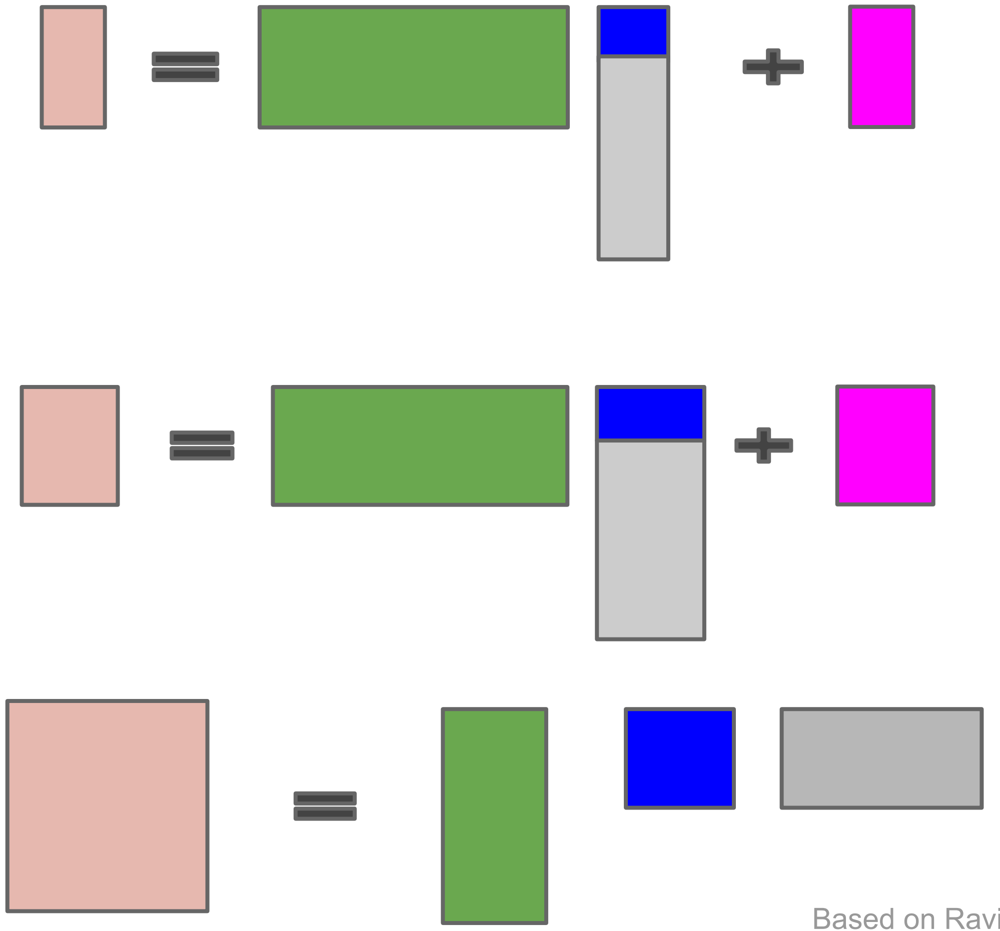
1. **Greedy Algorithms for Structurally Constrained High Dimensional Problems**

2. **Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization**

Presenter: Konstantina Christakopoulou
CSci 8980

Motivation



Goal: Unifying computational framework for high-dimensional structured problems

Based on Ravikumar's presentation in IMA 2013

Statistical Framework: Atomic Sets

Given a set of 'atoms' A

$$x = \sum_{i=1}^{k} c_i a_i,\ a_i \in A$$

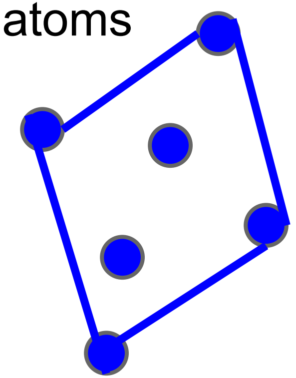
If C_A convex hull of A, **Gauge function**: $\|x\|_A = \inf\{t \geq 0 : x \in tC_A\}$

When A is centrosymmetric,

$$\|x\|_A = \inf\left\{\sum_{a \in A} |c_a| : x = \sum_{a \in A} c_a \cdot a\right\}$$

Support function: $\|x\|_A^* = \sup\{\langle x, a\rangle : a \in A\}$

D: convex hull of atoms



Any *linear* function will attain its *minimum* over *D* at an *atom* s: *A*

Greedy algorithm (Tewari et al. 2011)

$$\min_{x:\|x\|_A \leq \kappa} f(x)$$

where f is **convex, smooth** and atomic norm is bounded $\{x : \|x\|_A \leq \kappa\}$ and f : goodness of fit measure.

Greedy Algorithm to minimize convex function f over κ scaled atomic norm ball

Let $x^0 = \kappa a_0$ for an aribitrary atom $a_0 \in A$

for $t = 0, \ldots,$ **do**

$$a_t = \arg\min_{a \in A} \langle \nabla f(x_t), a \rangle \quad (1)$$

(Line search for step size) $\gamma_t = \arg\min_{\gamma \in [0,1]} f(x_t + \gamma(\kappa a_t - x_t))$

(Update as linear combination) $x_{t+1} = x_t + \gamma_t(\kappa a_t - x_t)$

end for

- ➔ Add atom at every step
- ➔ Iterate x_t : conv. combination of at most t+1 atoms
- ➔ Select atom that makes the optimization problem easy

Contributions (Tewari et al. 2011)

Restricted Smoothness

<u>in high dimensions not good smoothness constants</u>

Restricted Smoothness Property

$$L_{\|\cdot\|}(f;S) = \sup_{x,y\in S,\alpha\in(0,1]} \frac{f((1-\alpha x)+\alpha y) - f(x) - \langle \nabla f(x), \alpha(y-x)\rangle}{\alpha^2\|y-x\|^2}$$

$L_{\|\cdot\|}(f;S) \geq 0$ (since f is convex). How smooth in S with respect to $\|\cdot\|$
Connection with L-Lipschitz: $L_{\|\cdot\|}(f;S) \leq L$ (cares about smoothness just in S).
Connection with Hessian: $\||\nabla^2 f(z)\|| \leq H$

Convergence results:

Convergence

$$f(x_T) - f(x^*) \leq \frac{8\kappa^2 L_{\|\|}(f;\kappa C_A)\|A\|^2}{T}$$

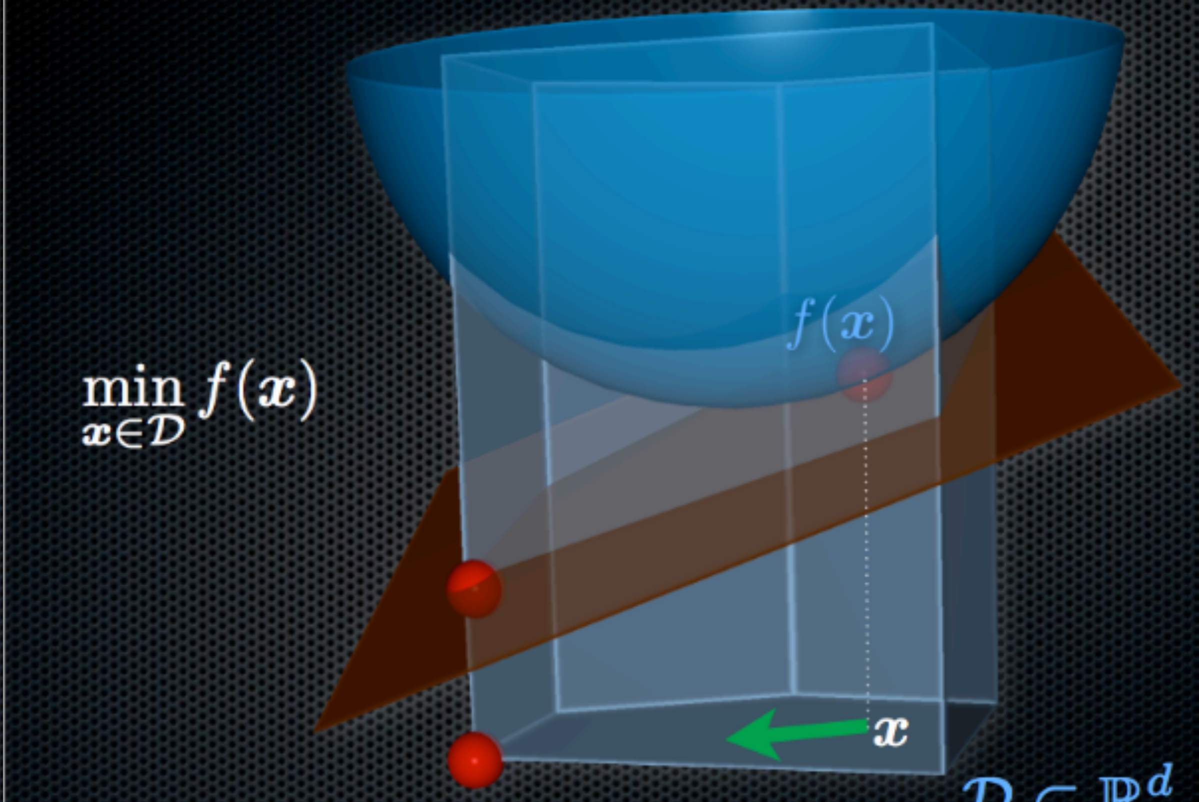
$\|A\| = sup_{a\in A}\|a\|$

For Banach spaces:

Extension to infinite dimensional Banach spaces

V: Banach space equipped with inner product. ∇f : Fechel derivative: elements of the dual space $V*$, inner product $\langle X, x\rangle = X(x)$,x $\in V, X \in V*$
$L_{\|\cdot\|}(f;S) = \sup_{x,y\in S,\alpha\in(0,1]} \frac{f((1-\alpha x)+\alpha y)-f(x)-\langle\nabla f(x),\alpha(y-x)\rangle}{\frac{1}{r}\alpha^r\|y-x\|^r}, r \in [1,2]$
$\rightarrow O(\frac{1}{T^{r-1}})$

Revisiting Frank - Wolfe (Jaggi 2013)



Visualization: from Jaggi's presentation in Smile, Paris Seminar, 2013

Frank - Wolfe

Algorithm 1 Frank-Wolfe (1956)

Let $x^0 \in D$

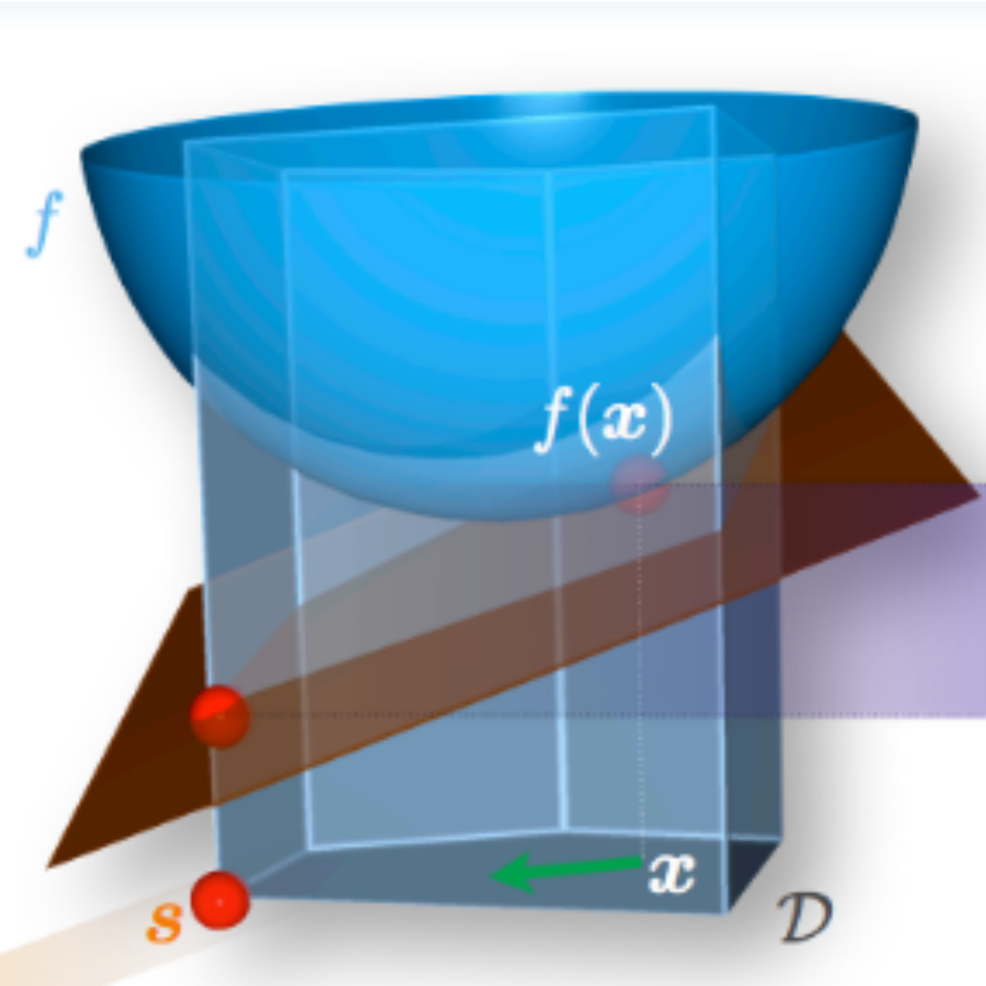
for $k = 0, \ldots, K$ **do**

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

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

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

end for



The Duality gap and Certificates

Primal problem:

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

Surrogate duality gap.

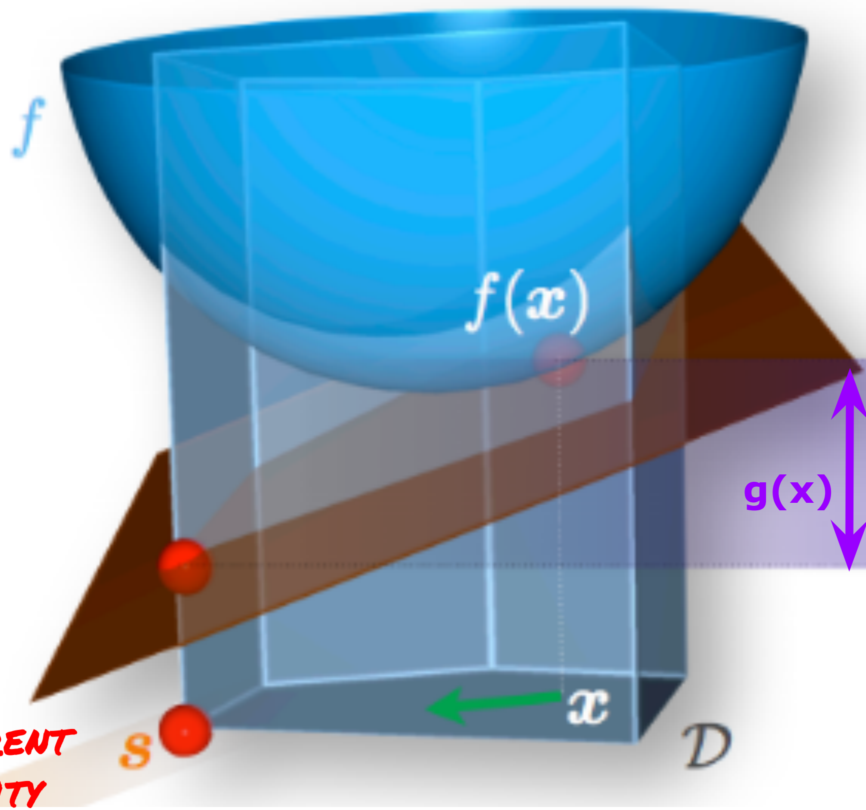
$$g(x) = \max_{s' \in D} \langle x - s', \nabla f(x) \rangle$$

Linearized problem

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

$$g(x) \geq f(x) - f(x*)$$

CERTIFICATE FOR CURRENT APPROXIMATION QUALITY



Based on Jaggi's presentation in Smile, Paris Seminar, 2013

Frank-Wolfe Variants

Algorithm 1 Frank-Wolfe (1956)

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

Algorithm 2 Frank-Wolfe with Approximate Linear Subproblems, for Quality $\delta \geq 0$

Let $x^0 \in D$
for $k = 0, \dots, K$ **do**
Let $\gamma = \frac{2}{k+2}$
Find $s \in D$ s.t

$$\langle s, \nabla f(x^{(k)}) \rangle \leq \min_{s' \in D} \langle s', \nabla f(x^{(k)}) \rangle + \frac{1}{2}\delta\gamma C_f$$

a) Optionally: perform line search for γ
b) Update $x^{(k+1)} = (1-\gamma)x^{(k)} + \gamma s$
end for

Algorithm 3 Line-Search for the step size γ

As Algorithm 2, except replacing line a) with :

$$\gamma = \arg\min_{\gamma \in [0,1]} f\left(x^{(k)} + \gamma(s - x^{(k)})\right)$$

Algorithm 4 Fully Corrective, Re-optimizing over all previous directions

As Algorithm 2, except replacing line b) with :
Do the update

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

Contributions (Jaggi 2013)

The Curvature

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

Connection with L-Lipschitz: $C_f \leq \mathrm{diam}_{\|\cdot\|}(D^2)L$

Convergence results:

Primal Convergence

$$f(x^{(k)}) - f(x^*) \leq \frac{2C_f}{k+2}(1+\delta)$$

δ accuracy for solving the linear sub-problems.

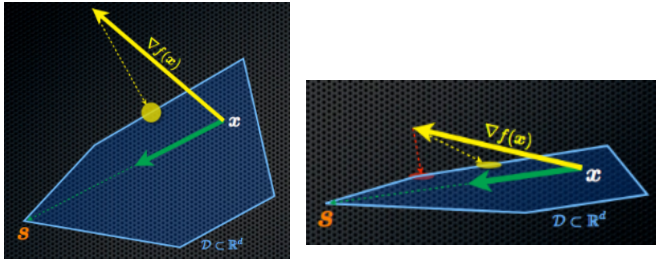
Primal-Dual Convergence

$$g(x^{(\hat{k})}) \leq \frac{7C_f}{K+2}(1+\delta)$$

where $1 \leq \hat{k} \leq K$

CONTRIBUTION NO.1

1. Duality gap convergence guarantee
2. Affine invariance



3. Optimality in terms of sparsity

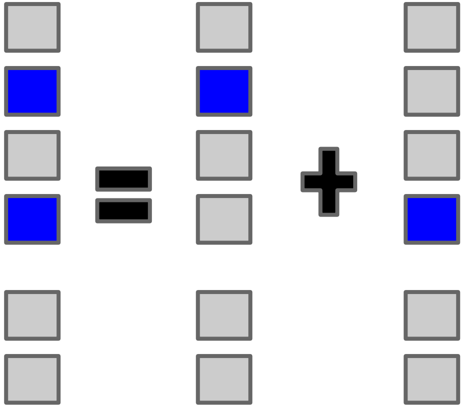
The obtained sparsity k is **optimal** for an approximation quality of 1/k

Frank - Wolfe on Atomic Domains

$\mathcal{X}$	Optimization Domain Atoms $\mathcal{A}$	$\mathcal{D} = \text{conv}(\mathcal{A})$	Complexity of one Frank-Wolfe Iteration $\sup_{s \in \mathcal{D}} \langle s, y \rangle$	Complexity
$\mathbb{R}^n$	Sparse Vectors	$\lVert . \rVert_1$-ball	$\lVert y \rVert_\infty$	$O(n)$
$\mathbb{R}^n$	Sign-Vectors	$\lVert . \rVert_\infty$-ball	$\lVert y \rVert_1$	$O(n)$
$\mathbb{R}^n$	ℓ_p-Sphere	$\lVert . \rVert_p$-ball	$\lVert y \rVert_q$	$O(n)$
$\mathbb{R}^n$	Sparse Non-neg. Vectors	Simplex Δ_n	$\max_i \{y_i\}$	$O(n)$
$\mathbb{R}^n$	Latent Group Sparse Vec.	$\lVert . \rVert_{\mathcal{G}}$-ball	$\max_{g \in \mathcal{G}} \lVert y_{(g)} \rVert_g^*$	$\sum_{g \in \mathcal{G}} \lvert g \rvert$
$\mathbb{R}^{m \times n}$	Matrix Trace Norm	$\lVert . \rVert_{tr}$-ball	$\lVert y \rVert_{op} = \sigma_1(y)$	$\tilde{O}(N_f / \sqrt{\varepsilon'})$ (Lanczos)
$\mathbb{R}^{m \times n}$	Matrix Operator Norm	$\lVert . \rVert_{op}$-ball	$\lVert y \rVert_{tr} = \lVert (\sigma_i(y)) \rVert_1$	SVD
$\mathbb{R}^{m \times n}$	Schatten Matrix Norms	$\lVert (\sigma_i(.)) \rVert_p$-ball	$\lVert (\sigma_i(y)) \rVert_q$	SVD
$\mathbb{R}^{m \times n}$	Matrix Max-Norm	$\lVert . \rVert_{\max}$-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	$\{x \succeq 0, \text{Tr}(x) = 1\}$	$\lambda_{\max}(y)$	$\tilde{O}(N_f / \sqrt{\varepsilon'})$ (Lanczos)
$\mathbb{S}^{n \times n}$	PSD matrices of bounded diagonal	$\{x \succeq 0, x_{ii} \leq 1\}$		$\tilde{O}(N_f n^{1.5} / \varepsilon'^{2.5})$

Table: from Jaggi's presentation in Smile, Paris Seminar, 2013

Special case: Sparse vectors



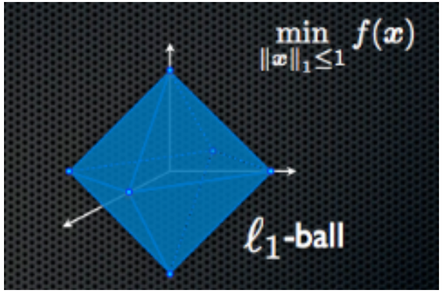
Set of atoms: $A = \{\pm e_i | i \in [n]\}$

$\mathrm{Conv}(A)$ = unit ball of ℓ_1 norm

Compute

$$a_t = \arg \min_{a \in \pm\{e_1, \ldots, e_p\}} \langle \nabla f(x_t), a \rangle$$

Greedy coordinate descent: Find $j = \arg\max_{j' \in \{1,\ldots,p\}} |[\nabla f(x_t)]_{j'}|$ and set $a_t = -\mathrm{sign}([\nabla f(x_t)]_j])e_j$



Obtain $O(1/k)$ approximate solution of sparsity k
Equivalent to (Orthogonal) Matching Pursuit.
Original Frank-Wolfe to polyhedral sets.

Visualization: from Jaggi's presentation in Smile, Paris Seminar, 2013

Special case: sparse non-negative vectors

Set of atoms: $A = \{e_i | i \in [n]\}$

$\text{Conv}(A) = \text{Simplex}$

Compute

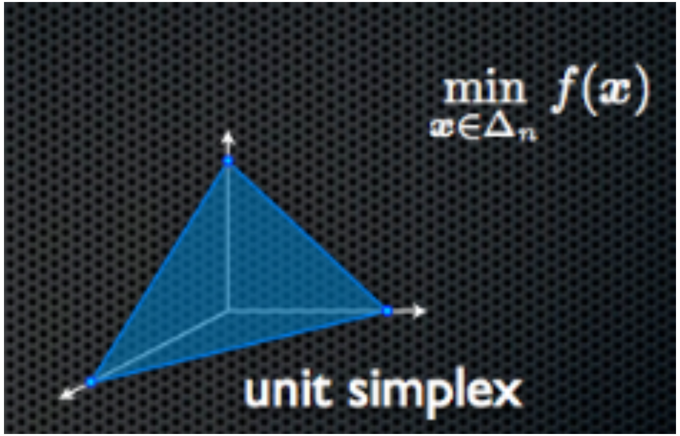
$$a_t = \arg\min_{a \in \pm\{e_1,\dots,e_p\}} \langle \nabla f(x_t), a \rangle$$

Find $j = \arg\min j' \in \{1,\dots,p\}[\nabla f(x_t)]_{j'}$ and set $a_t = e_j$

Clarkson 2010.

Tradeoff between **sparsity** and **approximation quality**

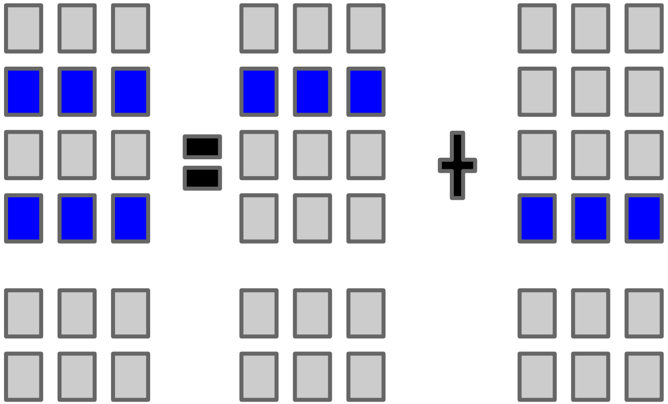
Sparsity of FW iterates is **Optimal** in primal and dual approximation quality



If we choose || || to be l_1 norm, then restricted smoothness constant is similar to C_f

Visualization: from Jaggi's presentation in Smile, Paris Seminar, 2013

Special case: Group Sparse Matrices

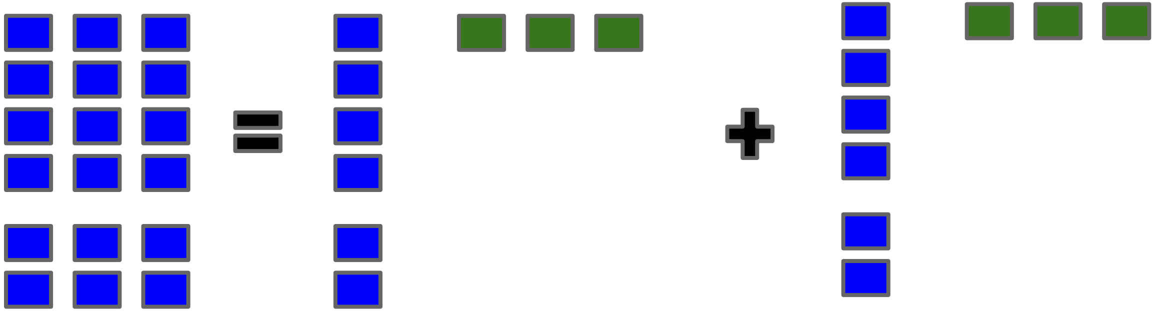


Infinite Set of atoms: All matrices with a single non-zero row where the row has ℓ_q norm 1 for some $q > 1$
Conv(A) = unit ball of the $\|\cdot\|_{q,1}$ group norm on $R^{p\times k}$ (sum of ℓ_q norms on the rows)
Compute

$$a_t = \arg\min_{a:\mathrm{nnzrows}(a)=1, \|a\|_{q,1}=1} \langle \nabla f(x_t), a\rangle$$

Find row j of $\nabla f(x_t)$ with maximal $\ell_{q'}$ norm. Set a_t to be the matrix with the only non zero row in j equal to u^T such that $\langle u, [\nabla f(x_t)]^T_{j,:}\rangle = -\|[\nabla f(x_t)]^T_{j,:}\|_{q'}$

Special case: Low rank matrices



Infinite Set of atoms: All rank 1 matrices with Frobenius norm $= 1$

$$A := \left\{ uv^T | u \in R^n, \|u\|_2 = 1, v \in R^n, \|v\|_2 = 1 \right\}$$

Conv(A) $=$ unit ball of the Trace norm (Schatten ℓ_1 norm)
Greedy step: Compute

$$a_t = \arg \min_{a:\mathrm{rank}(a)=1, \|a\|_F=1} \langle \nabla f(x_t), a \rangle$$

Compute SVD $\nabla f(x_t) = U \Sigma V^T$ and set $a = -u_1 v_1^T$, where u_1, v_1 left and right singular vectors corresponding to largest singular value.

Polynomial time

- (Tewari et al 2011) Not polynomial time to compute greedy step for non negative low rank matrices
- Permutation matrices: Efficient optimization over Birkoff polytope, Hungarian Algorithm (Conv(A) = set of doubly stochastic matrices)
- (Jaggi 2013)

Novel framework for factorized matrix norms

Novel framework for matrix factorization:

$$A = \left\{ LR^T \middle| L \in A_{\text{left}}, R \in A_{\text{right}} \right\},$$

$A_{\text{left}} \subseteq R^{m \times r}, A_{\text{right}} \subseteq R^{n \times r}$

All Frank-Wolfe iterates are low-rank updates: $s = LR^T$

Thanks

References

1. Greedy Algorithms for Structurally Constrained High Dimensional Problems, A. Tewari, P. Ravikumar, I. Dhillon. In Advances in Neural Information Processing Systems (NIPS) 24, 2011.

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