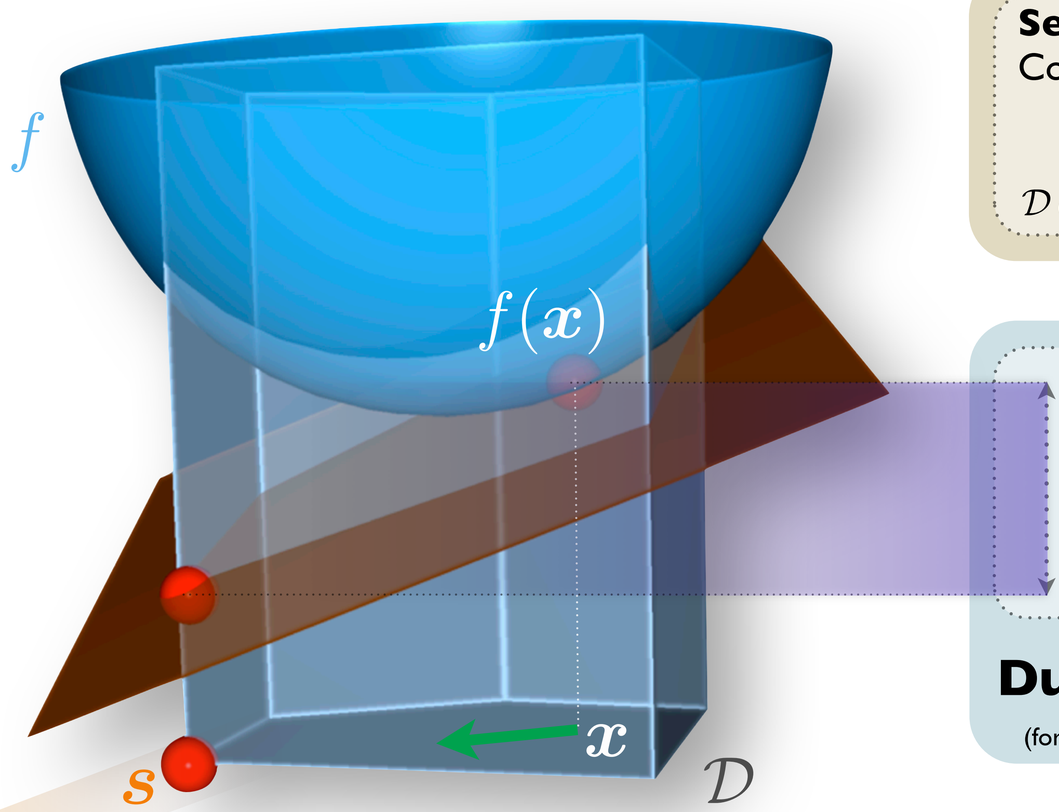
Revisiting Frank-Wolfe:
Projection-Free Sparse Convex Optimization

Martin Jaggi
CMAP - École Polytechnique,
ERC SIPA project,
CNRS UMR 7641, Paris, France

Introduction: There are two types of ***first-order methods*** for constrained convex optimization. One of them became nearly forgotten in the last decades.

Contributions: Stronger and more general primal-dual convergence results for Frank-Wolfe methods, and a unified view on many variants and applications.



Setup
Constrained convex optim.

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

$\mathcal{D}$ compact, f differentiable

useful & efficient ***certificate*** *for the approximation quality*

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

Duality Gap

(for any bounded constrained probem)

The Frank-Wolfe Algorithm

(or conditional gradient)

Idea: Minimize a *linear* approximation of f

Algorithm 1 Frank-Wolfe

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

Convergence Results

Primal Rate

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

new

Primal-Dual Rate

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

For some $\hat{k} \leq k$

For all algorithm variants!

Curvature constant C_f is bounded by the Lipschitz constant of the gradient, times diameter:
$C_f \leq \mathrm{diam}_{\|.\|}(\mathcal{D})^2 L$ w.r.t. any choice of norm! (our algorithms/analysis are norm-free)

Advantages

	Frank-Wolfe	***Gradient Methods***
Iterates	***Sparse*** ✓ *(using at most k atoms after k iterations)*	***Dense*** ✗
Iteration Cost	***Linear oracle*** *(can be much cheaper)*	***Projection step***
Example: trace-norm	top EV	full SVD

Affine Invariance

The ***algorithms*** and our ***analysis*** are fully ***invariant*** under (affine) transformations of the input task

Contrasting gradient methods which use projections and norms

Optimal Sparsity & Rate

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

No algorithm can do better in general
Lower bounds for *sparsity* (*l1*-domain) and *low-rank* (trace-norm domain)

Algorithm Variants

- ***Approximate subproblems*** and ***inexact gradients*** (or inexact domain) (using approximate linear minimizers $\boldsymbol{s}$)
- ***Line-search*** for the optimal step-size $\gamma \in [0,1]$
- ***Fully corrective*** (re-optimizing over all used $\boldsymbol{s}$)
- ***Away steps*** (removing the worst of all used $\boldsymbol{s}$)

Factorized Matrix Domains

For two sets
$\mathcal{A}_{\text{left}} \subseteq \mathbb{R}^{m\times r}$
$\mathcal{A}_{\text{right}} \subseteq \mathbb{R}^{n\times r}$

Consider the ***outer-product*** matrices

$$\mathcal{A} := \left\{ LR^T \;\middle|\; \begin{matrix} L\in\mathcal{A}_{\text{left}} \,, \\ R\in\mathcal{A}_{\text{right}} \end{matrix} \right\}$$

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

➔ every FW iteration is a ***low-rank*** update

Natural way to optimize over ***matrix factorizations*** (including *sparse* and *non-negative* ones)

r	$\mathcal{A}_{\text{left}} \subseteq \mathbb{R}^{m\times r}$	$\mathcal{A}_{\text{right}} \subseteq \mathbb{R}^{n\times r}$	$\Omega_{\mathrm{conv}(\mathcal{A})}(M)$	$\Omega^*_{\mathcal{A}}(M)$	**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
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 [MK87]
1	Simplex Δ_m	Simplex Δ_n	"non-neg. matrix ℓ_1-norm"		$O(nm)$

Applications *to sparse and low-rank optimization*

Generalized sparsity problems: Usually in machine learning and signal processing applications, we optimize over a domain $\mathcal{D} := \mathrm{conv}(\mathcal{A})$, that is the convex hull of a simple set of things/atoms (***atomic norm*** idea).

For such problems, Frank-Wolfe methods are ***particularly suitable***:

$\mathcal{X}$	**Optimization Domain** Atoms $\mathcal{A}$	$\mathcal{D} = \mathrm{conv}(\mathcal{A})$	**Complexity of one Frank-Wolfe Iteration** $\Omega^*_{\mathcal{D}}(\boldsymbol{y}) = \sup_{\boldsymbol{s}\in\mathcal{D}} \langle \boldsymbol{s}, \boldsymbol{y}\rangle$	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 vect.	$\|.\|_{\mathcal{G}}$-ball	$\max_{g\in\mathcal{G}} \|\boldsymbol{y}_{(g)}\|^*_g$	$\sum_{g\in\mathcal{G}} \|g\|$
$\mathbb{R}^{m\times n}$	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	$\{\boldsymbol{x}\succeq 0,\ \mathrm{Tr}(\boldsymbol{x})=1\}$	$\lambda_{\max}(\boldsymbol{y})$	$\tilde{O}\big(N_f/\sqrt{\varepsilon'}\big)$ (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* $\varepsilon' = \frac{\delta C_f}{k+2}$ *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.*

Many other application such as optimizing over ***structured atomic norms***, ***matrix factorizations***, and ***submodular*** optimization