Block-Coordinate Frank-Wolfe Optimization

with applications to structured prediction

Martin Jaggi
CMAP, Ecole Polytechnique, Paris

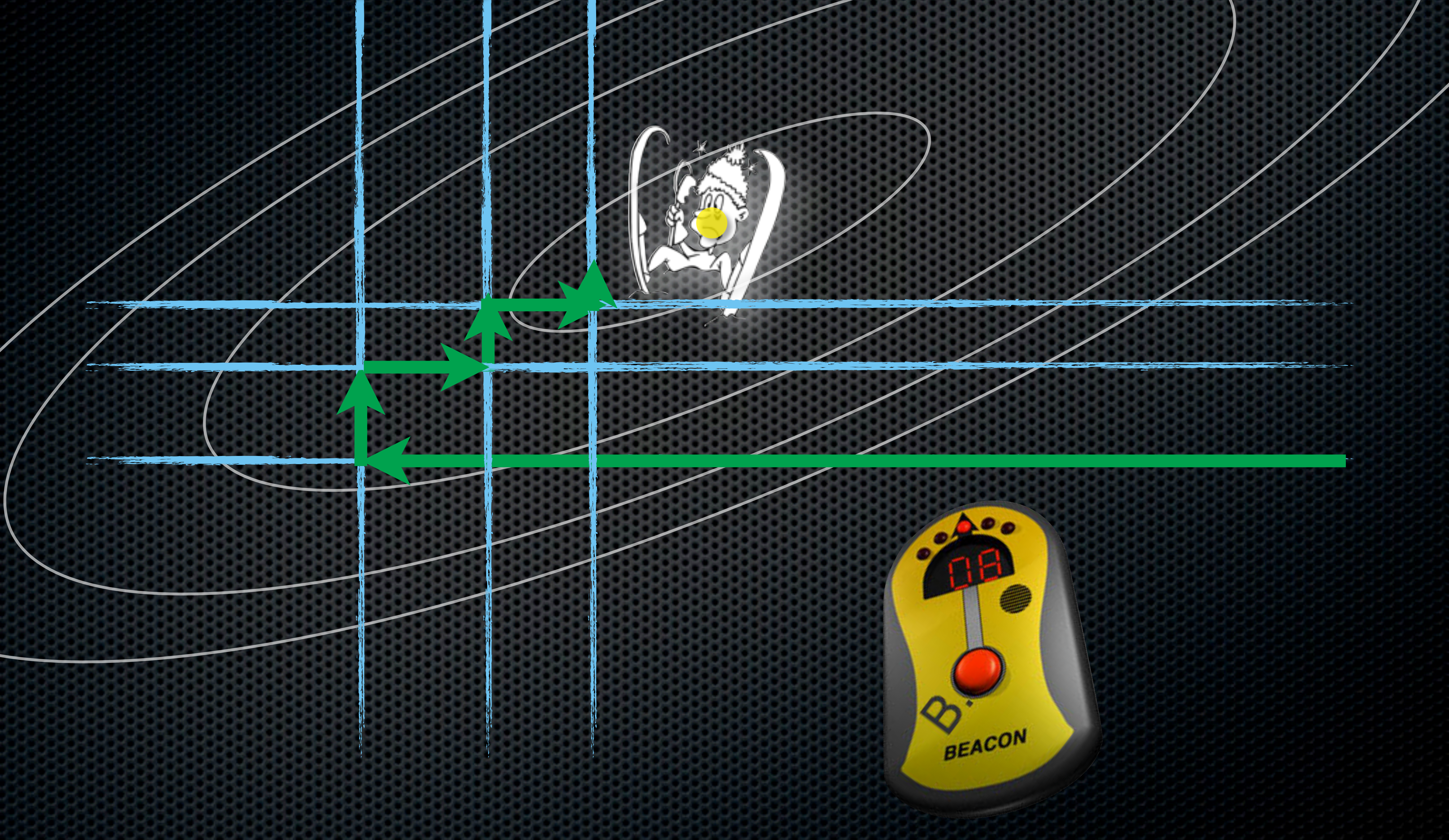
Optimization and Big Data Workshop, Edinburgh, 2013 / 5 / 2

Co-Authors: Simon Lacoste-Julien, Mark Schmidt and Patrick Pletscher

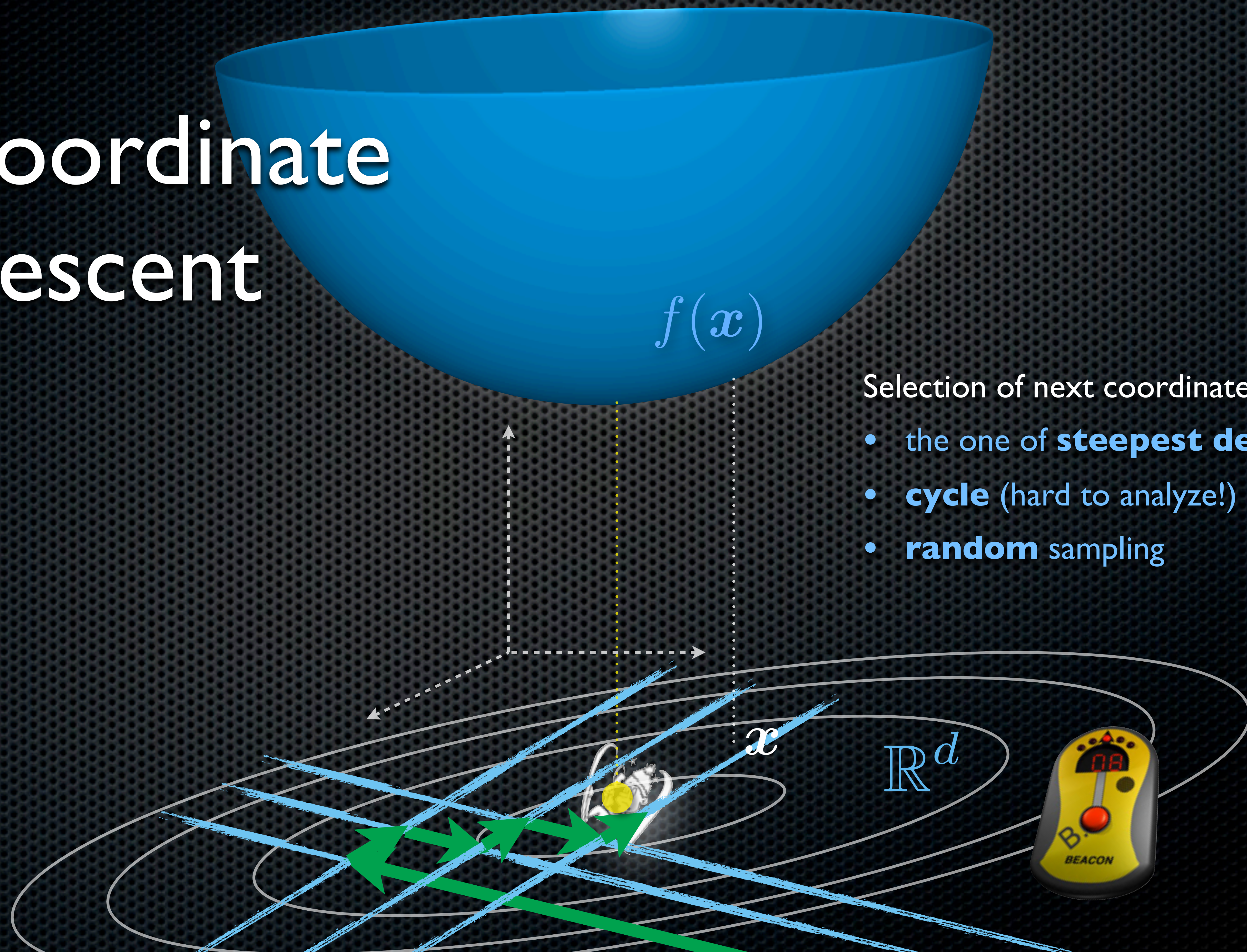
Outline

- Two Old First-Order Optimization Algorithms
  - Coordinate Descent
  - The Frank-Wolfe Algorithm
- Duality for Constrained Convex Optimization
- Combining *Frank-Wolfe* and *Coordinate Descent*
- Applications: Large Margin Prediction
  - binary SVMs
  - structural SVMs

Coordinate Descent



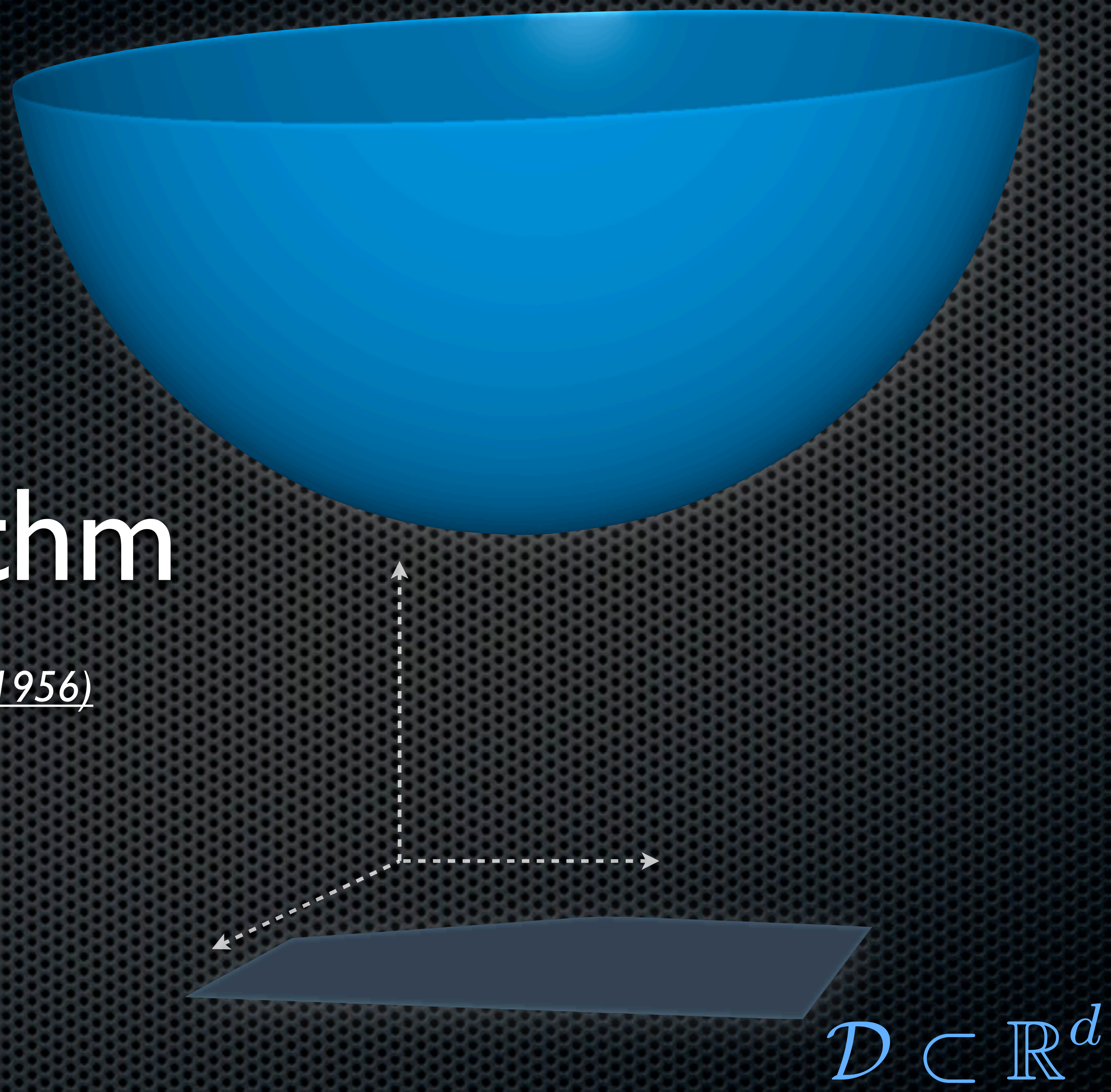
Coordinate Descent



Selection of next coordinate:

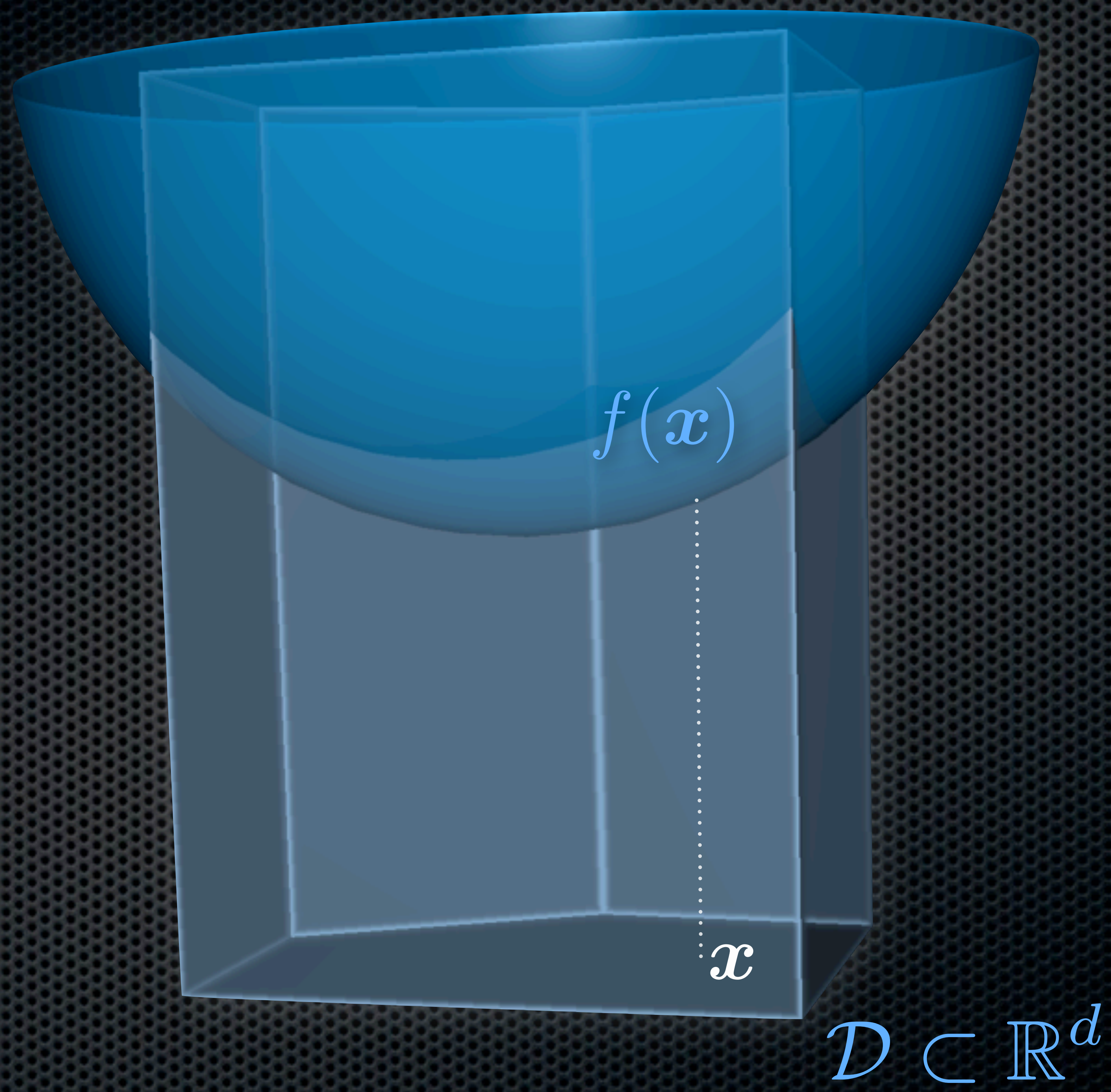
- the one of **steepest desc.**
- **cycle** (hard to analyze!)
- **random** sampling

The Frank-Wolfe Algorithm

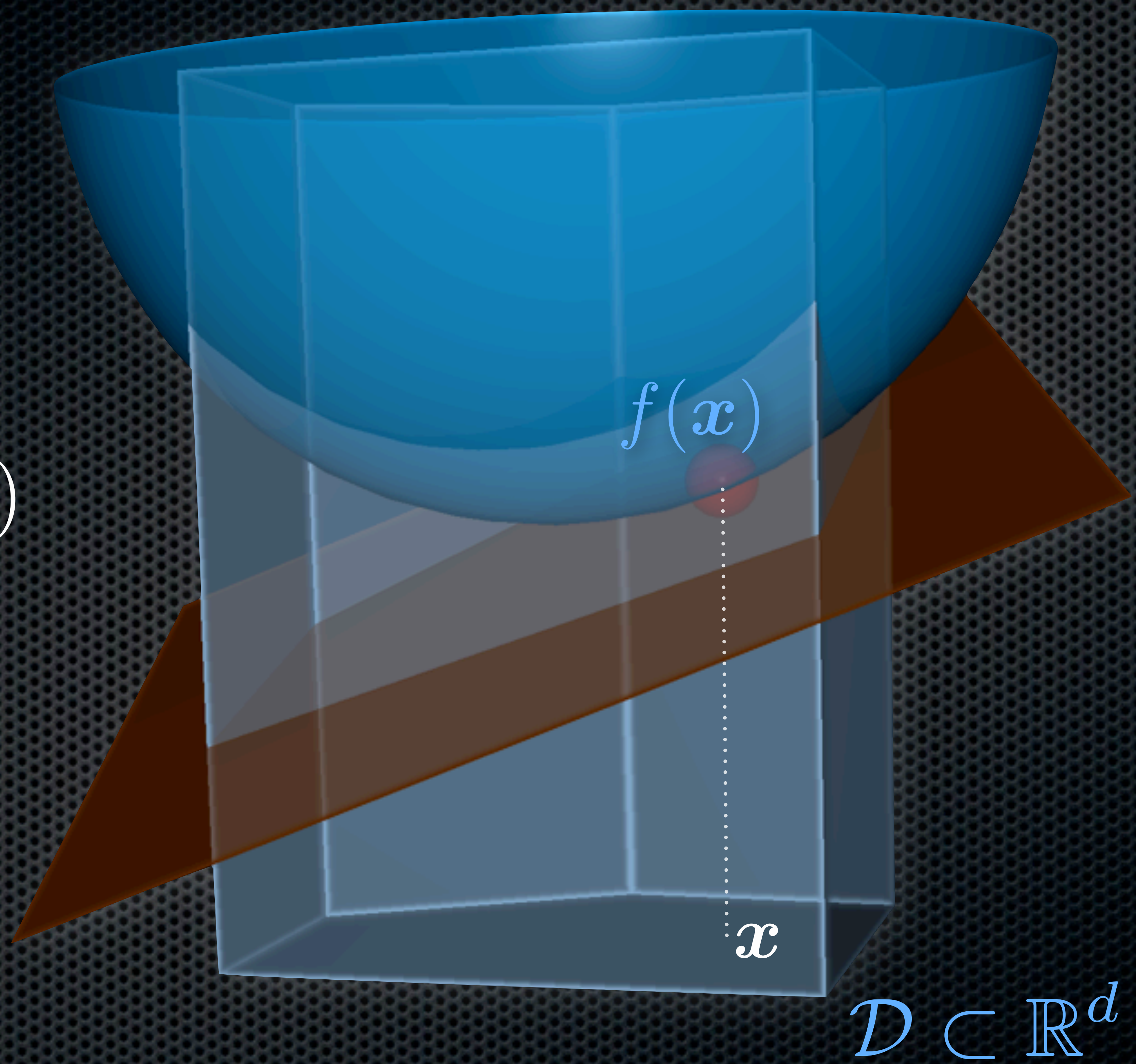


Frank and Wolfe (1956)

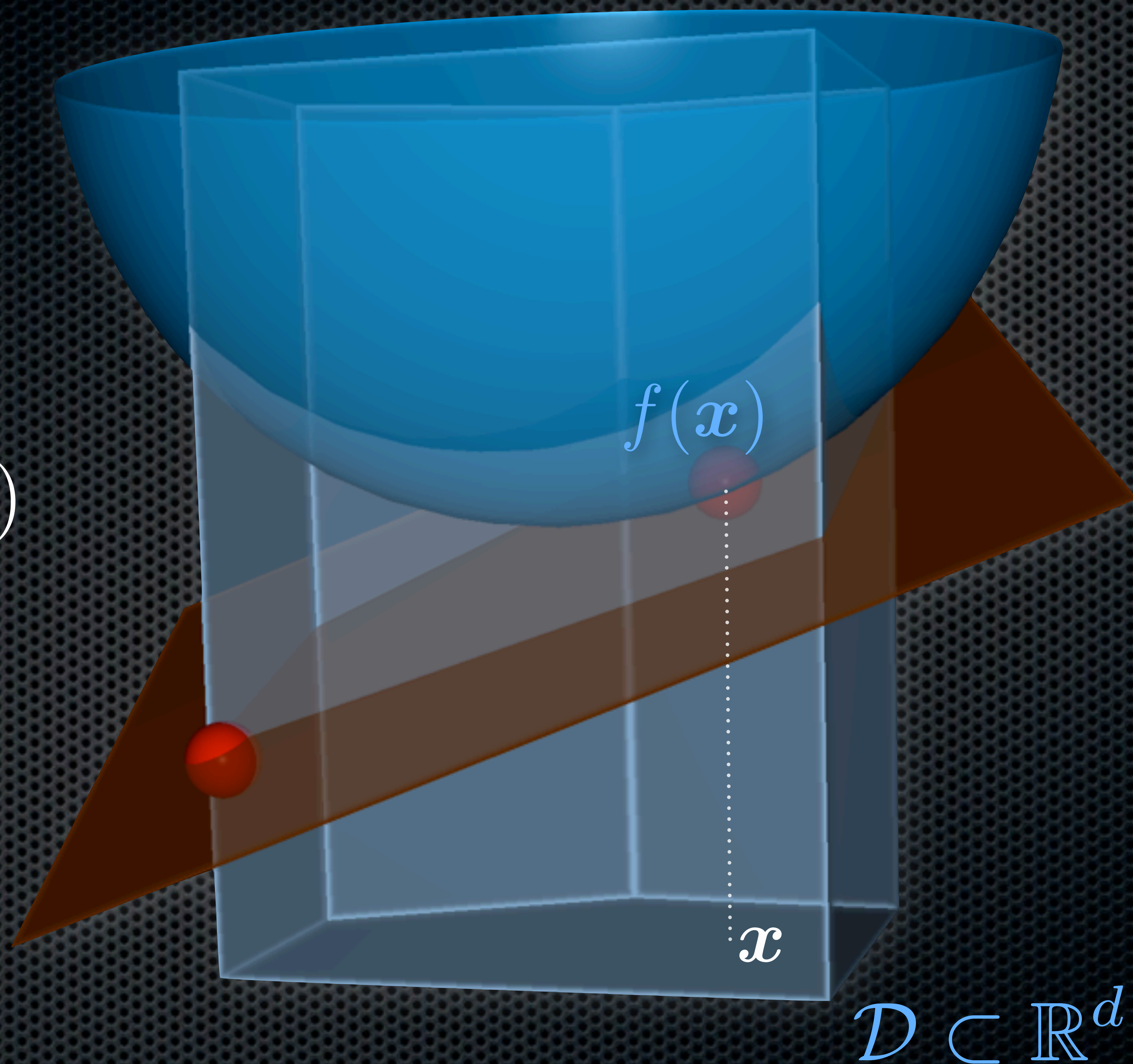
$$\min_{\boldsymbol{x} \in \mathcal{D}} f(\boldsymbol{x})$$

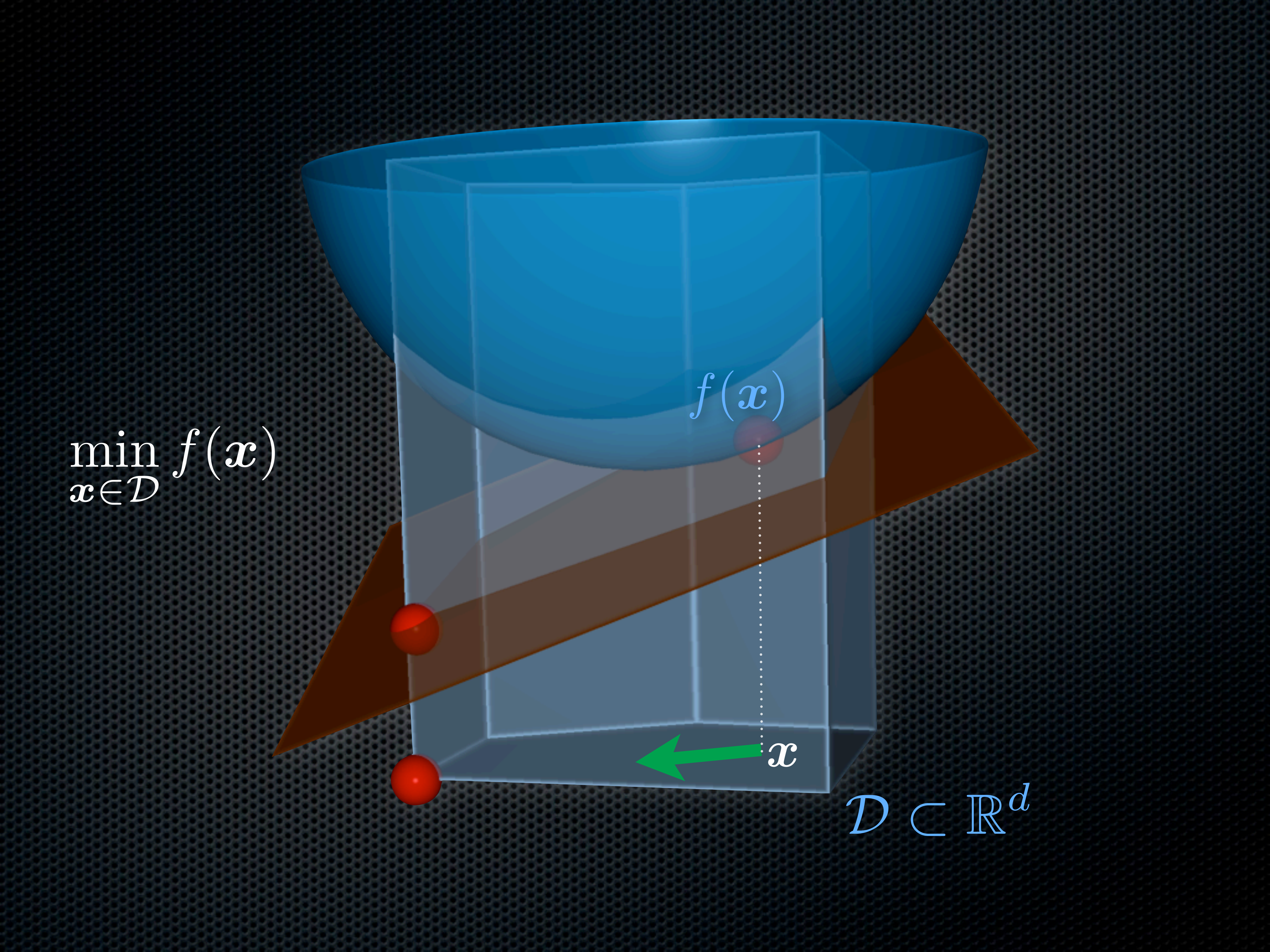


$$\min_{\boldsymbol{x} \in \mathcal{D}} f(\boldsymbol{x})$$



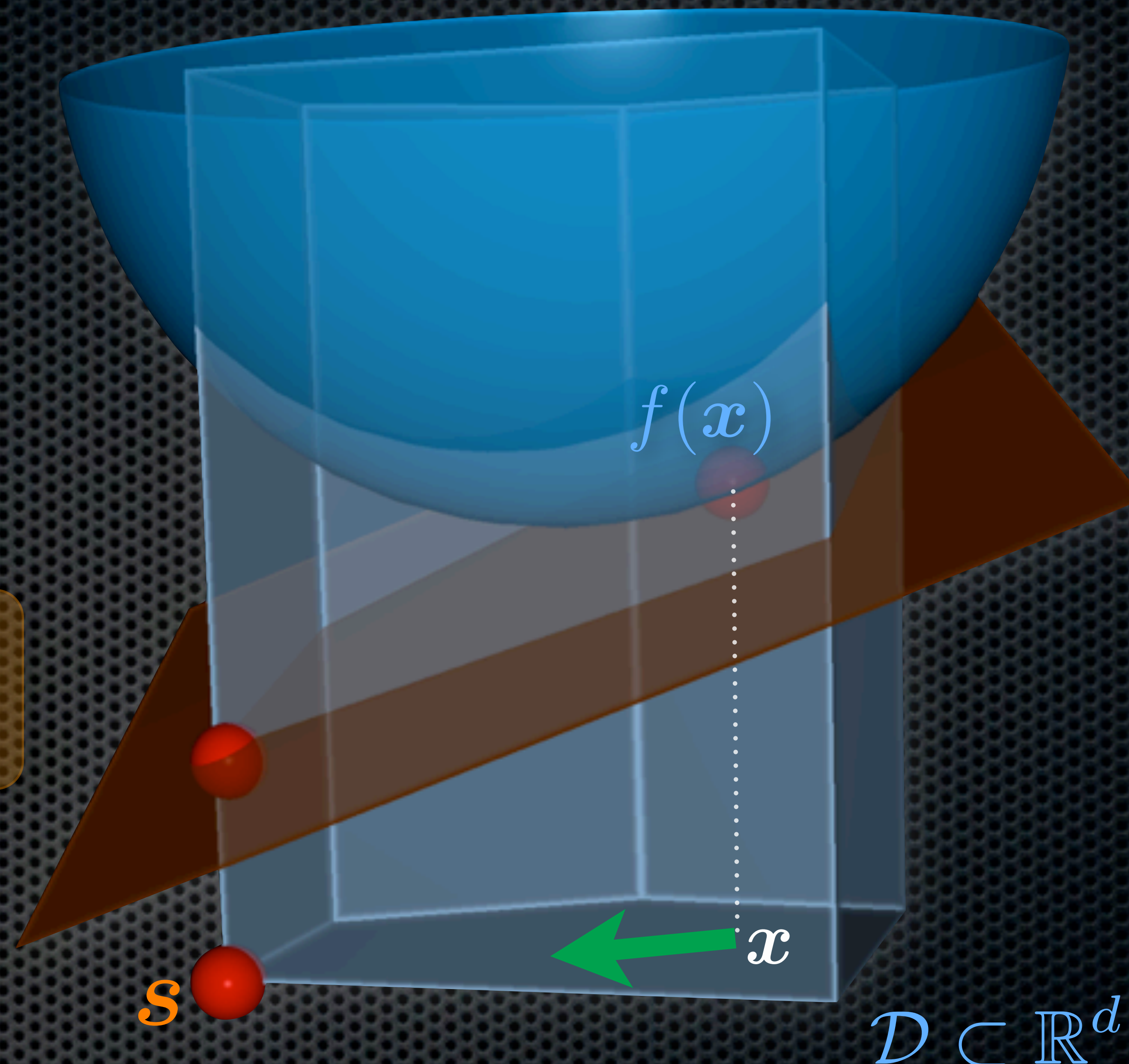
$$\min_{\boldsymbol{x} \in \mathcal{D}} f(\boldsymbol{x})$$




$\min_{x \in \mathcal{D}} f(x)$
$f(x)$
x
$\mathcal{D} \subset \mathbb{R}^d$

The Linearized Problem

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



Algorithm 1 Frank-Wolfe

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

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

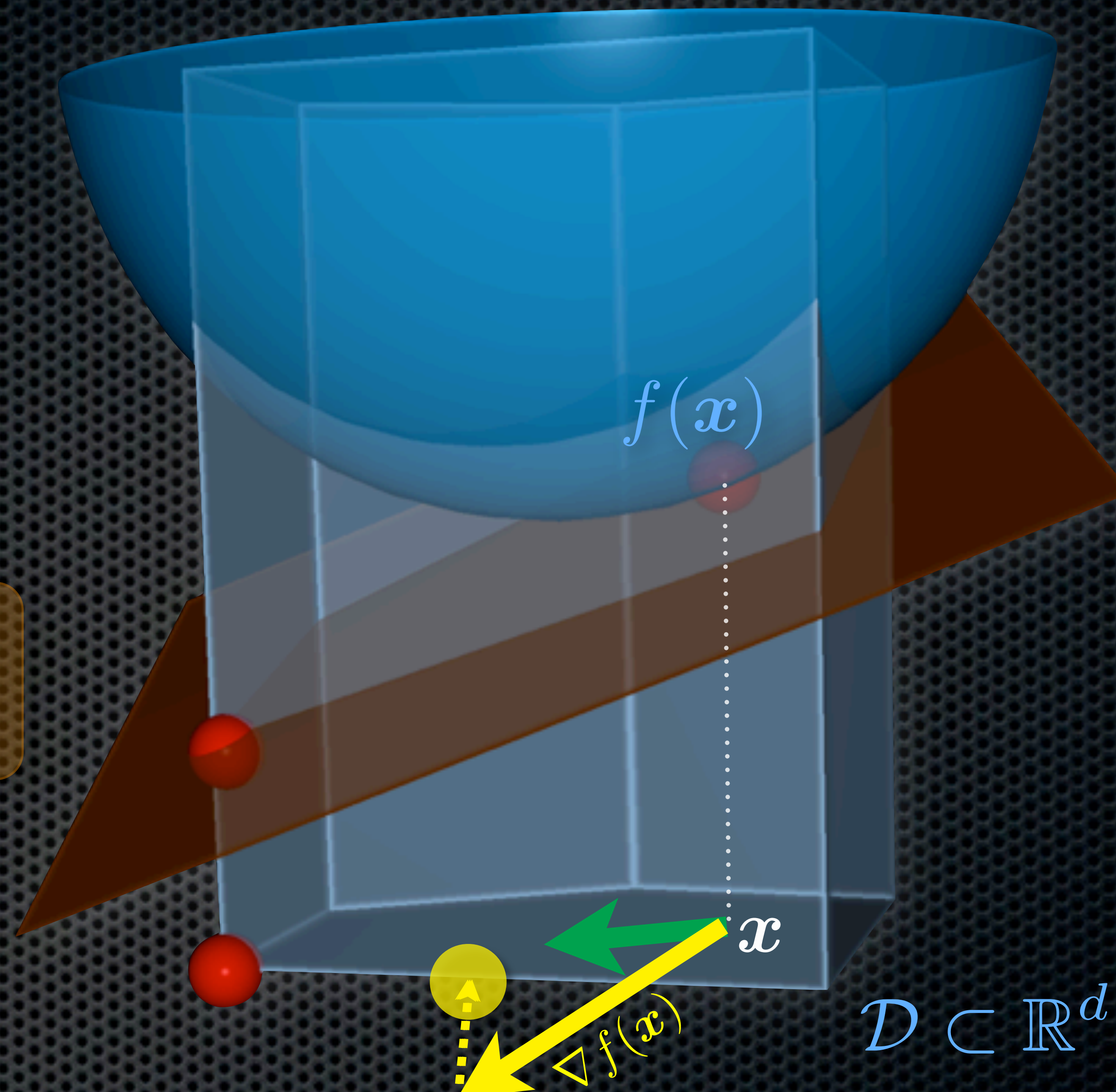
Let $\gamma := \frac{2}{k+2}$

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

end for

The Linearized Problem

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



	Frank-Wolfe	**Gradient Descent**
Cost per step	*(approx.)* solve linearized problem on D	Projection back to D
Sparse Solutions	✓ (in terms of used vertices)	✗

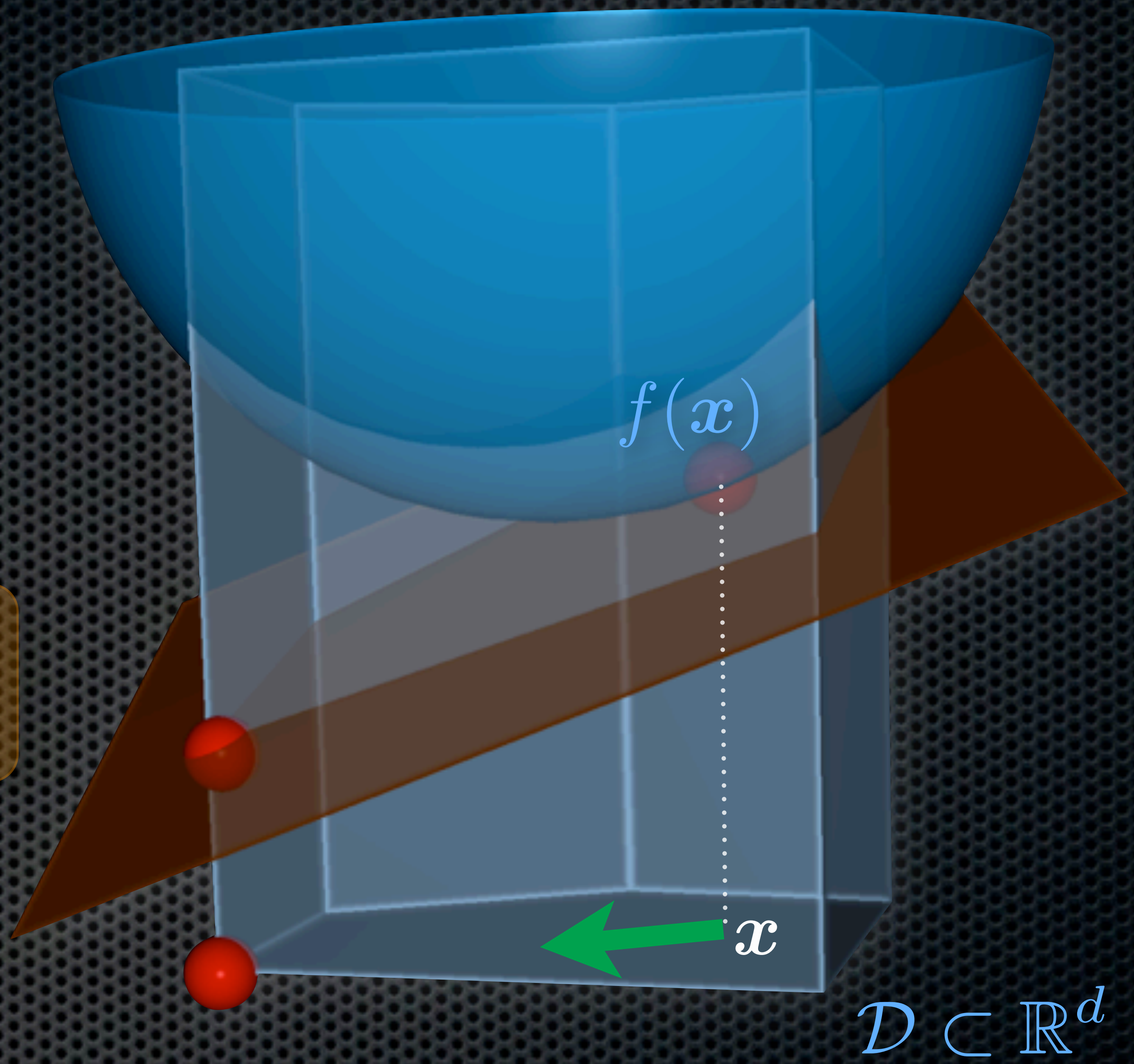
Some Examples of Atomic Domains Suitable for Frank-Wolfe

$\mathcal{X}$	**Optimization Domain**		**Complexity of one Frank-Wolfe Iteration**	
	Atoms $\mathcal{A}$	$\mathcal{D} = \operatorname{conv}(\mathcal{A})$	$\sup_{\boldsymbol{s}\in\mathcal{D}} \langle \boldsymbol{s}, \boldsymbol{y} \rangle$	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	$\max_{g\in\mathcal{G}} \left\\|\boldsymbol{y}_{(g)}\right\\|_g^*$	$\sum_{g\in\mathcal{G}} \lvert g \rvert$
$\mathbb{R}^{m\times n}$	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	$\{\boldsymbol{x}\succeq 0,\ \operatorname{Tr}(\boldsymbol{x})=1\}$	$\lambda_{\max}(\boldsymbol{y})$	$\tilde{O}\big(N_f/\sqrt{\varepsilon'}\big)$ (Lanczos)
$\mathbb{S}^{n\times n}$	PSD matrices of bounded diagonal	$\{\boldsymbol{x}\succeq 0,\ \boldsymbol{x}_{ii}\le 1\}$		$\tilde{O}\big(N_f\, n^{1.5}/\varepsilon'^{2.5}\big)$

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* ...

J. 2013

Dudık et al. 2011, Tewari et al. 2011, J. 2011



The Linearized Problem

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

Primal Convergence:

Algorithms obtain

$$f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^*) \leq O\left(\tfrac{1}{k}\right)$$

after k **steps**.

[Frank & Wolfe 1956]

Primal-Dual Convergence:

Algorithms obtain

$$gap(\boldsymbol{x}^{(k)}) \leq O\left(\tfrac{1}{k}\right)$$

after k **steps**.

[Clarkson 2008, J. 2013]

A Simple Optimization Duality

Original Problem

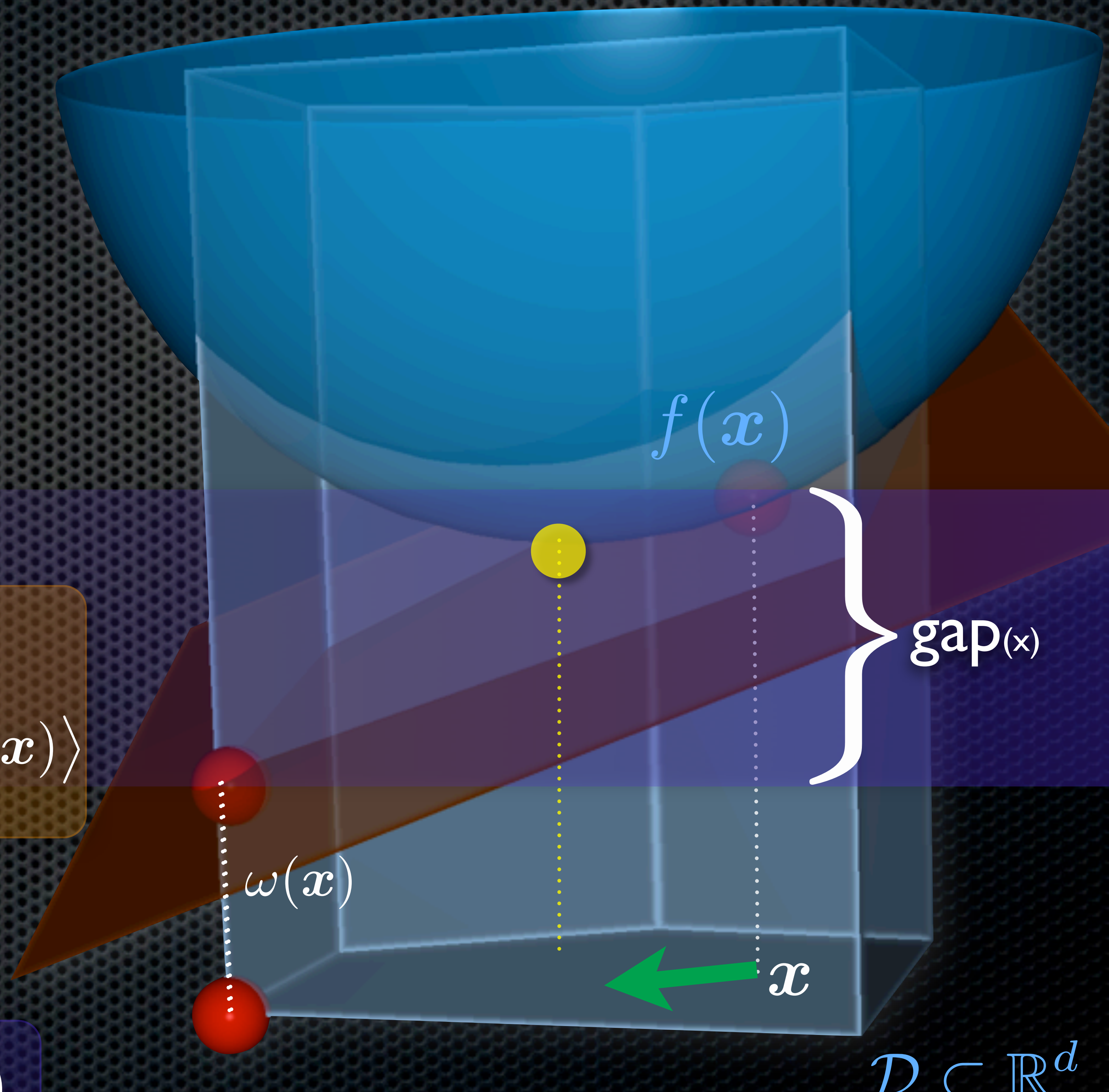
$$\min_{\boldsymbol{x} \in \mathcal{D}} f(\boldsymbol{x})$$

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

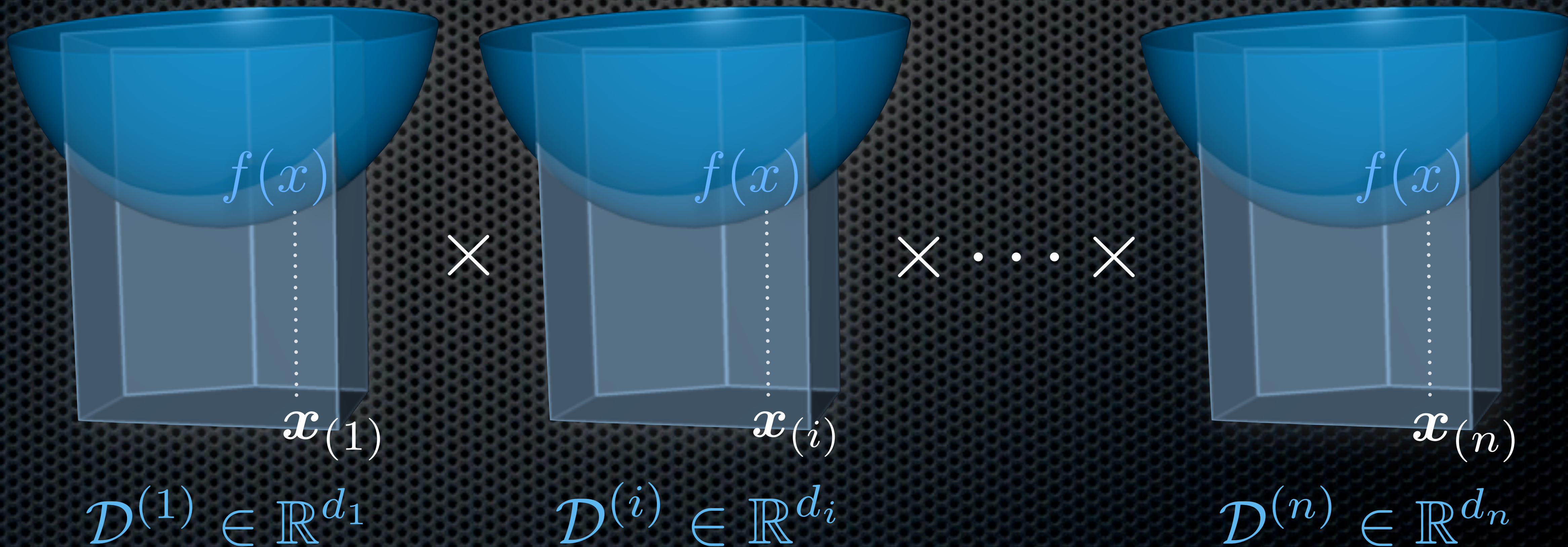
$$\omega(\boldsymbol{x}) \leq f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}')$$

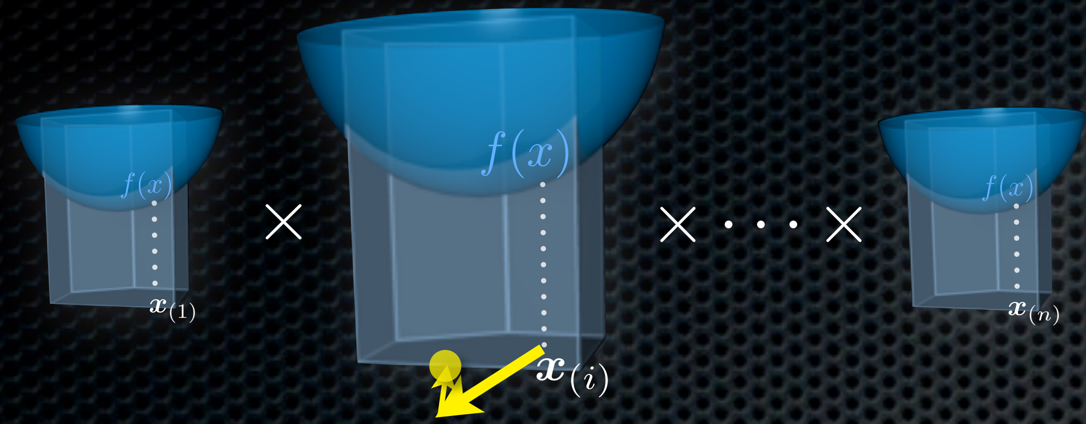


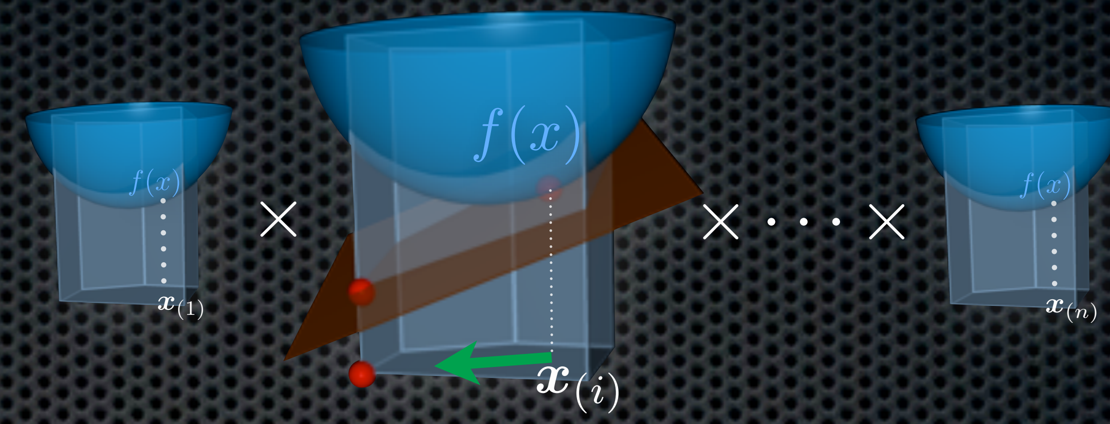
Block-Separable Optimization Problems

$$\min_{\boldsymbol{x} \in \mathcal{D}^{(1)} \times \cdots \times \mathcal{D}^{(n)}} f(\boldsymbol{x})$$

$$\boldsymbol{x} = (\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(n)})$$







Algorithm 2: Uniform Coordinate Descent

Let $\boldsymbol{x}^{(0)} \in \mathcal{D}$
for $k = 0 \dots K$ **do**
  Pick $i \in_{u.a.r.} [n]$
  Compute $\boldsymbol{s}_{(i)} := \underset{\boldsymbol{s}_{(i)} \in \mathcal{D}^{(i)}}{\arg\min} \left\langle \boldsymbol{s}_{(i)}, \nabla_{(i)} f(\boldsymbol{x}^{(k)}) \right\rangle + \frac{L_i}{2} \left\| \boldsymbol{s}_{(i)} - \boldsymbol{x}_{(i)} \right\|^2$
  Update $\boldsymbol{x}_{(i)}^{(k+1)} := \boldsymbol{x}_{(i)}^{(k)} + \left(\boldsymbol{s}_{(i)} - \boldsymbol{x}_{(i)}^{(k)}\right)$
end

Algorithm 3: Block-Coordinate "Frank-Wolfe"

Let $\boldsymbol{x}^{(0)} \in \mathcal{D}$
for $k = 0 \dots K$ **do**
  Pick $i \in_{u.a.r.} [n]$
  Compute $\boldsymbol{s}_{(i)} := \underset{\boldsymbol{s}_{(i)} \in \mathcal{D}^{(i)}}{\arg\min} \left\langle \boldsymbol{s}_{(i)}, \nabla_{(i)} f(\boldsymbol{x}^{(k)}) \right\rangle$
  Let $\gamma := \frac{2n}{k+2n}$, or optimize γ by line-search
  Update $\boldsymbol{x}_{(i)}^{(k+1)} := \boldsymbol{x}_{(i)}^{(k)} + \gamma\left(\boldsymbol{s}_{(i)} - \boldsymbol{x}_{(i)}^{(k)}\right)$
end

Nesterov (2012)
Richtárik, Takáč (2012)
``Huge-Scale" Coordinate Descent

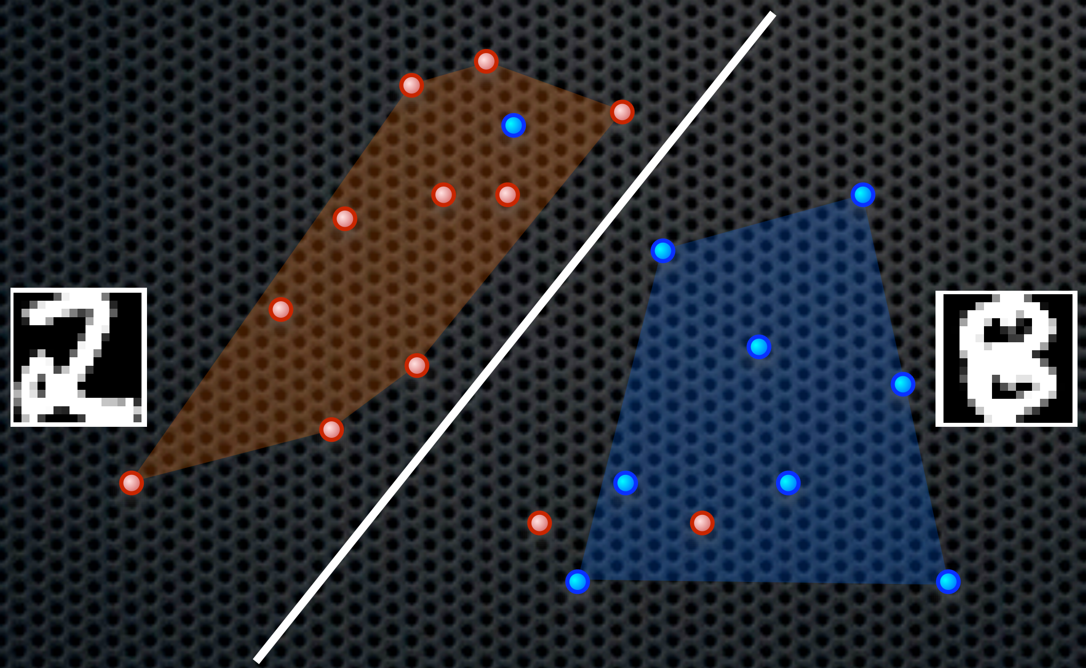
Theorem:
Algorithm obtains **accuracy**
$$O\left(\frac{2n}{k+2n}\right)$$
after k steps.

Hidden constant:
Curvature
$\leq \sum_i L_f \, diam^2(\mathcal{D}^{(i)})$

(also in **duality gap**, and with **inexact subproblems**)

Applications: Large Margin Prediction

- Binary Support Vector Machine
  (no bias)
- also: Ranking SVM

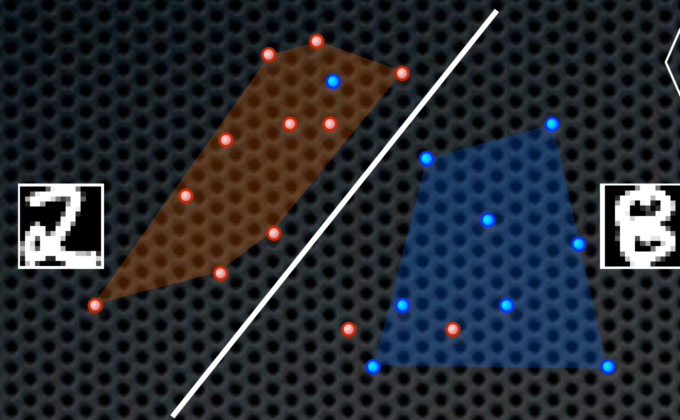


$$\langle \boldsymbol{w}, \phi(\boldsymbol{x}_i)\, \boldsymbol{y}_i \rangle \geq 1 - \xi_i$$

primal problem:

$$\min_{\boldsymbol{w}} \quad \frac{\lambda}{2} \|\boldsymbol{w}\|^2 + \frac{1}{n} \sum_{i=1}^{n} \max \left\{ 0, 1 - \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}_i)\, \boldsymbol{y}_i \rangle \right\}$$

Binary SVM



$\langle \boldsymbol{w}, \phi(\boldsymbol{x}_i)\, \boldsymbol{y}_i \rangle \geq 1 - \xi_i$

primal

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \quad \frac{\lambda}{2} \|\boldsymbol{w}\|^2 + \frac{1}{n} \sum_{i=1}^{n} \max \left\{ 0, 1 - \langle \boldsymbol{w}, \underbrace{\boldsymbol{\phi}(\boldsymbol{x}_i)\, \boldsymbol{y}_i}_{i\text{-th column of } A} \rangle \right\}$$

- *d*-dim
- unconstrained
- *non*-smooth, strongly convex

dual

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \quad f(\boldsymbol{\alpha}) := \frac{\lambda}{2} \|A\boldsymbol{\alpha}\|^2 - \boldsymbol{b}^T \boldsymbol{\alpha}$$
$$\text{s.t.} \quad 0 \leq \alpha_i \leq 1 \quad \forall i \in [n]$$

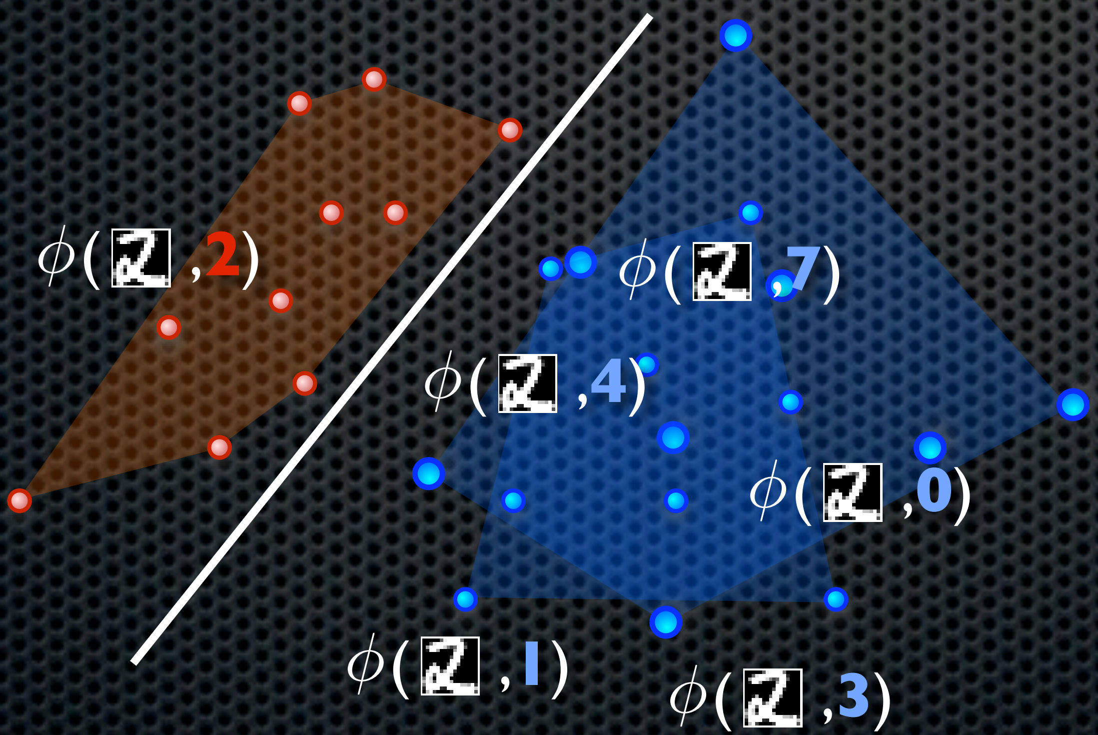
- *n*-dim
- box-constrained
- smooth, *not* strongly convex

Structural SVM

``joint'' feature map $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$

large margin ``separation''

$$\langle \boldsymbol{w}, \phi(\boldsymbol{x}_i, \boldsymbol{y}_i) - \phi(\boldsymbol{x}_i, \boldsymbol{y}) \rangle \geq L(\boldsymbol{y}, \boldsymbol{y}_i) - \xi_i \quad \forall \boldsymbol{y}$$



primal problem:

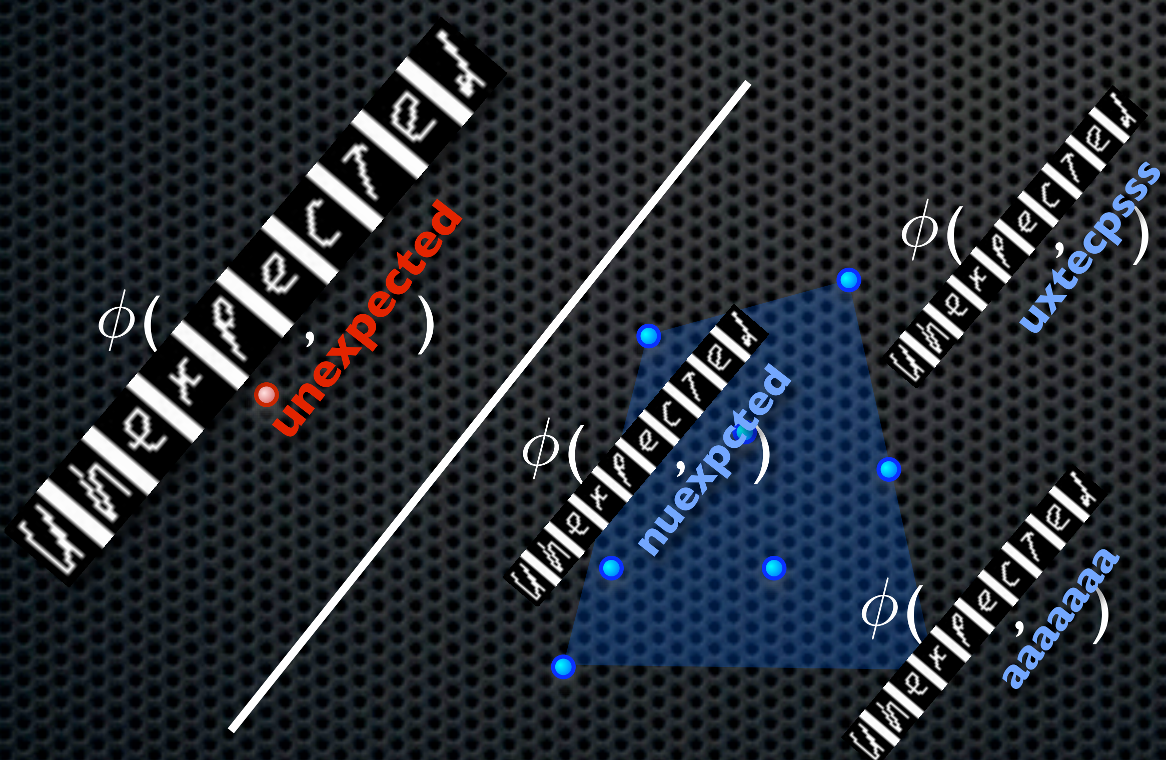
$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \quad \frac{\lambda}{2} \|\boldsymbol{w}\|^2 + \frac{1}{n} \sum_{i=1}^{n} \max_{\boldsymbol{y} \in \mathcal{Y}} \left\{ L(\boldsymbol{y}_i, \boldsymbol{y}) - \langle \boldsymbol{w}, \underbrace{\boldsymbol{\phi}(\boldsymbol{x}_i, \boldsymbol{y}_i) - \boldsymbol{\phi}(\boldsymbol{x}_i, \boldsymbol{y})}_{(i, \boldsymbol{y})\text{-th column of } A} \rangle \right\}$$

Structural SVM

``joint'' feature map $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$

large margin ``separation''

$$\langle \boldsymbol{w}, \phi(\boldsymbol{x}_i, \boldsymbol{y}_i) - \phi(\boldsymbol{x}_i, \boldsymbol{y}) \rangle \geq L(\boldsymbol{y}, \boldsymbol{y}_i) - \xi_i \quad \forall \boldsymbol{y}$$



primal problem:

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \quad \frac{\lambda}{2} \|\boldsymbol{w}\|^2 + \frac{1}{n} \sum_{i=1}^{n} \max_{\boldsymbol{y} \in \mathcal{Y}} \left\{ L(\boldsymbol{y}_i, \boldsymbol{y}) - \langle \boldsymbol{w}, \underbrace{\boldsymbol{\phi}(\boldsymbol{x}_i, \boldsymbol{y}_i) - \boldsymbol{\phi}(\boldsymbol{x}_i, \boldsymbol{y})}_{(i,\boldsymbol{y})\text{-th column of } A} \rangle \right\}$$

decoding oracle

donaudampfschifffahrtsgesellschaftskapitän $|\mathcal{Y}| = 26^{42}$

Binary SVM

primal

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \quad \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n} \max\left\{0, 1 - \langle \boldsymbol{w}, \underbrace{\boldsymbol{\phi}(\boldsymbol{x}_i)\,\boldsymbol{y}_i}_{i\text{-th column of } A} \rangle\right\}$$

dual

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \quad f(\boldsymbol{\alpha}) := \frac{\lambda}{2}\|A\boldsymbol{\alpha}\|^2 - \boldsymbol{b}^T\boldsymbol{\alpha}$$

$$\text{s.t.} \quad 0 \le \alpha_i \le 1 \quad \forall i \in [n]$$

primal-dual correspondence

$$\boldsymbol{w} = A\boldsymbol{\alpha}$$

Structural SVM

primal

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \quad \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n} \max_{\boldsymbol{y}\in\mathcal{Y}}\left\{L(\boldsymbol{y}_i, \boldsymbol{y}) - \langle \boldsymbol{w}, \underbrace{\boldsymbol{\phi}(\boldsymbol{x}_i, \boldsymbol{y}_i) - \boldsymbol{\phi}(\boldsymbol{x}_i, \boldsymbol{y})}_{(i,\boldsymbol{y})\text{-th column of } A} \rangle\right\}$$

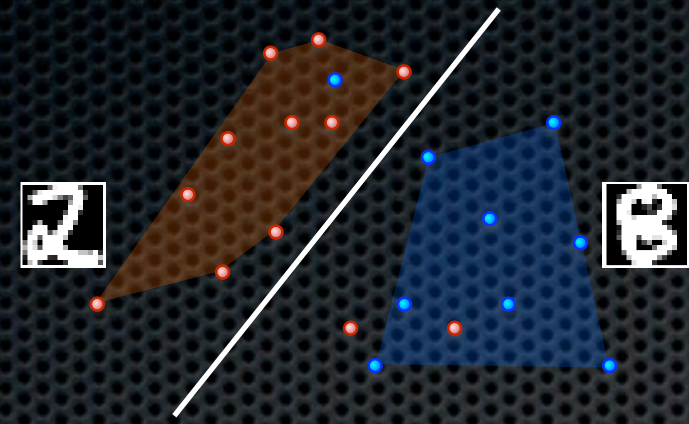
dual

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^{n\cdot|\mathcal{Y}|}} \quad f(\boldsymbol{\alpha}) := \frac{\lambda}{2}\|A\boldsymbol{\alpha}\|^2 - \boldsymbol{b}^T\boldsymbol{\alpha}$$

$$\text{s.t.} \quad \textstyle\sum_{\boldsymbol{y}\in\mathcal{Y}} \alpha_i(\boldsymbol{y}) = 1 \quad \forall i \in [n]$$

$$\text{and} \quad \alpha_i(\boldsymbol{y}) \ge 0 \quad \forall i \in [n], \forall \boldsymbol{y} \in \mathcal{Y}$$

Binary SVM



primal

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \quad \frac{\lambda}{2}\left\|\boldsymbol{w}\right\|^2 + \frac{1}{n}\sum_{i=1}^{n} \max\left\{0, 1 - \langle \boldsymbol{w}, \underbrace{\boldsymbol{\phi}(\boldsymbol{x}_i)\,\boldsymbol{y}_i}_{i\text{-th column of } A} \rangle\right\}$$

- d-dim
- unconstrained
- *non*-smooth, strongly convex

dual

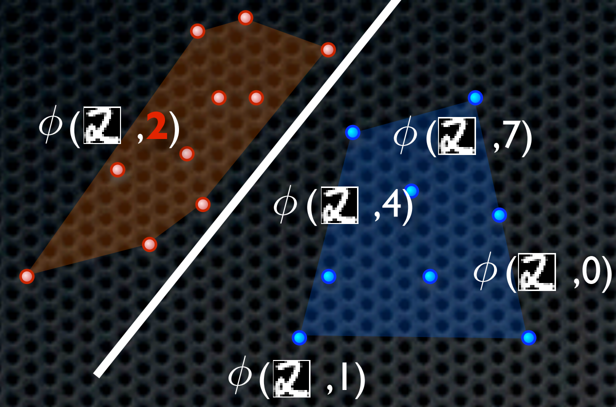
$$\begin{aligned} \min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \quad & f(\boldsymbol{\alpha}) \; := \; \frac{\lambda}{2}\left\|A\boldsymbol{\alpha}\right\|^2 - \boldsymbol{b}^T\boldsymbol{\alpha} \\ \text{s.t.} \quad & 0 \le \alpha_i \le 1 \quad \forall i \in [n] \end{aligned}$$

- n-dim
- box-constrained
- smooth, *not* strongly convex

Optimization Algorithms

	primal		***dual***
batch (**n** cost per iteration)	• subgradient descent	$O(\frac{R^2}{\lambda\varepsilon})$	• Frank-Wolfe =cutting plane (*SVM-light*)
online (**1** cost per iteration)	• stochastic subgradient *(SGD, Pegasos)*		• coordinate descent *(Hsieh 2008)* =block-coordinate descent =block-coordinate Frank-Wolfe

Structural SVM



primal

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \quad \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n}\max_{\boldsymbol{y}\in\mathcal{Y}}\left\{L(\boldsymbol{y}_i,\boldsymbol{y}) - \langle \boldsymbol{w},\ \underbrace{\boldsymbol{\phi}(\boldsymbol{x}_i,\boldsymbol{y}_i) - \boldsymbol{\phi}(\boldsymbol{x}_i,\boldsymbol{y})}_{(i,\boldsymbol{y})\text{-th column of } A}\rangle\right\}$$

- *d*-dim
- unconstrained
- *non*-smooth, strongly convex

dual

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^{n\cdot|\mathcal{Y}|}} \quad f(\boldsymbol{\alpha}) := \frac{\lambda}{2}\|A\boldsymbol{\alpha}\|^2 - \boldsymbol{b}^T\boldsymbol{\alpha}$$
$$\text{s.t.} \quad \textstyle\sum_{\boldsymbol{y}\in\mathcal{Y}} \alpha_i(\boldsymbol{y}) = 1 \quad \forall i\in[n]$$
$$\text{and} \quad \alpha_i(\boldsymbol{y}) \geq 0 \quad \forall i\in[n],\ \forall \boldsymbol{y}\in\mathcal{Y}$$

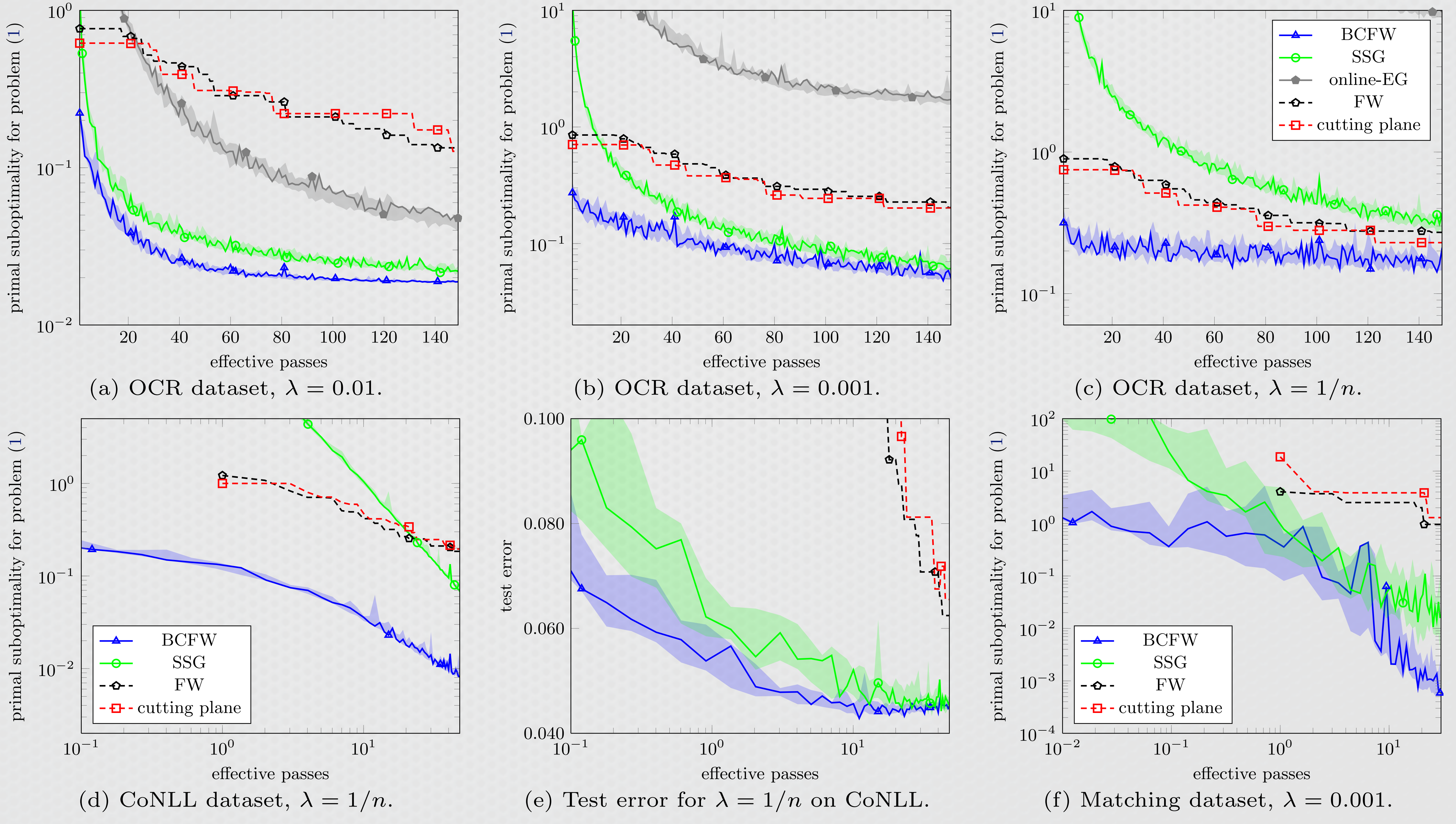
- *n* |*Y*| - dim
- block-constrained
- smooth, *not* strongly convex

Optimization Algorithms

	primal		***dual***
batch (**n** cost per iteration)	• subgradient descent	$O(\frac{R^2}{\lambda\varepsilon})$	• Frank-Wolfe =cutting plane (*SVM-struct*)
online (**1** cost per iteration)	• stochastic subgradient *(SGD, Pegasos)*		• ~~block coordinate descent (*Nesterov*)~~ • **block-coordinate Frank-Wolfe**

Experimental Results

dataset		n	d
OCR	*sequence labeling*	6251	4028
CoNLL	*POS sequence labeling*	8936	1643026
Matching	*word alignment*	5000	82



(a) OCR dataset, $\lambda = 0.01$.
(b) OCR dataset, $\lambda = 0.001$.
(c) OCR dataset, $\lambda = 1/n$.
(d) CoNLL dataset, $\lambda = 1/n$.
(e) Test error for $\lambda = 1/n$ on CoNLL.
(f) Matching dataset, $\lambda = 0.001$.

Thanks!

Co-Authors:
Simon Lacoste-Julien, Mark Schmidt and Patrick Pletscher

Block-Coordinate Frank-Wolfe Optimization for Structural SVMs
Lacoste-Julien, S., Jaggi, M*., Schmidt, M., & Pletscher, P.*
ICML 2013

Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization
Jaggi, M.
ICML 2013

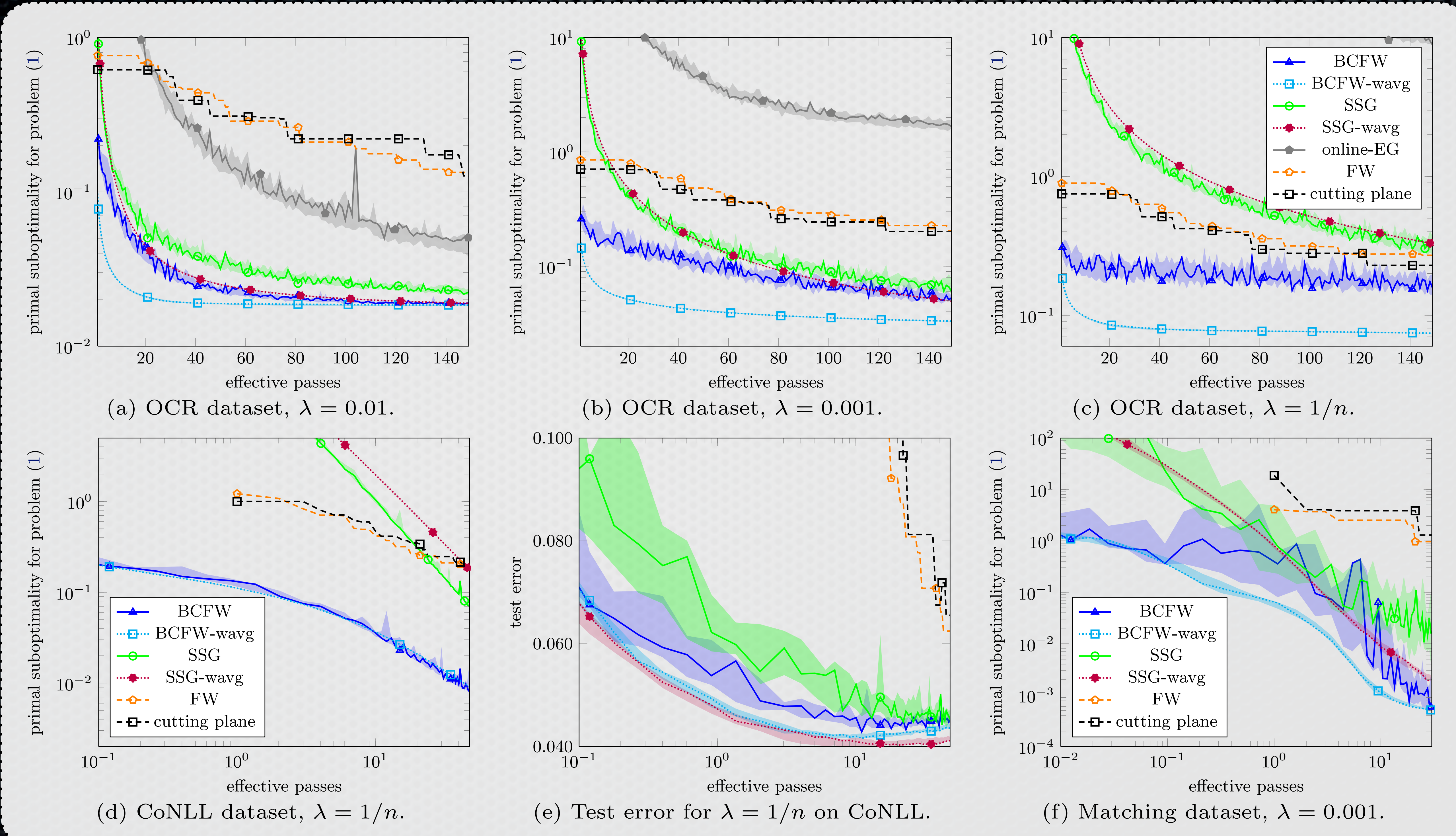
Related Work

Table 1. Convergence rates given in the *number of calls to the oracles* for different optimization algorithms for the structural SVM objective (1) in the case of a Markov random field structure, to reach a specific accuracy ε measured for different types of gaps, in term of the number of training examples n, regularization parameter λ, size of the label space $|\mathcal{Y}|$, maximum feature norm $R := \max_{i,\boldsymbol{y}} \|\boldsymbol{\psi}_i(\boldsymbol{y})\|_2$ (some minor terms were ignored for succinctness). Table inspired from (Zhang et al., 2011). Notice that only stochastic subgradient and our proposed algorithm have rates independent of n.

Optimization algorithm	Online	Primal/Dual	Type of guarantee	Oracle type	# Oracle calls
dual extragradient (Taskar et al., 2006)	no	primal-"dual"	saddle point gap	Bregman projection	$O\left(\frac{nR\log\|\mathcal{Y}\|}{\lambda\varepsilon}\right)$
online exponentiated gradient (Collins et al., 2008)	yes	dual	expected dual error	expectation	$O\left(\frac{(n+\log\|\mathcal{Y}\|)R^2}{\lambda\varepsilon}\right)$
excessive gap reduction (Zhang et al., 2011)	no	primal-dual	duality gap	expectation	$O\left(nR\sqrt{\frac{\log\|\mathcal{Y}\|}{\lambda\varepsilon}}\right)$
BMRM (Teo et al., 2010)	no	primal	$\geq$primal error	maximization	$O\left(\frac{nR^2}{\lambda\varepsilon}\right)$
1-slack SVM-Struct (Joachims et al., 2009)	no	primal-dual	duality gap	maximization	$O\left(\frac{nR^2}{\lambda\varepsilon}\right)$
stochastic subgradient (Shalev-Shwartz et al., 2010)	yes	primal	primal error w.h.p.	maximization	$\tilde{O}\left(\frac{R^2}{\lambda\varepsilon}\right)$
this paper: stochastic block-coordinate Frank-Wolfe	yes	primal-dual	expected duality gap	maximization	$O\left(\frac{R^2}{\lambda\varepsilon}\right)$ Thm. 3

Experimental Results (with averaging)

dataset		n	d
OCR	*sequence labeling*	6251	4028
CoNLL	*POS sequence labeling*	8936	1643026
Matching	*word alignment*	5000	82



(a) OCR dataset, $\lambda = 0.01$. (b) OCR dataset, $\lambda = 0.001$. (c) OCR dataset, $\lambda = 1/n$.

(d) CoNLL dataset, $\lambda = 1/n$. (e) Test error for $\lambda = 1/n$ on CoNLL. (f) Matching dataset, $\lambda = 0.001$.