

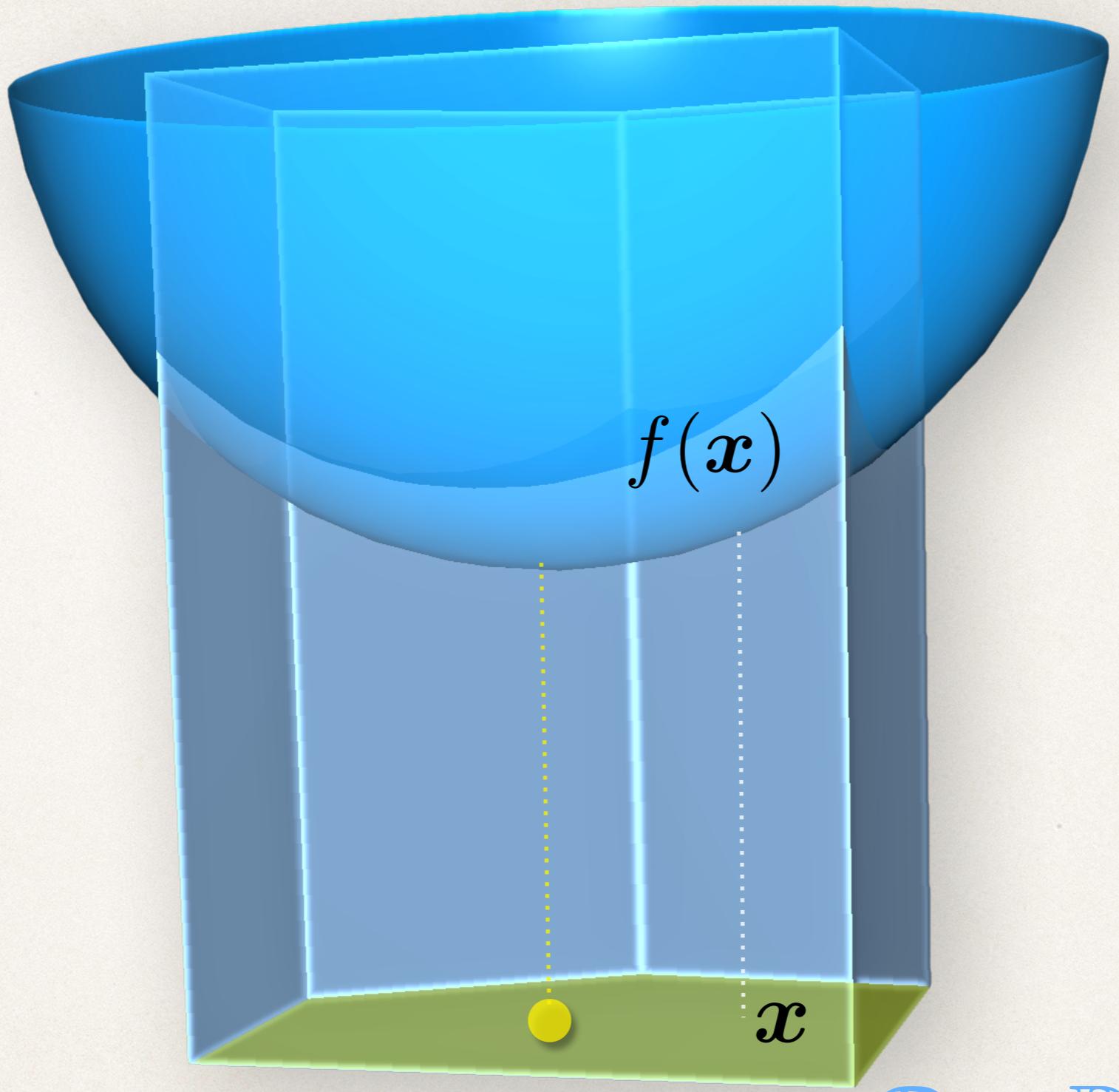
Frank-Wolfe and Greedy Optimization for Learning with Big Data

Part I: **Martin Jaggi**
ETH Zurich

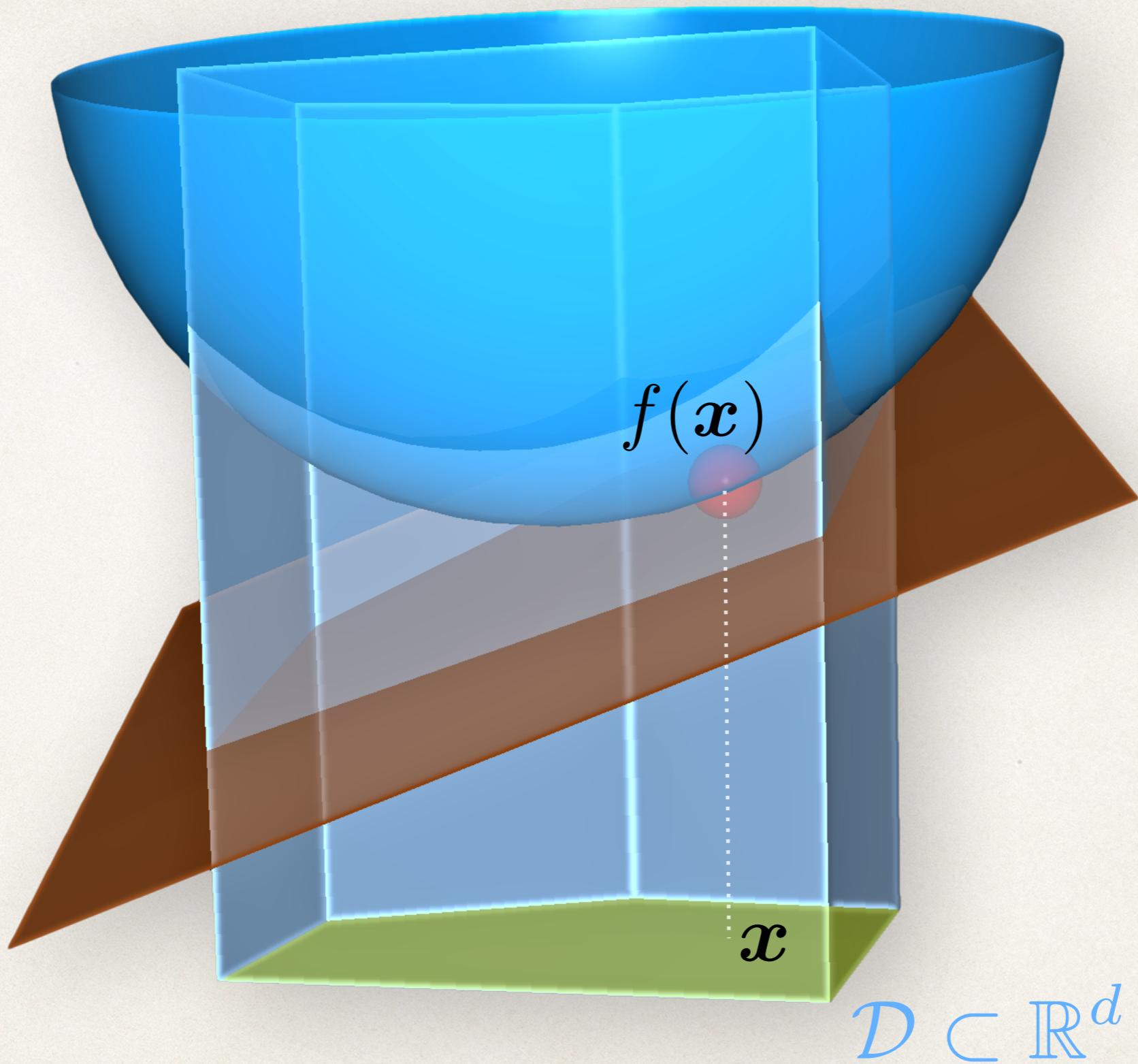
Part II: **Zaid Harchaoui**
Inria

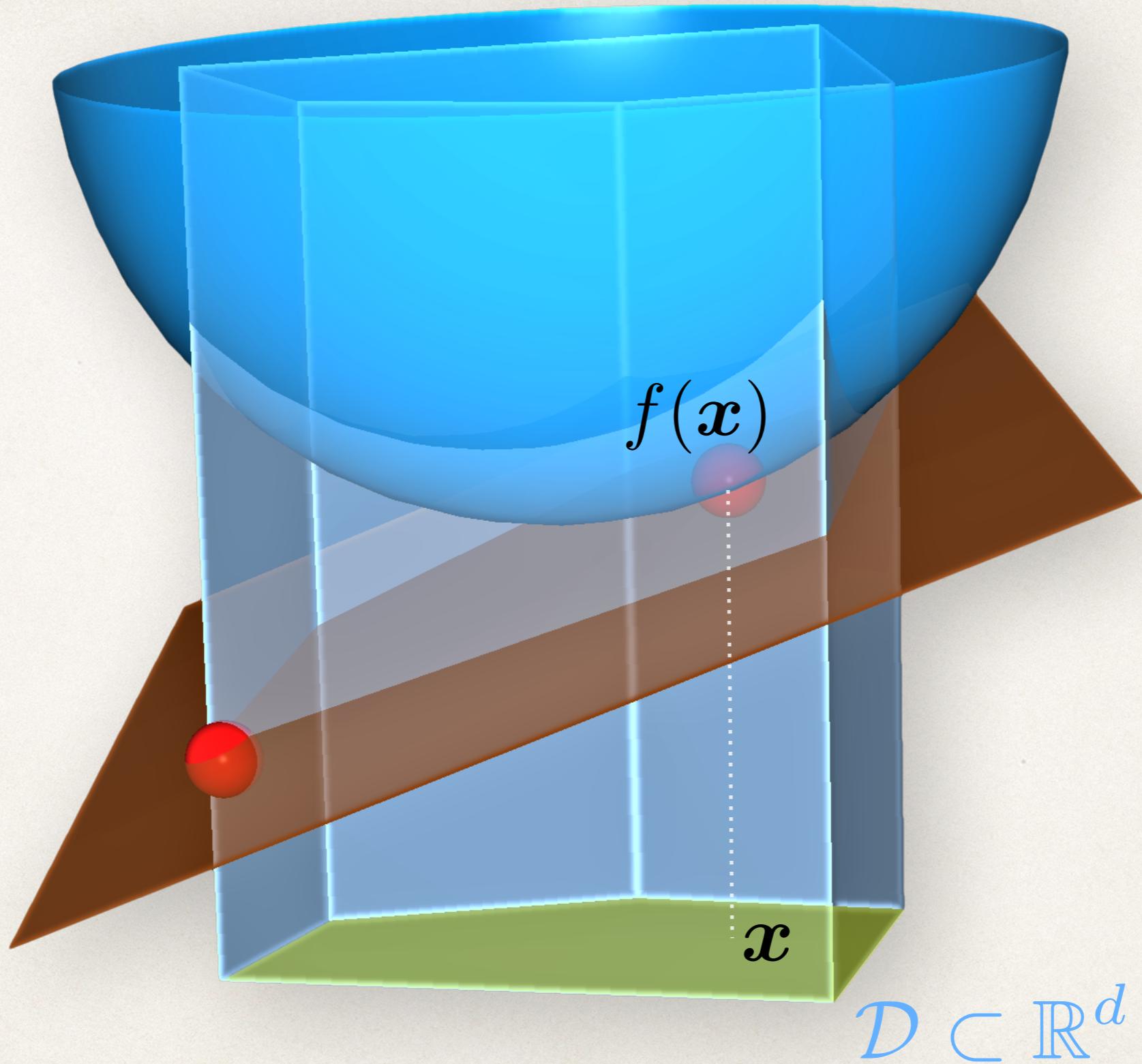


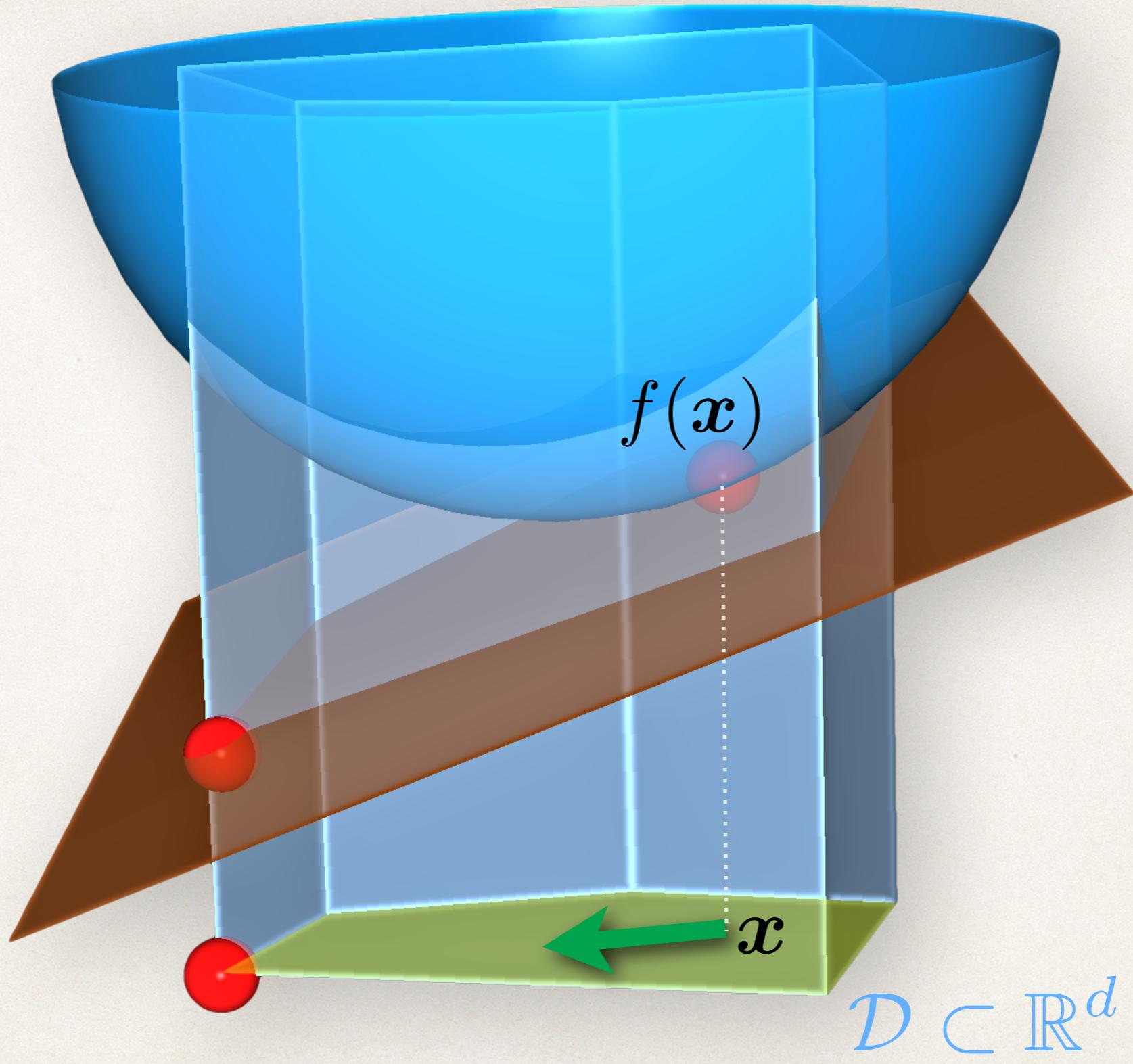
$$\mathcal{D} \subset \mathbb{R}^d$$



$$\mathcal{D} \subset \mathbb{R}^d$$







Tutorial Outline

- ✿ **Part I:** (*Martin Jaggi*)

- ✿ Frank-Wolfe basics and history
- ✿ convergence analysis and geometry independence
- ✿ optimality & lower bounds
- ✿ demo and applications for atomic domains
- ✿ faster rates under additional assumptions

- ✿ **Part II:** (*Zaid Harchaoui*)

- ✿ Greedy Optimization and Frank-Wolfe
- ✿ Frank-Wolfe methods for:
 - ✿ composite optimization
 - ✿ non-smooth optimization
 - ✿ block-structured optimization

Tutorial Website

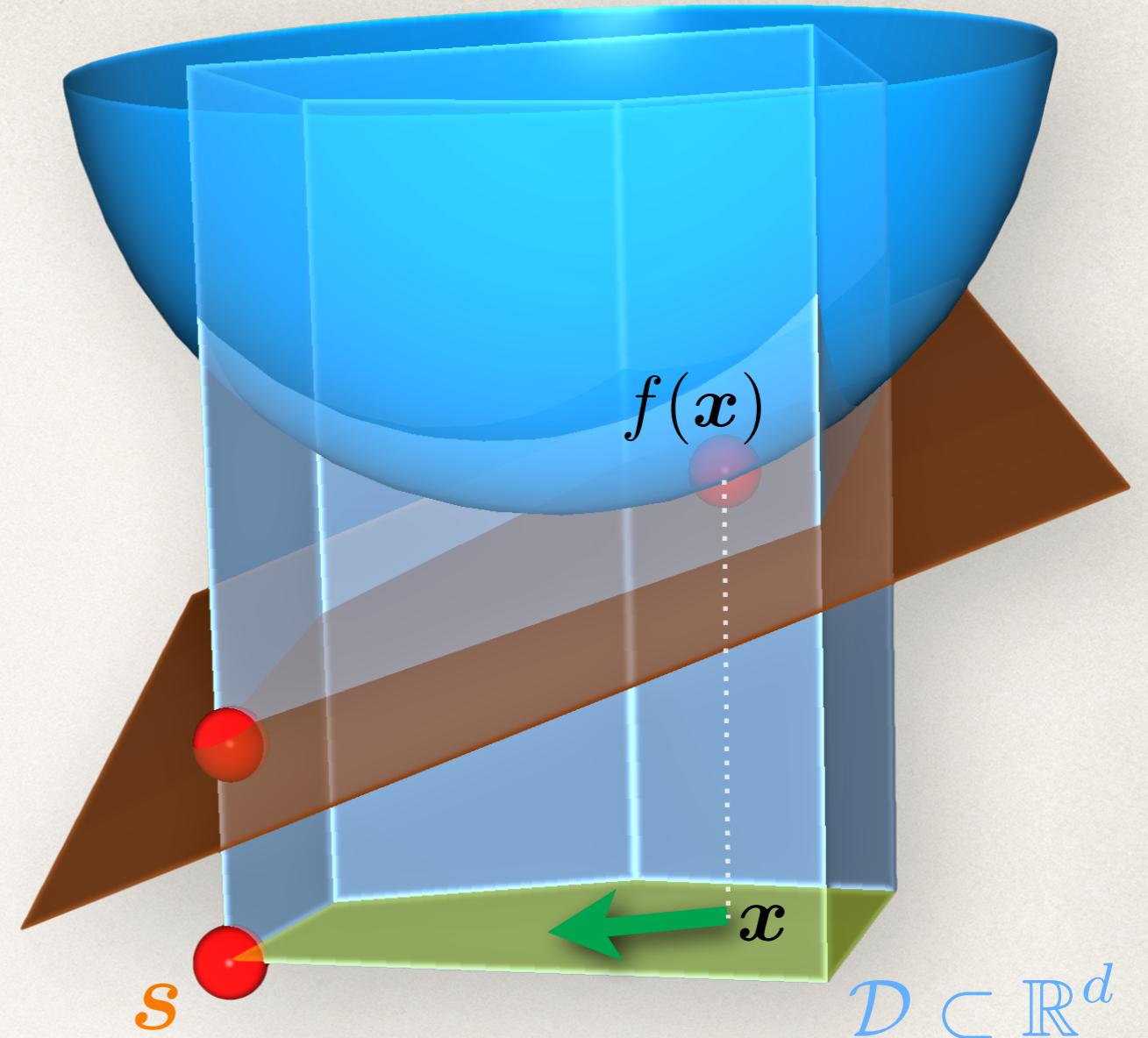
- ✿ Sample Code
- ✿ Bibliography
- ✿ Tutorial Slides
- ✿ Topics not covered

<https://sites.google.com/site/FrankWolfeGreedyTutorial>

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

The Linear Minimization Oracle

$$\text{LMO}_{\mathcal{D}}(\mathbf{d}) := \arg \min_{\mathbf{s} \in \mathcal{D}} \langle \mathbf{d}, \mathbf{s} \rangle$$



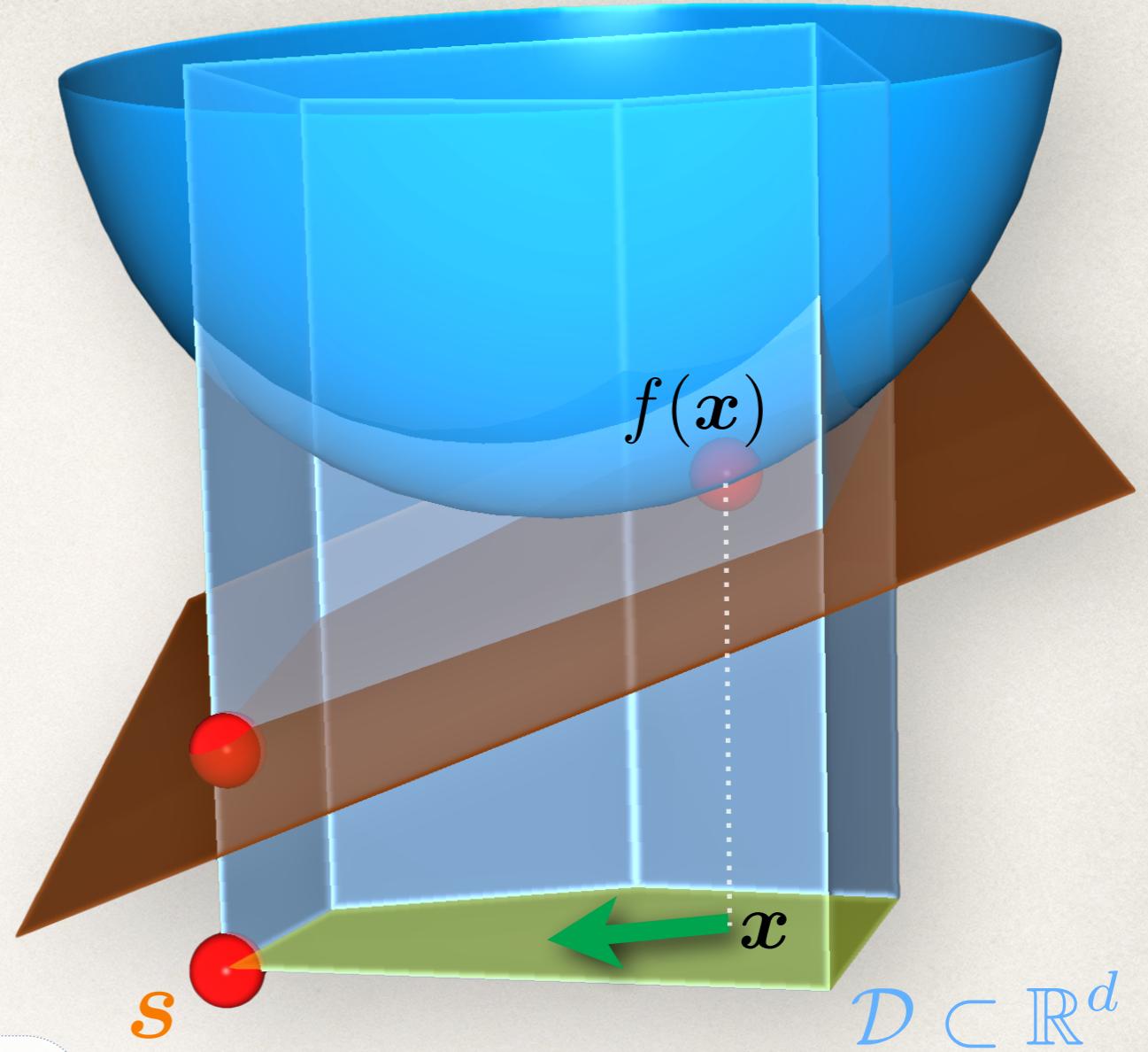
Linearization

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

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

The Linear Minimization Oracle

$$\text{LMO}_{\mathcal{D}}(\mathbf{d}) := \arg \min_{\mathbf{s} \in \mathcal{D}} \langle \mathbf{d}, \mathbf{s} \rangle$$



Algorithm 1: Frank-Wolfe

Let $\mathbf{x}^{(0)} \in \mathcal{D}$

for $t = 0 \dots T$ **do**

 Compute $\mathbf{s} := \text{LMO}_{\mathcal{D}}(\nabla f(\mathbf{x}^{(t)}))$

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

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

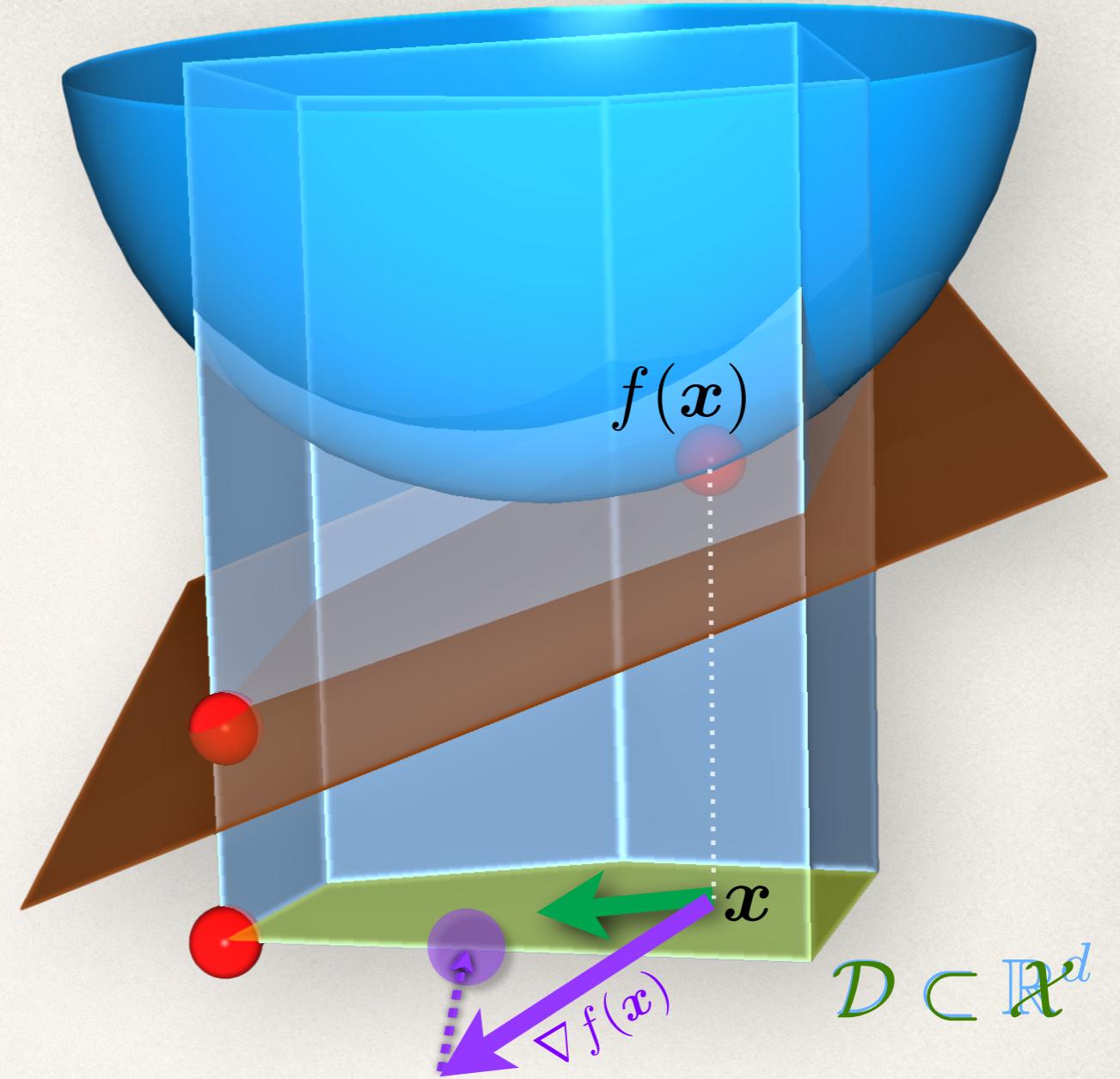
end

Convergence

$$f(\mathbf{x}^{(t)}) - f^* \leq O(1/t)$$

convexity, smoothness of f , boundedness of \mathcal{D}

Two kinds of first-order methods

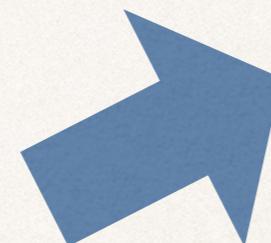


	Frank-Wolfe	Gradient Descent and Proximal Methods
Iteration cost	(approx.) solve <i>linear problem</i> on D	<i>projection</i> back to D (or prox operator)
Iterates	<i>sparse</i> ✓ (in terms of used vertices)	<i>dense</i> ✗
∞ -dim domain	<i>yes</i> ✓ (if oracle available)	<i>no</i> ✗

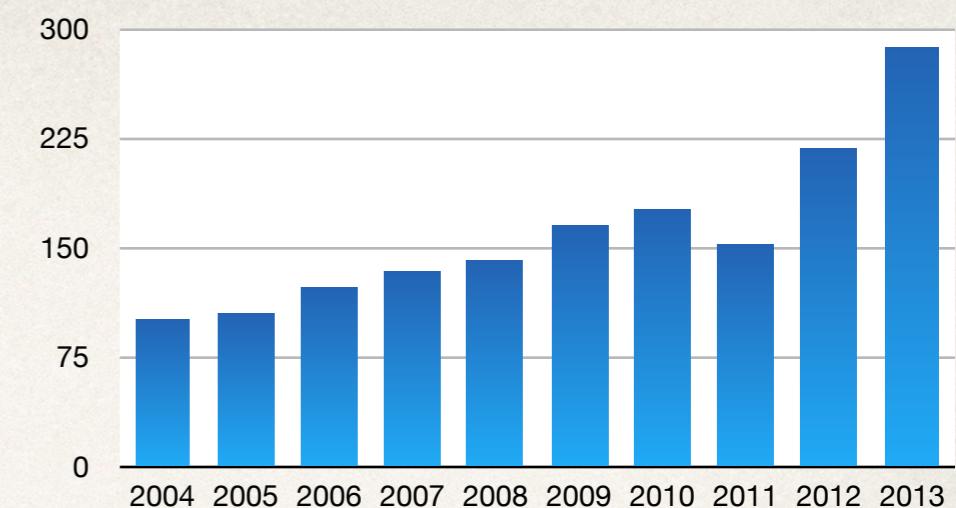
Origins

of Frank-Wolfe /
Conditional Gradient

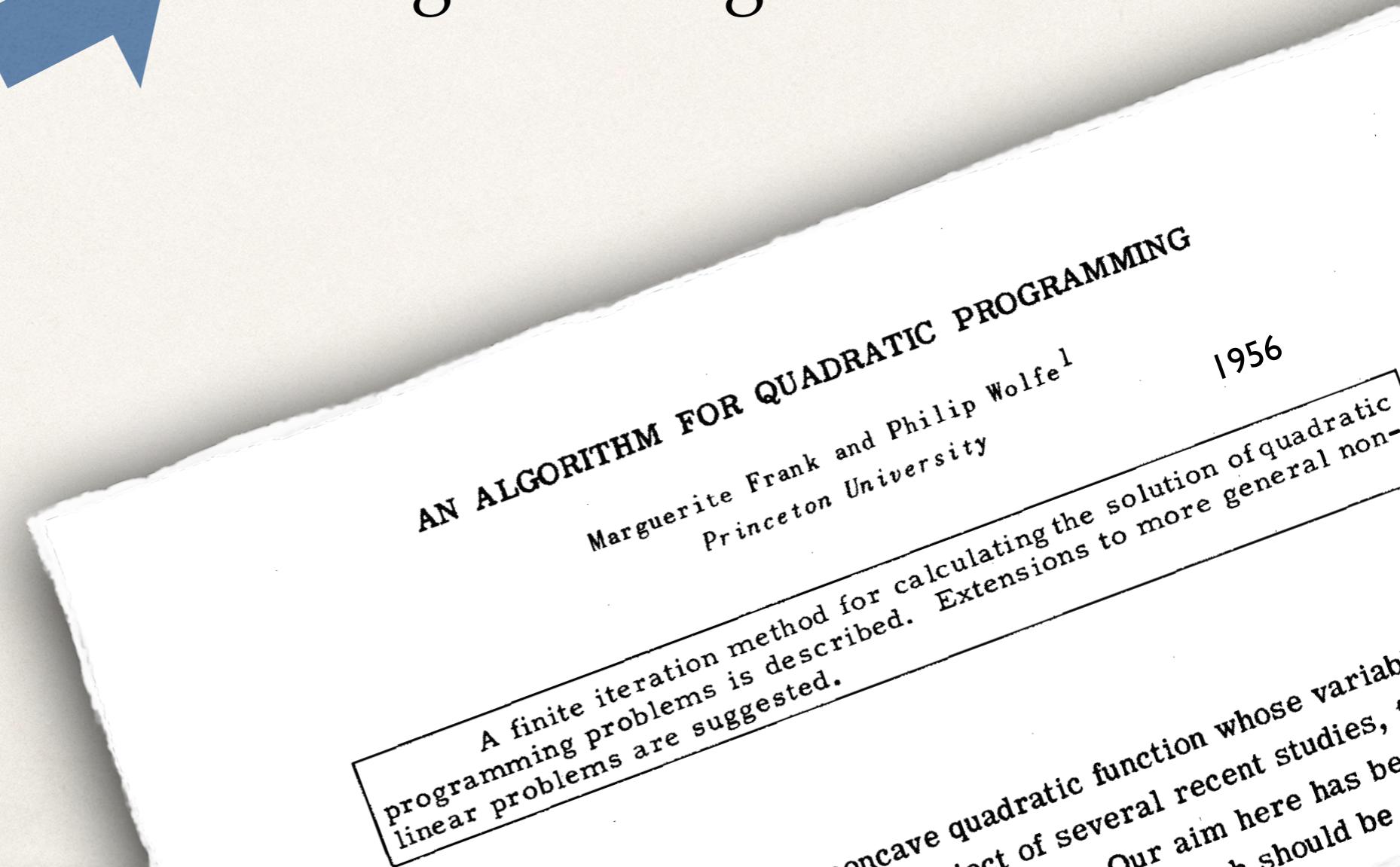
Linear
Programming



Convex
Programming



“Marguerite Frank”



Numerous Applications

- ❖ network flows / transportation problems
- ❖ greedy selection and sparse optimization
- ❖ low-rank matrix factorizations, many other matrices
- ❖ wavelets (infinite-dimensional)
- ❖ structured sparsity and structured prediction
- ❖ total-variation-norm for image de-noising
- ❖ submodular optimization
- ❖ boosting
- ❖ training deep networks

Variants of Frank-Wolfe

✿ Simple Step-Size

Algorithm 1: Frank-Wolfe

Let $\mathbf{x}^{(0)} \in \mathcal{D}$

for $t = 0 \dots T$ **do**

 Compute $\mathbf{s} := \text{LMO}_{\mathcal{D}}(\nabla f(\mathbf{x}^{(t)}))$

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

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

end

✿ Line-Search

Optimize γ by line-search

$$\gamma := \arg \min_{\gamma \in [0,1]} f((1 - \gamma)\mathbf{x}^{(t)} + \gamma\mathbf{s})$$

✿ Fully Corrective / Restricted Simplicial Decomposition

$$\text{Update } \mathbf{x}^{(t+1)} := \arg \min_{\mathbf{x} \in \text{conv}(\mathbf{s}^{(0)}, \dots, \mathbf{s}^{(t+1)})} f(\mathbf{x})$$

[Holloway:1974]

Approximate Oracles and Inexact Gradients

$$\varepsilon_t := \frac{C_f}{2+t}$$

Algorithm 1: Frank-Wolfe

```

Let  $\mathbf{x}^{(0)} \in \mathcal{D}$ 
for  $t = 0 \dots T$  do
    Compute  $\mathbf{s} := \text{LMO}_{\mathcal{D}}(\nabla f(\mathbf{x}^{(t)}))$ 
    Let  $\gamma := \frac{2}{t+2}$ 
    Update  $\mathbf{x}^{(t+1)} := (1 - \gamma)\mathbf{x}^{(t)} + \gamma\mathbf{s}$ 
end
```

⊕ Approximate Oracle

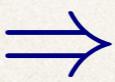
$\text{LMO}_{\mathcal{D}, \varepsilon_t}(\mathbf{d})$: returns \mathbf{s} s.t.

$$\langle \mathbf{d}, \mathbf{s} \rangle \leq \min_{\mathbf{s} \in \mathcal{D}} \langle \mathbf{d}, \mathbf{s} \rangle + \varepsilon_t$$

⊕ Inexact Gradients

$\text{LMO}_{\mathcal{D}}(\hat{\mathbf{d}}_x)$

$$\mathbb{E} \left\| \hat{\mathbf{d}}_x - \mathbf{d}_x \right\|^* \leq \frac{\varepsilon_t}{\text{diam}_{\|\cdot\|}(\mathcal{D})}$$



$$\mathbb{E} [\left| \langle \hat{\mathbf{d}}_x - \mathbf{d}_x, \mathbf{s}' - \mathbf{s}'' \rangle \right|] \leq \varepsilon_t \quad \forall \mathbf{s}', \mathbf{s}'' \in \mathcal{D}$$

[Jaggi 2011, 2013]

where $\mathbf{d}_x := \nabla f(\mathbf{x})$ is the true gradient.

[Freund & Grigas 2013]

Same Convergence Bounds!

Convergence Analysis

Primal Rate

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

Primal-Dual

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

efficient *certificates* for
approximation quality

[Frank & Wolfe 1956,
Demyanov & Rubinov 1967
Dunn et al. 1978, 1980]

[asymptotically: Demyanov & R. 1967
simplex domain: Clarkson 2008
general: Jaggi 2011,2013]

Theorem 1 (Primal Convergence Rate). *For each $t \geq 1$, the iterates $\mathbf{x}^{(t)}$ of the Frank-Wolfe algorithm satisfy*

$$f(\mathbf{x}^{(t)}) - f^* \leq \frac{2C_f}{t+2} .$$

Proof. Let C_f be a constant s.t.

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \underbrace{\gamma \langle \mathbf{s} - \mathbf{x}, \nabla f(\mathbf{x}) \rangle}_{-g(\mathbf{x})} + \frac{\gamma^2}{2} C_f$$

for all $\mathbf{x}, \mathbf{s} \in \mathcal{D}, \mathbf{y} := \mathbf{x} + \gamma(\mathbf{s} - \mathbf{x}), \gamma \in [0, 1]$

Writing $h(\mathbf{x}^{(t)}) := f(\mathbf{x}^{(t)}) - f^*$ for the objective error, we have

$$\begin{aligned} h(\mathbf{x}^{(t+1)}) &\leq h(\mathbf{x}^{(t)}) - \gamma g(\mathbf{x}^{(t)}) + \frac{\gamma^2}{2} C_f && \text{definition of } C_f \\ &\leq h(\mathbf{x}^{(t)}) - \gamma h(\mathbf{x}^{(t)}) + \frac{\gamma^2}{2} C_f && h \leq g \text{ by convexity} \\ &= (1 - \gamma)h(\mathbf{x}^{(t)}) + \frac{\gamma^2}{2} C_f , \end{aligned}$$

From here, the decrease rate follows from a simple induction as in Lemma 2. □

Lemma 2. *Suppose a sequence of numbers h_t satisfies*

$$h_{t+1} \leq (1 - \gamma^{(t)})h_t + \gamma^{(t)} C ,$$

for $\gamma^{(t)} = \frac{2}{t+2}$, and $t = 0, 1, \dots$, and a constant C . Then

$$h_t \leq \frac{4C}{t+2} \quad t = 0, 1, \dots$$

A Simple Duality Gap

Original Problem

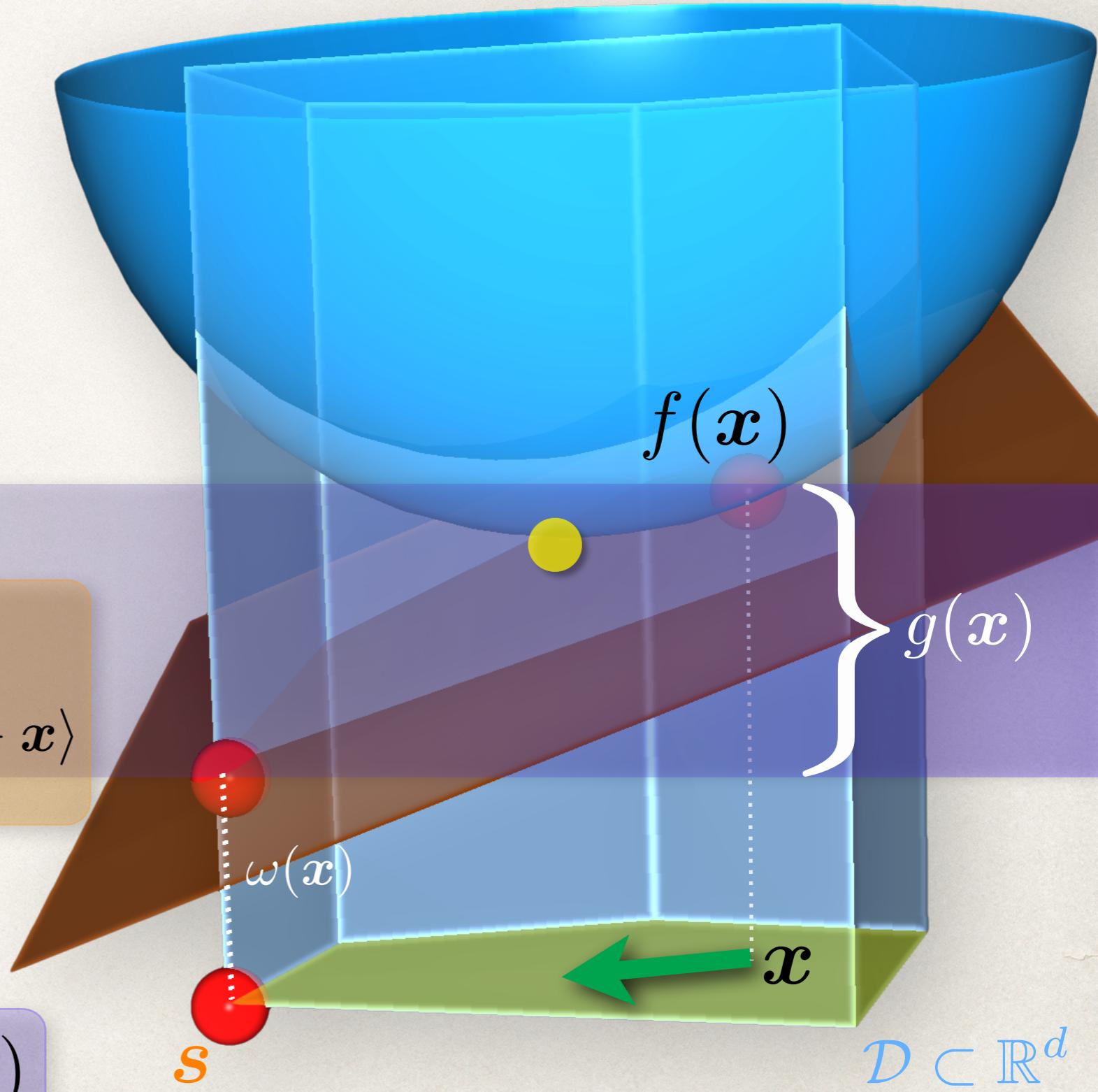
$$\min_{\mathbf{x} \in \mathcal{D}} f(\mathbf{x})$$

The Dual Function

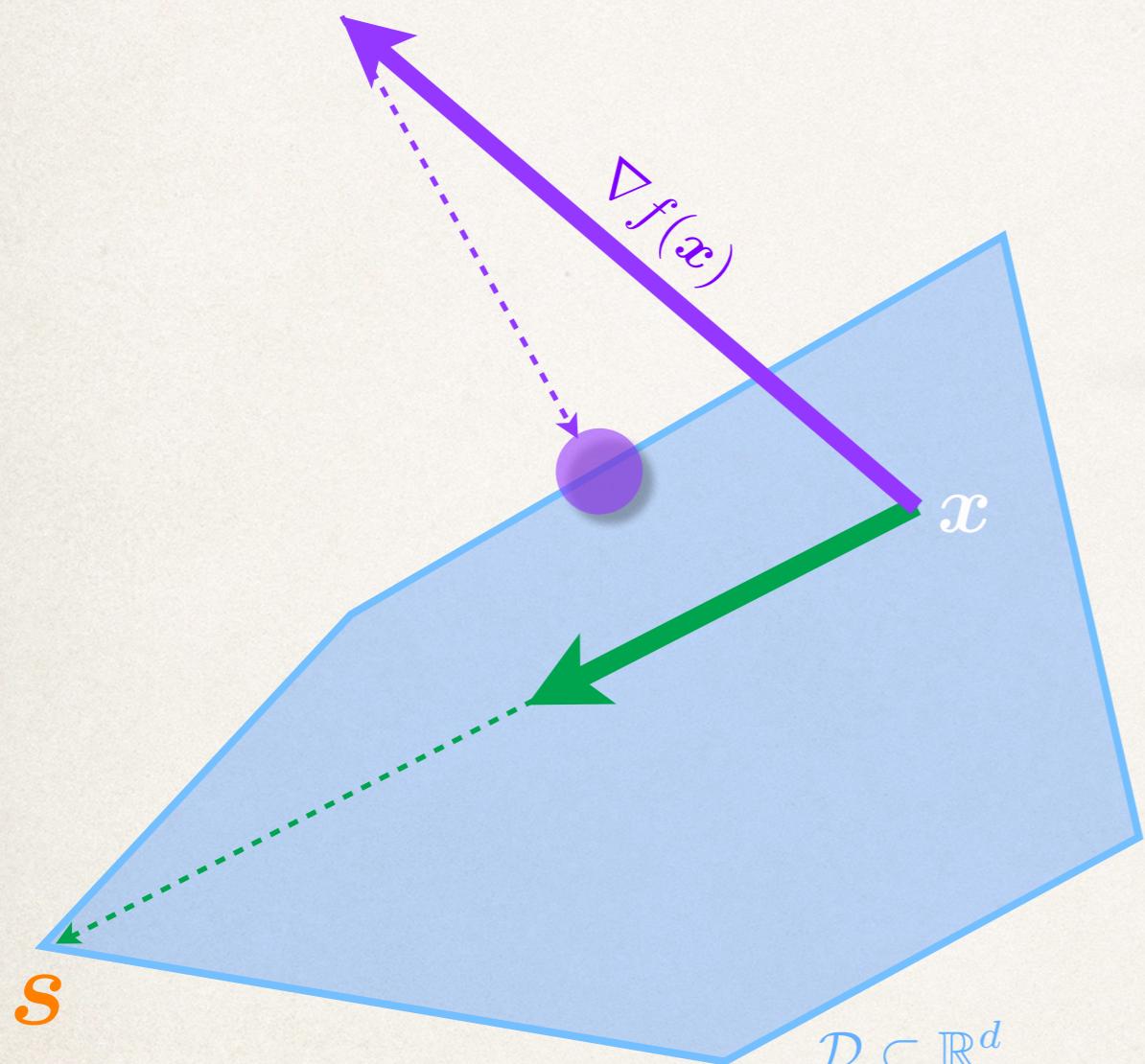
$$\begin{aligned}\omega(\mathbf{x}) := \\ \min_{\mathbf{s} \in \mathcal{D}} f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{s} - \mathbf{x} \rangle\end{aligned}$$

Weak Duality

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



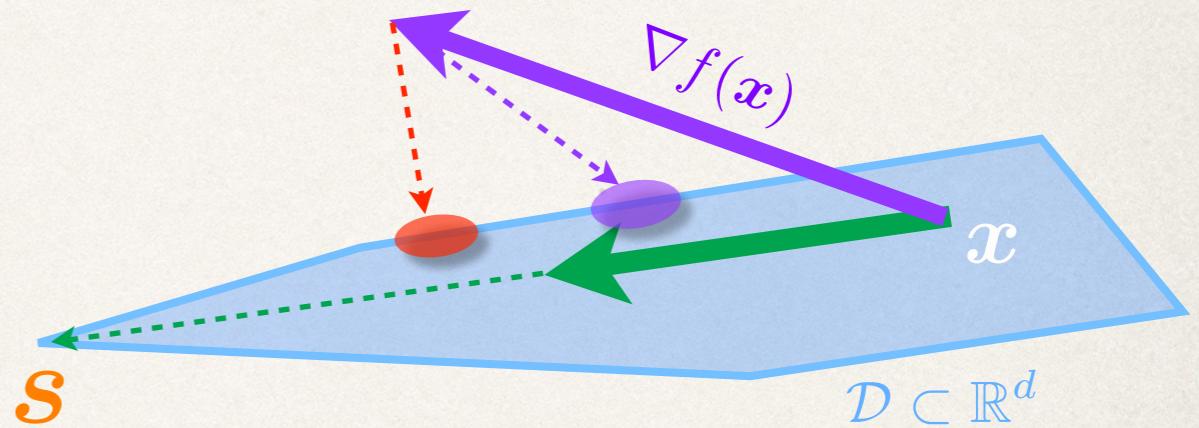
Affine Invariance (Geometry Independence)



Curvature constant

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

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



Convergence

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

Lower Bound for Sparsity, and Optimality of the Convergence Rate

$$\text{Error} \geq \frac{C_f}{t}$$

For an *optimization accuracy* of $\frac{C_f}{t}$,
any x must have *at least t non-zeros* (use *at least t corners*)

Trade-Off: Optimization Accuracy vs Sparsity
(algorithm-independent)

$$f(x) := \|x\|^2$$

$$\mathcal{D} := \Delta$$

Greedy Optimization over Atomic Sets

Atoms \mathcal{A}



$$\min_{\mathbf{x}} f(\mathbf{x})$$

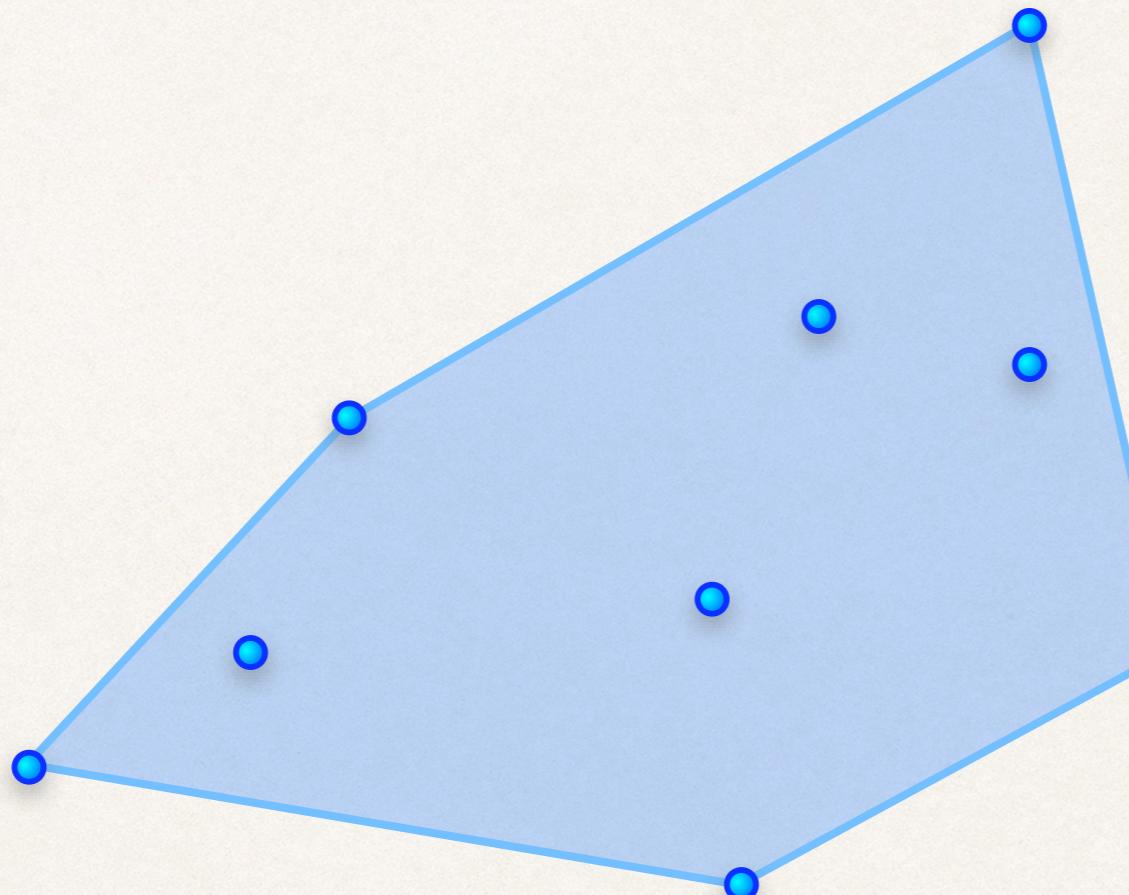
$$\mathbf{x}^{(t+1)} := \alpha \mathbf{x}^{(t)} + \beta \mathbf{s}$$

for $\mathbf{s} \in \mathcal{A}$

Greedy Optimization over Atomic Sets

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

Atoms \mathcal{A}



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

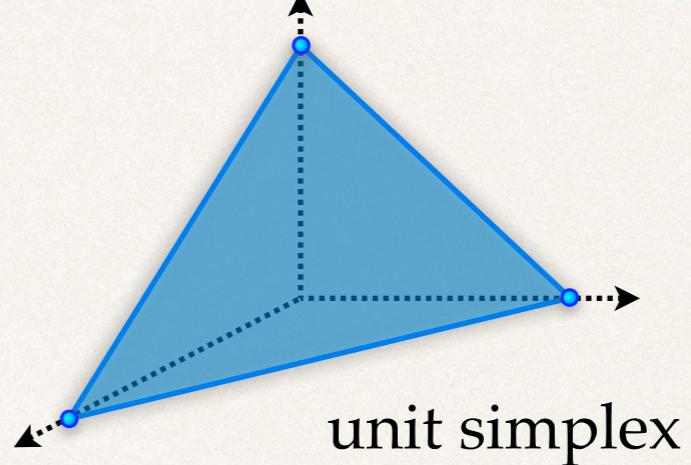
Frank-Wolfe:

$$\mathbf{x}^{(t+1)} := (1 - \gamma)\mathbf{x}^{(t)} + \gamma \mathbf{s}$$

for $\mathbf{s} \in \mathcal{A}$

Sparse Approximation

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



$$\min_{\mathbf{x} \in \Delta} f(\mathbf{x})$$

Corollary:
Obtain $\frac{2C_f}{t}$ -approximate
solution of **sparsity** t .

[Clarkson 2008]

Trade-Off:
Optimization accuracy vs **sparsity**

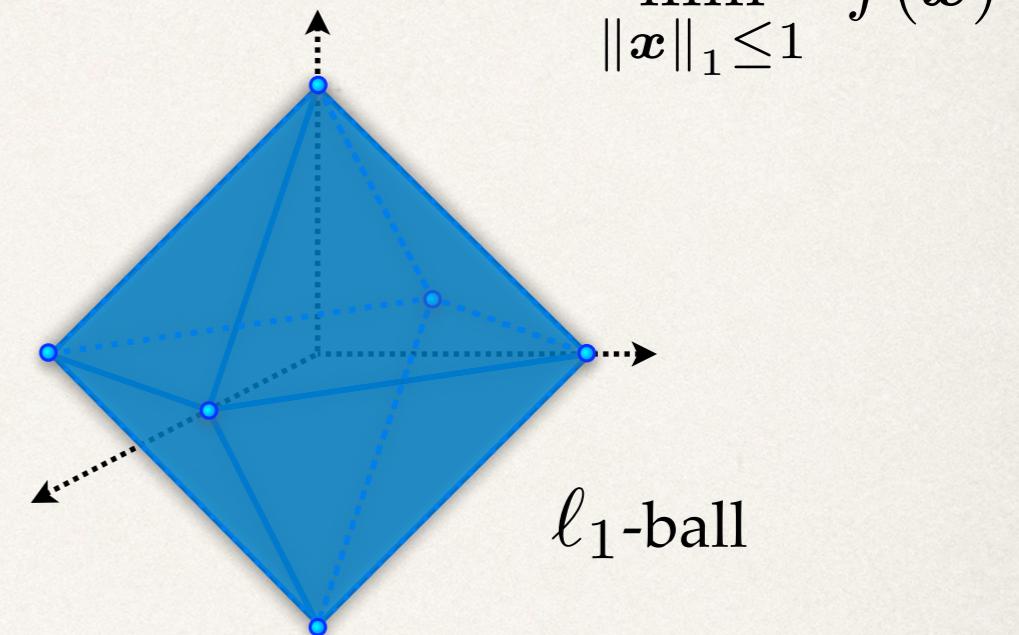
lower bound:
 $\frac{C_f}{t}$

Projection:
 $n \log n$

LMO:
linear time

Sparse Approximation

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



Corollary:

Obtain $\frac{2C_f}{t}$ -approximate solution of **sparsity** t .

Trade-Off:

Optimization accuracy vs **sparsity**

lower bound:

$$\frac{C_f}{t}$$

Projection:
 $n \log n$

LMO:
linear time

Demo & Code

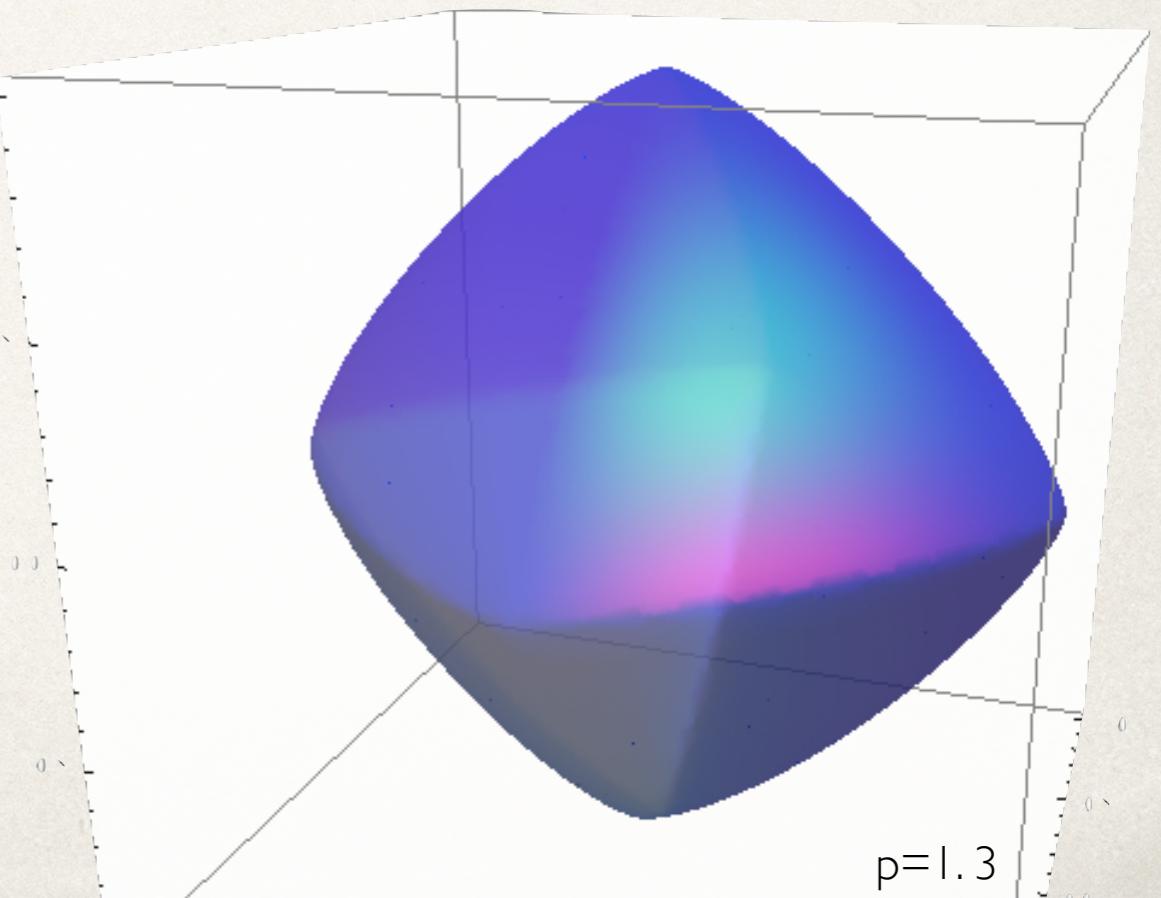
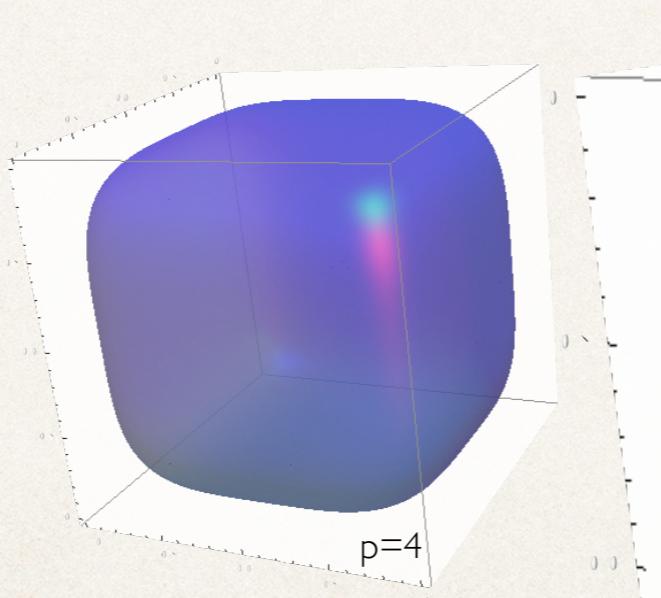
<https://sites.google.com/site/FrankWolfeGreedyTutorial>

ℓ_p -Norm Problems

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

$\mathcal{D} := \ell_p\text{-ball}$

$$1 < p < \infty$$



Projection:
unknown?

LMO:
linear time

Low Rank Approximation

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

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

nuclear-norm-ball

Corollary:

Obtain $\frac{2C_f}{t}$ -approximate solution of **rank** t .

[Jaggi & Sulovský 2010]

Trade-Off:

Optimization accuracy vs **rank**

lower bound:

$$\frac{C_f}{t}$$

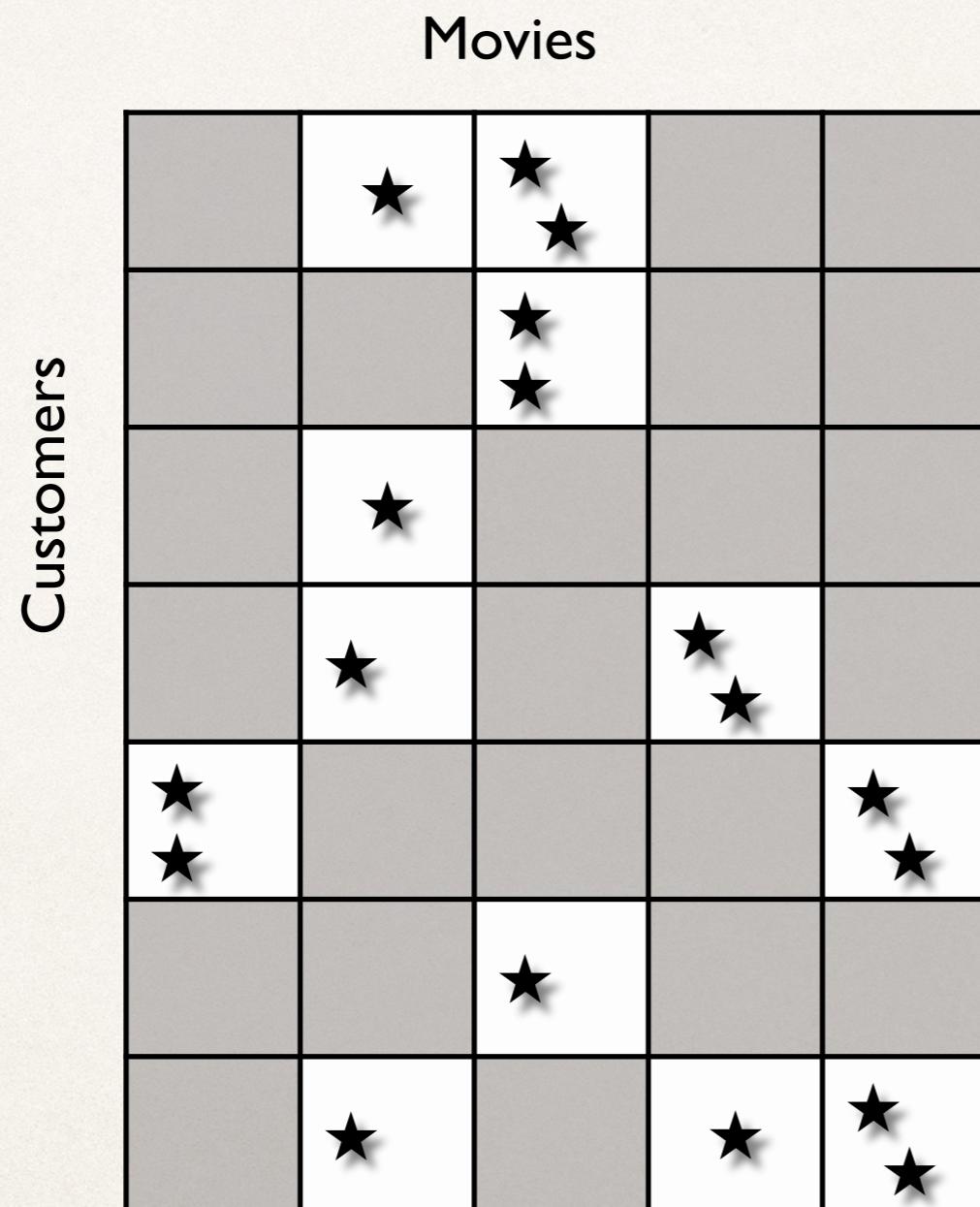
Projection:

full SVD

LMO:

approx. top singular vector

Low Rank Approximation



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

$$\approx UV^T$$

Image Classification with nuclear norm regularization



Star Anise (92.54 %)



Pulp Magazine (83.01 %)



Sea Snake (10.00 %)



Geyser (85.45 %)



Carricot (81.48 %)



Paintbrush (4.68 %)

ImageNet top-1 accuracy for each class

[Harchaoui et al. 2012]

Examples of Atomic Domains Suitable for Frank-Wolfe

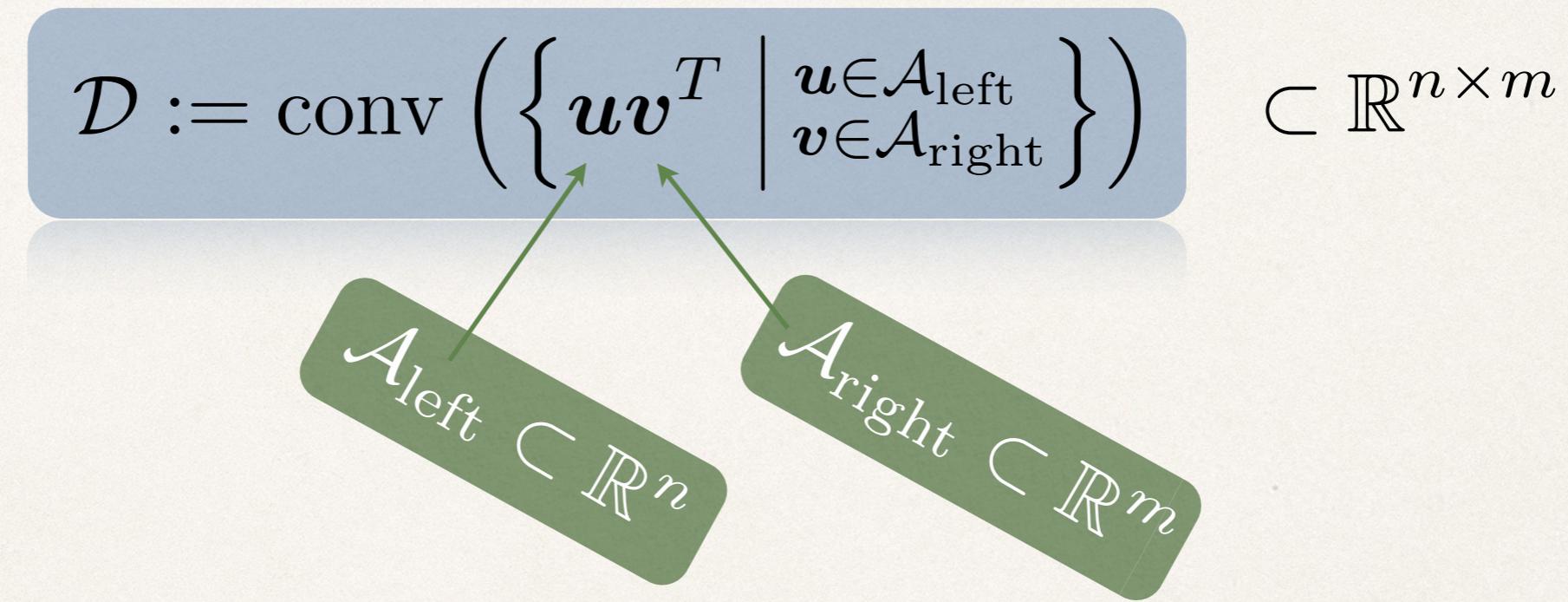
\mathcal{X}	Optimization Domain Atoms \mathcal{A}	$\mathcal{D} = \text{conv}(\mathcal{A})$	Complexity of one Frank-Wolfe Iteration $\sup_{\mathbf{s} \in \mathcal{D}} \langle \mathbf{s}, \mathbf{y} \rangle$	Complexity
\mathbb{R}^n	Sparse Vectors	$\ \cdot\ _1$ -ball	$\ \mathbf{y}\ _\infty$	$O(n)$
\mathbb{R}^n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

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 $p = 1$ and vice versa.

Factorized Matrix Domains (and *tensor*)

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

(nuclear norm)



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	$\ \cdot\ _2$ -sphere	$\ \cdot\ _2$ -sphere	Trace norm $\ M\ _{tr}$	$\ M\ _{op}$	Lanczos, see Table 1
1	$\ \cdot\ _1$ -sphere	$\ \cdot\ _1$ -sphere	Vector ℓ_1 -norm $\ \vec{M}\ _1$	$\ \vec{M}\ _\infty$	$O(nm)$
1	$\ \cdot\ _\infty$ -sphere	$\ \cdot\ _\infty$ -sphere		Cut-norm $\ \cdot\ _{\infty \rightarrow 1}$	NP-hard (Alon & Naor, 2006)
$n+m$	$\ \cdot\ _{2,\infty}$	$\ \cdot\ _{2,\infty}$	Max-norm $\ M\ _{\max}$		SDP, see Table 1
1	$\ \cdot\ _2 \cap \mathbb{R}_{>0}^m$	$\ \cdot\ _2 \cap \mathbb{R}_{>0}^n$	“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

Examples of *Atomic Domains* Suitable for Frank-Wolfe

- ✿ shortest paths / network flows / transportation
- ✿ greedy selection and sparse optimization
- ✿ low-rank matrix factorizations, many other matrices
- ✿ wavelets (infinite-dimensional)
- ✿ structured sparsity and structured prediction
- ✿ total-variation-norm for image de-noising
- ✿ submodular optimization
- ✿ boosting
- ✿ training deep networks

Faster Convergence

under additional assumptions

An affine invariant notion of *smoothness* and *strong convexity*

Curvature constant

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

$$C_f \leq L \cdot \text{diam}(\mathcal{D})^2$$

Strong convexity

$$\mu_f^{\text{int}} := \inf_{\substack{\mathbf{x} \in \mathcal{D} \setminus \{\mathbf{x}^*\}, \\ \gamma \in (0,1], \\ \bar{\mathbf{s}} := \text{ray}(\mathbf{x}, \mathbf{x}^*) \cap \partial \mathcal{D}, \\ \mathbf{y} = \mathbf{x} + \gamma(\bar{\mathbf{s}} - \mathbf{x})}} \frac{2}{\gamma^2} (f(\mathbf{y}) - f(\mathbf{x}) - \langle \mathbf{y} - \mathbf{x}, \nabla f(\mathbf{x}) \rangle)$$

$$\mu_f^{\text{int}} \geq \mu \cdot \min dist(x^*, \partial \mathcal{D})^2$$

Linear Convergence for strongly convex f , under additional assumptions

[Guélat & Marcotte, 1986, Garber & Hazan 2013]

Thm: The iterates of **standard FW** (with line-search) satisfy

$$f(\mathbf{x}^{(t)}) - f^* \leq h_0 \exp\left(-\frac{1}{2} \frac{\mu_f^{\text{int}}}{C_f} t\right)$$

[Lacoste-Julien & Jaggi 2013]

$$\mu_f^{\text{int}} \geq \mu \cdot \min dist(x^*, \partial \mathcal{D})^2$$

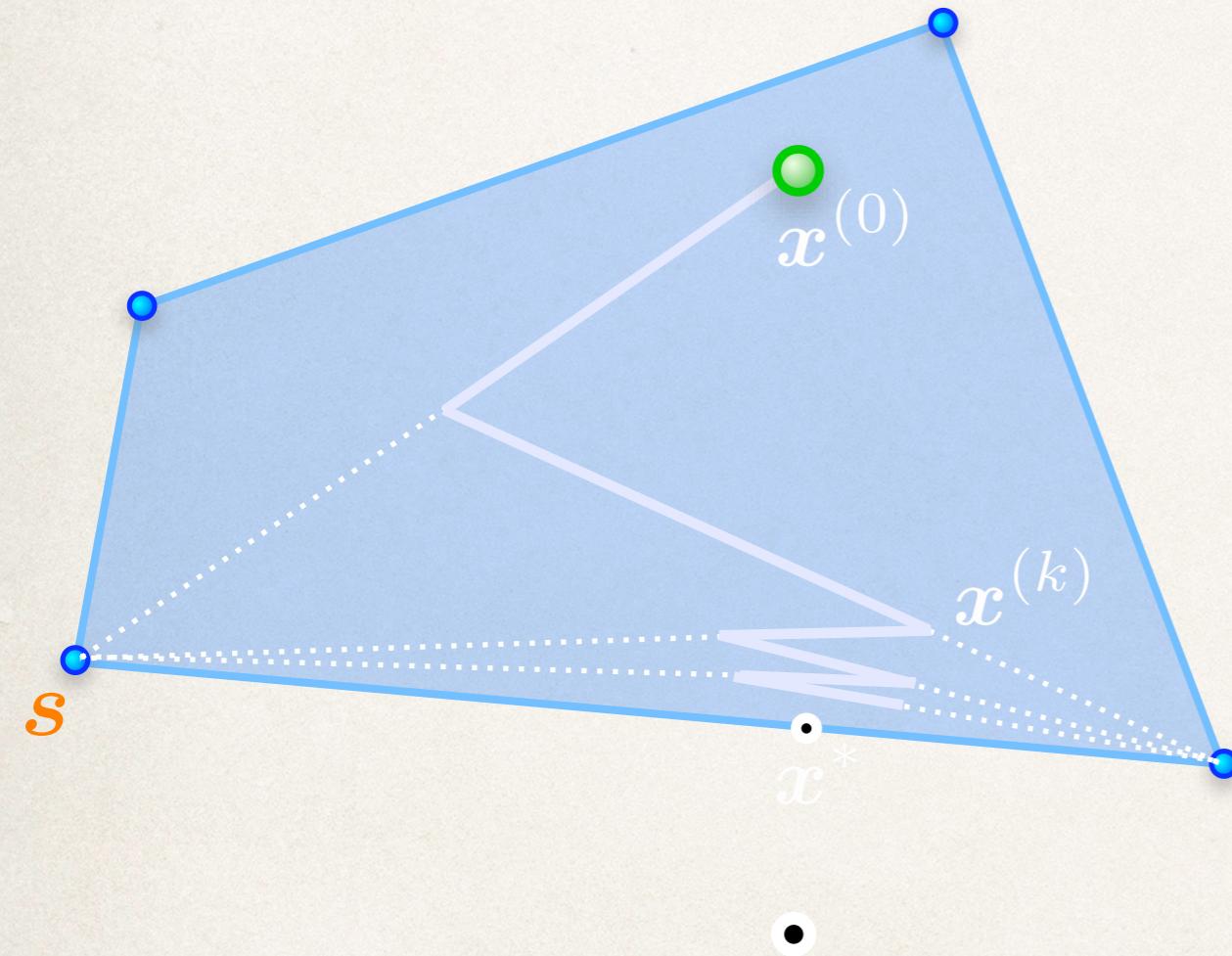
$$C_f \leq L \cdot \text{diam}(\mathcal{D})^2$$

complexity measure

↑
of function, of domain (and opt x)

Frank-Wolfe with Away-Steps

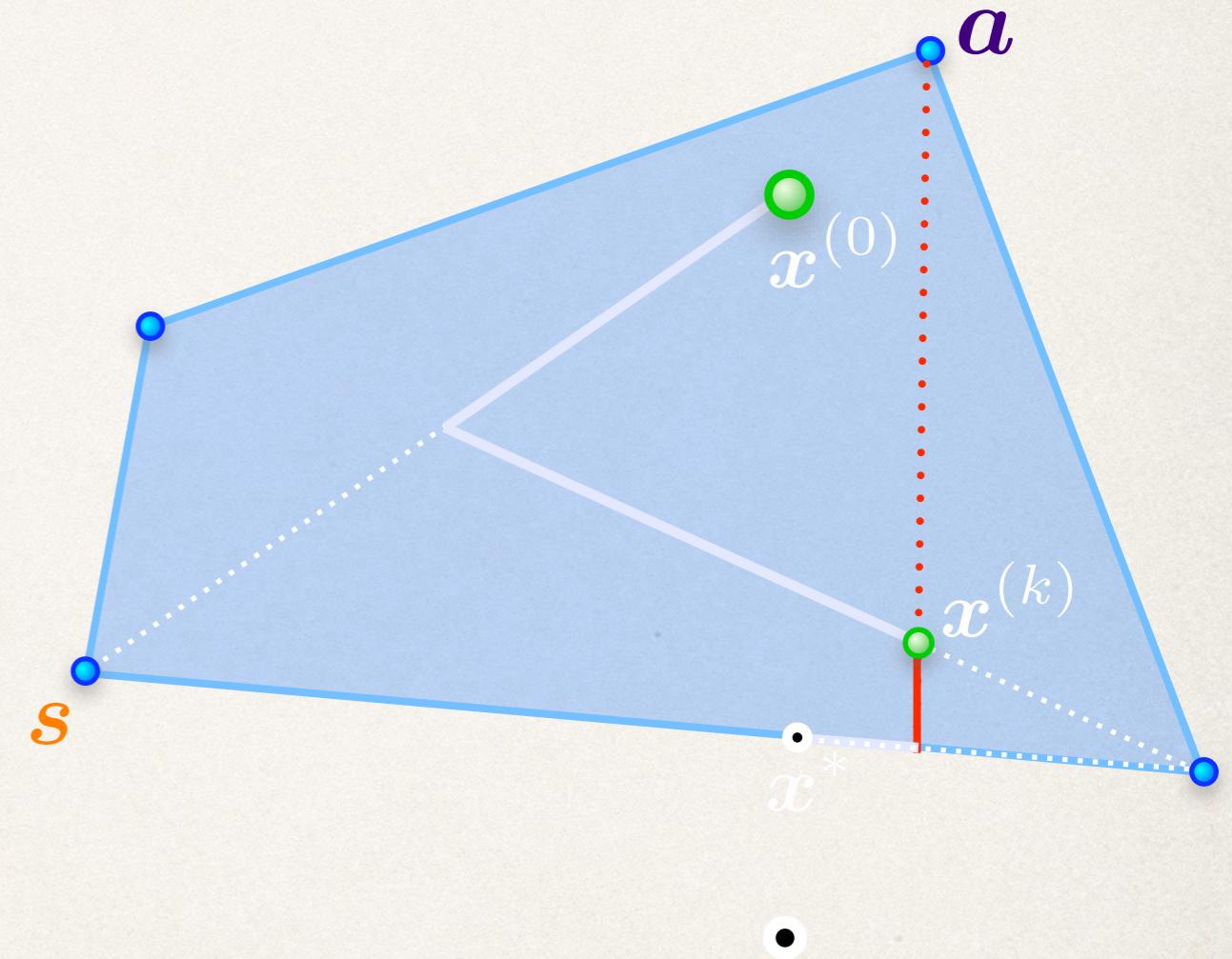
Standard FW



$$\mathbf{s} := \text{LMO}_{\mathcal{D}}(\mathbf{d}) = \arg \min_{\mathbf{s} \in \mathcal{D}} \langle \mathbf{d}, \mathbf{s} \rangle$$

FW with away steps

[Wolfe 1970, Guélat et al. 1986]



$$\mathbf{a} := \arg \max_{\mathbf{a} \in \{\mathbf{s}^{(0)}, \dots, \mathbf{s}^{(t)}\}} \langle \mathbf{d}, \mathbf{a} \rangle$$

Linear Convergence for strongly convex f , under additional assumptions

[Guélat & Marcotte, 1986, Garber & Hazan 2013]

Thm: The iterates of **standard FW** (with line-search) satisfy

$$f(\mathbf{x}^{(t)}) - f^* \leq h_0 \exp\left(-\frac{1}{2} \frac{\mu_f^{\text{int}}}{C_f} t\right)$$

[Lacoste-Julien & Jaggi 2013]

$$\mu_f^{\text{int}} \geq \mu \cdot \min dist(x^*, \partial \mathcal{D})^2$$

$$C_f \leq L \cdot \text{diam}(\mathcal{D})^2$$

complexity measure

↑
of function, of domain (and opt x)

Linear Convergence for strongly convex f , under additional assumptions

Thm: The iterates of FW with away-steps satisfy

$$f(\mathbf{x}^{(\textcolor{red}{t})}) - f^* \leq h_0 \exp\left(-\frac{1}{2} \frac{\mu_f^{\text{away}}}{C_f} \textcolor{red}{t}\right)$$

[Lacoste-Julien & Jaggi 2013]

$$\frac{\mu_f^{\text{away}} \geq \mu \cdot \text{pyr width}(\mathcal{D})^2}{C_f \leq L \cdot \text{diam}(\mathcal{D})^2}$$

complexity measure of function, of domain ~~(and opt x)~~

$1/t^2$ -convergence for strongly convex D

[Garber & Hazan 2014, Levitin and Polyak 1966]

Tutorial Website

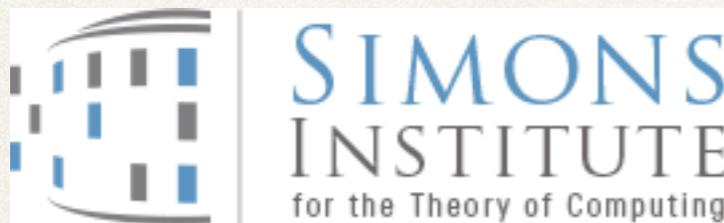
- ✿ Sample Code
- ✿ Bibliography
- ✿ Tutorial Slides
- ✿ Topics not covered

<https://sites.google.com/site/FrankWolfeGreedyTutorial>

Thanks



Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich



Microsoft Research - Inria
JOINT CENTRE



Tutorial Outline

- ✿ **Part I:** (*Martin Jaggi*)

- ✿ Frank-Wolfe basics and history
- ✿ convergence analysis and geometry independence
- ✿ optimality & lower bounds
- ✿ demo and applications for atomic domains
- ✿ faster rates under additional assumptions

- ✿ **Part II:** (*Zaid Harchaoui*)

- ✿ Greedy Optimization and Frank-Wolfe
- ✿ Frank-Wolfe methods for:
 - ✿ composite optimization
 - ✿ non-smooth optimization
 - ✿ block-structured optimization