# Frank-Wolfe Splitting via Augmented Lagrangian Method


**Gauthier Gidel**[1], **Fabian Pedregosa**[2,3] and **Simon Lacoste-Julien**[1]

[1]Université de Montréal, MILA [2]UC Berkeley [3]ETH Zurich


## Overview

### Summary


**Task:** Minimize a function over an *intersection* of convex sets.
**Problem:** Projections or linear minimization oracle (LMO) over the intersection is *expensive*. Projection onto each individual set is *expensive*.
**Our solution:**

- Minimize a **smooth** function over an intersection of constraints.
- Requires **linear minimization oracles** over **individual constraints**.
- Based on the **Augmented Lagrangian Method**.

**Motivating applications:** Multiple sequence alignment. Structured SVM. Simultaneously sparse and low rank matrices.


### Related work

Algorithm (GDMM) proposed in Yen et al. [4, 3], Huang et al. [2] but restricted to polytopes and simple (linear and quadratic) functions.

### Contributions

- Extension of GDMM for general convex sets.
- Major fix of the previous proofs.

## Problem Formulations

### Initial Problem

$$\underset{\boldsymbol{x}^{(1)},\dots,\boldsymbol{x}^{(k)}}{\text{minimize}}\; f(\boldsymbol{x}^{(1)},\dots,\boldsymbol{x}^{(k)})\,,$$

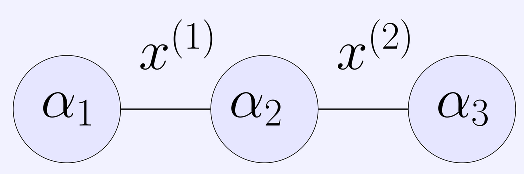
$$\boldsymbol{x}^{(k)} \in \mathcal{X}_k,\; k\in[K],\; \sum_{k=1}^{K} A_k\boldsymbol{x}^{(k)} = 0\,,$$

- $f$ convex smooth function.
- $\mathcal{X}_k$, $k \in [K]$, convex compact sets.
- $A_k \in \mathbb{R}^{d\times d_k}$, $k \in [K]$.

**Examples**

Intersection of two sets:
$n=2,\; A_1 = -A_2 = I_n$
$\Rightarrow \boldsymbol{x}^{(1)} = \boldsymbol{x}^{(2)}$

Marginal polytope:




$A_1x^{(1)} = A_2x^{(2)}$

### Augmented Lagrangian formulation

- Augmented Lagrangian trick to get rid of $\sum_{k=1}^K A_k\boldsymbol{x}^{(k)} = 0$.
- Introduce $M$ s.t. $M\boldsymbol{x} = 0 \Leftrightarrow \sum_{k=1}^K A_k\boldsymbol{x}^{(k)} = 0$ and the function

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{y}) := f(\boldsymbol{x}) + \langle \boldsymbol{y}, M\boldsymbol{x}\rangle + \tfrac{\lambda}{2}\|M\boldsymbol{x}\|^2.$$

- $\max_{\boldsymbol{y}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{x},\boldsymbol{y}) = f(\boldsymbol{x})$ if $M\boldsymbol{x}=0$ or $+\infty$ otherwise.

**Augmented Lagrangian formulation** of our initial problem,

$$\underset{(\boldsymbol{x}^{(1)},\dots,\boldsymbol{x}^{(K)})}{\text{minimize}}\; \max_{\boldsymbol{y}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{x}^{(1)},\dots,\boldsymbol{x}^{(K)},\boldsymbol{y})$$
$$\text{s.t.}\quad \boldsymbol{x}^{(k)} \in \mathcal{X}_k,\quad k\in\{1,\dots,K\}\,.$$

## Algorithm

### FW-AL

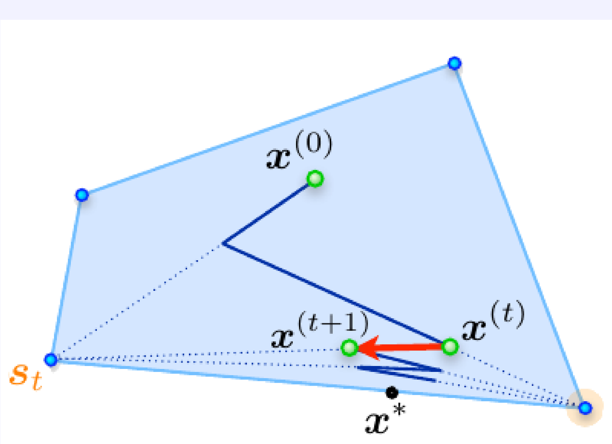
**Standard AL method**

$$\boldsymbol{x}_{t+1} = \arg\min_{\boldsymbol{x}\in\mathcal{X}} \mathcal{L}(\boldsymbol{x},\boldsymbol{y}_t))\,,$$
$$\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \eta_t M\boldsymbol{x}_{t+1}\,,$$

Replace $\arg\min$ steps by FW steps:

$$\boldsymbol{x}_{t+1} = \mathcal{FW}(\boldsymbol{x}_t; \mathcal{L}(\cdot,\boldsymbol{y}_t))\,.$$
$$\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \eta_t M\boldsymbol{x}_{t+1}\,.$$

### FW algorithm




**Algorithm 1** FW (one step) : [1]

**input:** $(\boldsymbol{x},\varphi)$
Compute $\boldsymbol{r} \leftarrow \nabla\varphi(\boldsymbol{x})$
Compute $\boldsymbol{s} \leftarrow \arg\min_{\boldsymbol{s}\in\mathcal{X}} \langle \boldsymbol{s},\boldsymbol{r}\rangle$
$\gamma \in \arg\min_{\gamma\in[0,1]} \varphi(\boldsymbol{x}+\gamma(\boldsymbol{s}-\boldsymbol{x}))$
Update $\boldsymbol{x} \leftarrow (1-\gamma)\boldsymbol{x} + \gamma\boldsymbol{s}$
**return:** $\boldsymbol{x}$

## Theoretical Results

### Assumption

Slater's condition: $\exists\, \boldsymbol{x}^{(k)} \in \operatorname{relint}(\mathcal{X}_k),\; k\in[K]$ s.t. $\sum_{k=1}^{K} A_k\boldsymbol{x}^{(k)} = 0\,.$

### For general convex sets

With decreasing step size $\eta_t := O\left(\frac{1}{t+1}\right)$,

$$\text{subopt:}\;\; \Delta_t \le \frac{O(1)}{t}\,,\qquad \text{feasibility:}\;\; \min_{t_0\le s\le t} \|M\boldsymbol{x}_s\|^2 \le \frac{O(1)}{t}\,.$$

### For $\mathcal{X}$ a polytope

With small enough constant step size $\eta_t$:

$$\Delta_t \le \frac{\Delta_{t_0}}{(1+\kappa)^{t-t_0}}\,,\qquad \|M\boldsymbol{x}_{t+1}\|^2 \le \frac{O(1)}{(1+\kappa)^{t-t_0}}\,.$$

Uses a variant of FW with away-step and holds only for generalized strongly convex function.

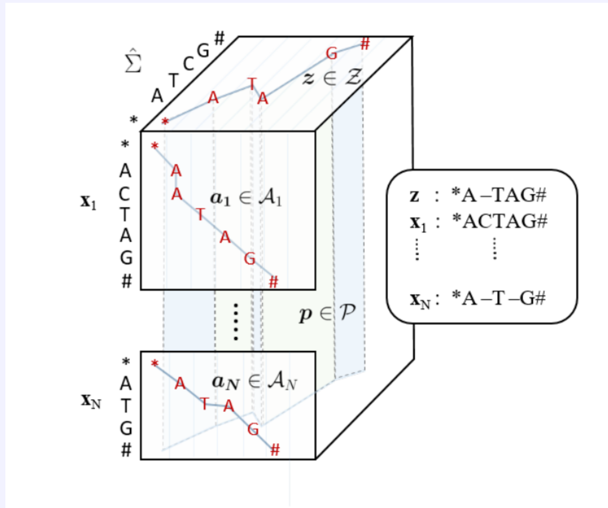
## Motivating Applications

### Multiple Sequence Alignment

Proposed by Yen et al. [4]:

$$\min_{W\in\mathcal{A}\cap\mathcal{P}} \langle W, D\rangle$$

$D$: cost matrix. $\mathcal{A}$ : alignment constraint. $\mathcal{P}$ : consensus constraint.



### Structured SVM

Proposed by Yen et al. [3]:

$$\text{dual problem:}\quad \min_{\alpha_f\in\Delta^{|\mathcal{Y}_f|}} \frac{1}{2}\sum_{F\in\mathcal{T}} \|A_F\alpha\|_2^2 - \sum_{j\in\mathcal{V}} \delta_j^\top\alpha_j$$
$$\text{s.t.}\quad M_{fi}\alpha_f = \alpha_i\,,\quad f\in F,\; F\in\mathcal{T},\; i\in\mathcal{N}(f)\,.$$

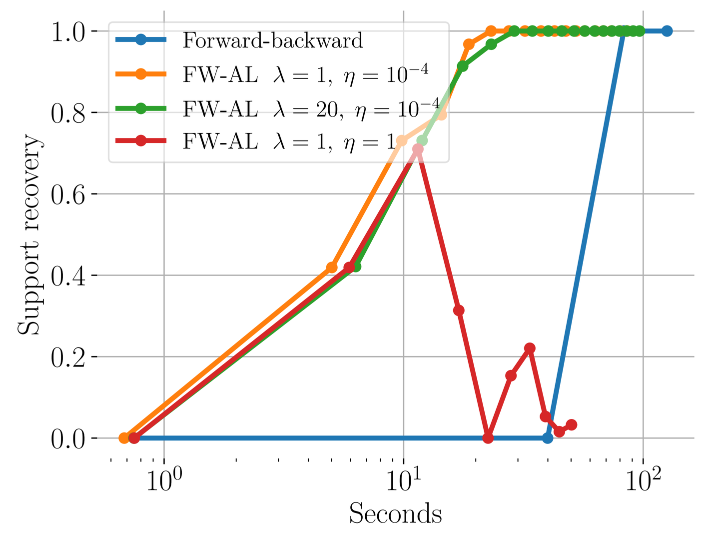
$\mathcal{V}$ : set of variables. $\mathcal{T}$ : set of factor templates. $\mathcal{N}(f)$: neighbors of $f$.

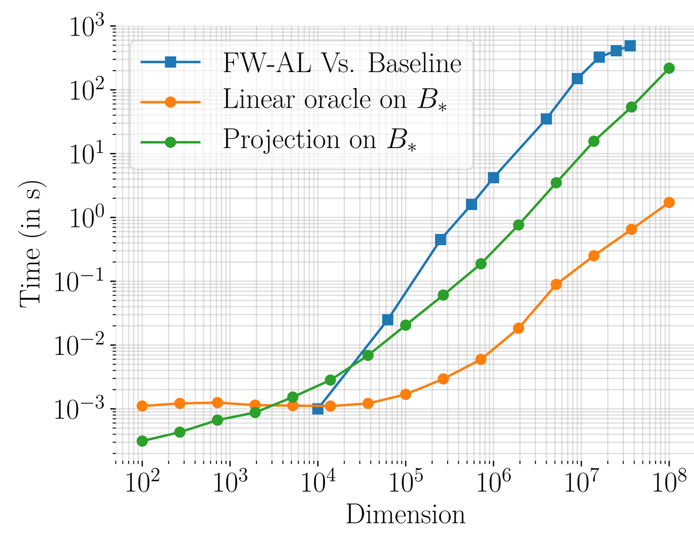
## Simultaneously Sparse and Low Rank Matrices

Given $\hat{\Sigma} \succeq 0$, the objective function is defined as

$$\min_{S\succeq 0,\; \|S\|_1\le\beta_1,\; \|S\|_*\le\beta_2} \|S-\hat{\Sigma}\|_2^2\,.$$

- Oracle over $\ell_1$-ball: largest coefficient $O(d^2)$.
- Oracle over trace-norm all: largest eigen-value $O(d^2/\sqrt{\varepsilon})$ (Lanczos).




Percentage of the support recovered by FW-AL and the Forward backward algorithm as a function of time. As the dimension increase the projection take more and more time to be performed.

## References


[1] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics*, 1956.

[2] X. Huang, I. E.-H. Yen, R. Zhang, Q. Huang, P. Ravikumar, and I. Dhillon. Greedy direction method of multiplier for MAP inference of large output domain. In *AISTATS*, 2017.

[3] I. Yen, X. Huang, K. Zhong, R. Zhang, P. Ravikumar, and I. Dhillon. Dual decomposed learning with factorwise oracle for structural SVM with large output domain. In *NIPS*, 2016b.

[4] I. E.-H. Yen, X. Lin, J. Zhang, P. Ravikumar, and I. Dhillon. A convex atomic-norm approach to multiple sequence alignment and motif discovery. In *ICML*, 2016a.