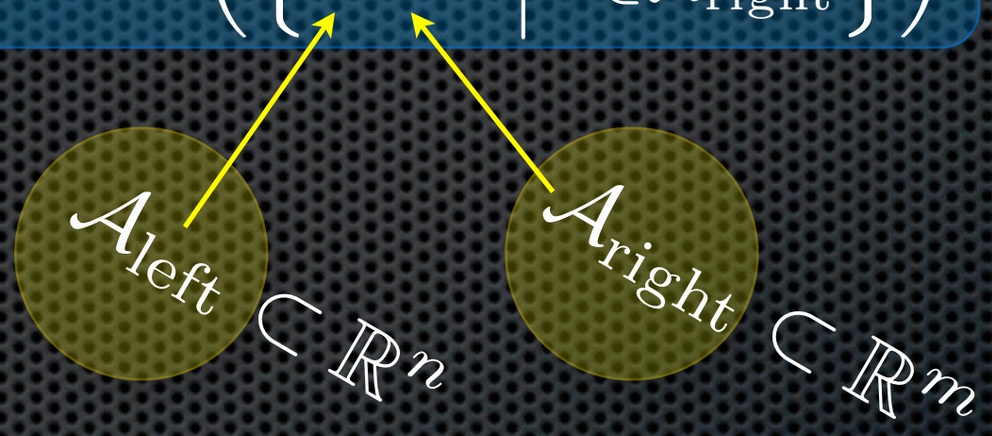
Table 1: Some examples of atomic domains suitable for optimization using the Frank-Wolfe algorithm. *Here SVD refers to the complexity of computing a singular value decomposition, which is* $O(\min\{mn^2, m^2n\})$. *N_f is the number of non-zero entries in the gradient of the objective function f, and $\varepsilon' = \frac{2\delta C_f}{k+2}$ is the required accuracy for the linear subproblems. For any $p \in [1,\infty]$ the conjugate value q is meant to satisfy $\frac{1}{p} + \frac{1}{q} = 1$, allowing $q = \infty$ for $p = 1$.*

Factorized Matrix Domains

$$\mathcal{D} := \operatorname{conv}\left(\left\{ \boldsymbol{u}\boldsymbol{v}^T \;\middle|\; \begin{matrix} \boldsymbol{u}\in\mathbb{R}^n,\ \|\boldsymbol{u}\|_2=1 \\ \boldsymbol{v}\in\mathbb{R}^m,\ \|\boldsymbol{v}\|_2=1 \end{matrix} \right\}\right)$$

(trace norm)

$$\mathcal{D} := \operatorname{conv}\left(\left\{ \boldsymbol{u}\boldsymbol{v}^T \;\middle|\; \begin{matrix} \boldsymbol{u}\in\mathcal{A}_{\text{left}} \\ \boldsymbol{v}\in\mathcal{A}_{\text{right}} \end{matrix} \right\}\right)$$



r	$\mathcal{A}_{\text{left}} \subseteq \mathbb{R}^{m\times r}$	$\mathcal{A}_{\text{right}} \subseteq \mathbb{R}^{n\times r}$	$\Omega_{\text{conv}(\mathcal{A})}(M)$	$\Omega^*_{\mathcal{A}}(M)$	**FW step**
1	$\|.\|_2$-sphere	$\|.\|_2$-sphere	Trace norm $\|M\|_{tr}$	$\|M\|_{op}$	Lanczos, see Table 1
1	$\|.\|_1$-sphere	$\|.\|_1$-sphere	Vector ℓ_1-norm $\|\vec{M}\|_1$	$\|\vec{M}\|_\infty$	$O(nm)$
1	$\|.\|_\infty$-sphere	$\|.\|_\infty$-sphere		Cut-norm $\|.\|_{\infty\to 1}$	NP-hard (Alon & Naor, 2006)
n+m	$\|.\|_{2,\infty}$	$\|.\|_{2,\infty}$	Max-norm $\|M\|_{\max}$		SDP, see Table 1
1	$\|.\|_2 \cap \mathbb{R}^m_{\geq 0}$	$\|.\|_2 \cap \mathbb{R}^n_{\geq 0}$	"non-neg. trace norm"		NP-hard (Murty & Kabadi, 1987)
1	Simplex Δ_m	Simplex Δ_n	"non-neg. matrix ℓ_1-norm"		$O(nm)$

Table 2. Examples of some factorized matrix norms on $\mathbb{R}^{m\times n}$, each induced by two atomic norms (last two rows giving

Active Research

- Faster Convergence for Strongly Convex f under additional assumptions
  [Talk by Dan Garber, Guélat&Marcotte 1986, Beck&Teboulle 2004]

- Penalized Version $\min_{\boldsymbol{x}} f(\boldsymbol{x}) + \lambda \|\boldsymbol{x}\|_{\mathcal{A}}$
  [Talk by Zaid Harchaoui]

- Non-Smooth f
  [Talks by Federico Pierucci and Guanghui Lan]

- Block-Wise Version $\min_{\boldsymbol{x} \in \mathcal{D}^{(1)} \times \cdots \times \mathcal{D}^{(n)}} f(\boldsymbol{x})$
  [Talk by Simon Lacoste-Julien] $\boldsymbol{x} = (\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(n)})$

- Find More Applications!

Thanks