Frank-Wolfe type methods for nonconvex inequality-constrained problems

Ting Kei Pong
Department of Applied Mathematics
The Hong Kong Polytechnic University
Hong Kong

Workshop on Optimization, Equilibrium and Complementarity
August 2023
(Joint work with Guoyin Li, Liaoyuan Zeng & Yongle Zhang)

Motivating applications

- Matrix completion: (Candés, Recht '09)

$$\min_{x \in \mathbb{R}^{m \times n}} \sum_{(i,j) \in \Omega} (x_{ij} - \bar{x}_{ij})^2 \ \text{subject to} \ \Phi(x) \leq \sigma,$$

where $\bar{x}$ comes from observation, Ω is the index set of observed entries, $\sigma > 0$, and typical choices of $\Phi : \mathbb{R}^{m \times n} \to \mathbb{R}_+$ are:

  - ⋆ $\Phi(x) = \|x\|_*$, the nuclear norm of x;
  - ⋆ $\Phi(x) = \|x\|_* - \mu\|x\|_F$, $\mu \in (0, 1)$.

Motivating applications

- Matrix completion: (Candés, Recht '09)

$$\min_{x \in \mathbb{R}^{m\times n}} \sum_{(i,j)\in\Omega} (x_{ij} - \bar{x}_{ij})^2 \;\text{ subject to }\; \Phi(x) \le \sigma,$$

  where $\bar{x}$ comes from observation, Ω is the index set of observed entries, $\sigma > 0$, and typical choices of $\Phi : \mathbb{R}^{m\times n} \to \mathbb{R}_+$ are:

  ⋆ $\Phi(x) = \|x\|_*$, the nuclear norm of x;

  ⋆ $\Phi(x) = \|x\|_* - \mu\|x\|_F$, $\mu \in (0, 1)$.

- Adversarial (ℓ_p) attack: (Chen, Zhou, Yi, Gu '20)

$$\min_{x\in\mathbb{R}^n} \; h(\bar{x} + x) \;\text{ subject to }\; \|x\|_p^p \le \sigma,$$

  where $\bar{x}$ is a correctly classified data point, h is smooth, $\sigma > 0$, $\|x\|_p^p = \sum_{i=1}^n |x_i|^p$, $p > 0$.

Motivating applications

- Matrix completion: (Candés, Recht '09)

$$\min_{x \in \mathbb{R}^{m \times n}} \sum_{(i,j)\in\Omega} (x_{ij} - \bar{x}_{ij})^2 \ \text{ subject to } \ \Phi(x) \le \sigma,$$

  where $\bar{x}$ comes from observation, Ω is the index set of observed entries, $\sigma > 0$, and typical choices of $\Phi : \mathbb{R}^{m\times n} \to \mathbb{R}_+$ are:

  ⋆ $\Phi(x) = \|x\|_*$, the nuclear norm of x;

  ⋆ $\Phi(x) = \|x\|_* - \mu\|x\|_F$, $\mu \in (0, 1)$.

- Adversarial (ℓ_p) attack: (Chen, Zhou, Yi, Gu '20)

$$\min_{x \in \mathbb{R}^n} \ h(\bar{x} + x) \ \text{ subject to } \ \|x\|_p^p \le \sigma,$$

  where $\bar{x}$ is a correctly classified data point, h is smooth, $\sigma > 0$, $\|x\|_p^p = \sum_{i=1}^n |x_i|^p$, $p > 0$.

- Project onto the constraint sets?

Motivating applications

- Matrix completion: (Candés, Recht '09)

$$\min_{x \in \mathbb{R}^{m \times n}} \sum_{(i,j) \in \Omega} (x_{ij} - \bar{x}_{ij})^2 \text{ subject to } \Phi(x) \le \sigma,$$

  where $\bar{x}$ comes from observation, Ω is the index set of observed entries, $\sigma > 0$, and typical choices of $\Phi : \mathbb{R}^{m \times n} \to \mathbb{R}_+$ are:
  - ⋆ $\Phi(x) = \|x\|_*$, the nuclear norm of x;
  - ⋆ $\Phi(x) = \|x\|_* - \mu \|x\|_F$, $\mu \in (0, 1)$.

- Adversarial (ℓ_p) attack: (Chen, Zhou, Yi, Gu '20)

$$\min_{x \in \mathbb{R}^n} h(\bar{x} + x) \text{ subject to } \|x\|_p^p \le \sigma,$$

  where $\bar{x}$ is a correctly classified data point, h is smooth, $\sigma > 0$, $\|x\|_p^p = \sum_{i=1}^n |x_i|^p$, $p > 0$.

- Project onto the constraint sets?

Motivating applications

- Matrix completion: (Candés, Recht '09)

$$\min_{x \in \mathbb{R}^{m\times n}} \sum_{(i,j)\in\Omega} (x_{ij} - \bar{x}_{ij})^2 \ \text{ subject to } \ \Phi(x) \leq \sigma,$$

  where $\bar{x}$ comes from observation, Ω is the index set of observed entries, $\sigma > 0$, and typical choices of $\Phi : \mathbb{R}^{m\times n} \to \mathbb{R}_+$ are:
  - ⋆ $\Phi(x) = \|x\|_*$, the nuclear norm of x;
  - ⋆ $\Phi(x) = \|x\|_* - \mu\|x\|_F$, $\mu \in (0, 1)$.

- Adversarial (ℓ_p) attack: (Chen, Zhou, Yi, Gu '20)

$$\min_{x \in \mathbb{R}^n} \ h(\bar{x} + x) \ \text{ subject to } \ \|x\|_p^p \leq \sigma,$$

  where $\bar{x}$ is a correctly classified data point, h is smooth, $\sigma > 0$, $\|x\|_p^p = \sum_{i=1}^n |x_i|^p$, $p > 0$.

- Project onto the constraint sets? Alternatives?

Frank-Wolfe method

Let $\mathbb{X}$ be a finite dimensional Hilbert space. Consider

$$\min_{x \in \mathbb{X}} f(x) \text{ subject to } x \in D,$$

where $f \in C^1(\mathbb{X})$ and D is compact convex such that for any $v \in \mathbb{X}$, a

$$u \in \operatorname*{Arg\,min}_{x \in D} \langle v, x \rangle$$

can be *easily* obtained.

Frank-Wolfe method

Let $\mathbb{X}$ be a finite dimensional Hilbert space. Consider

$$\min_{x\in\mathbb{X}} \ f(x) \ \text{ subject to } \ x \in D,$$

where $f \in C^1(\mathbb{X})$ and D is compact convex such that for any $v \in \mathbb{X}$, a

$$u \in \operatorname*{Arg\,min}_{x\in D} \ \langle v, x\rangle$$

can be *easily* obtained.

Examples of D:

- $D = \{x \in \mathbb{R}^n : \ \|x\|_p \leq \sigma\}$ for $p \in [1, \infty]$ and some $\sigma > 0$.
  Then u can be computed by considering the dual norm.

Frank-Wolfe method

Let $\mathbb{X}$ be a finite dimensional Hilbert space. Consider

$$\min_{x\in\mathbb{X}} \; f(x) \;\; \text{subject to} \;\; x \in D,$$

where $f \in C^1(\mathbb{X})$ and D is compact convex such that for any $v \in \mathbb{X}$, a

$$u \in \operatorname*{Arg\,min}_{x\in D} \; \langle v, x\rangle$$

can be *easily* obtained.

Examples of D:

- $D = \{x \in \mathbb{R}^n : \; \|x\|_p \le \sigma\}$ for $p \in [1, \infty]$ and some $\sigma > 0$. Then u can be computed by considering the dual norm.
- $D = \{x \in \mathbb{R}^{m\times n} : \; \|x\|_* \le \sigma\}$ for some $\sigma > 0$. Then $u = -\sigma r_1 s_1^T$, where r_1 and s_1 are the left and right unit singular vectors, respectively, corresponding to the largest singular value of $-v$, obtained via Lanzcos method. In contrast, projecting onto D requires full SVD of v.

Frank-Wolfe method cont.

Frank-Wolfe method for convex D: (Frank, Wolfe '56)

Step 1. Choose $x^0 \in D$. Pick any $c \in (0,1)$ and set $k = 0$.

Step 2. Compute $u^k \in \operatorname{Arg\,min}_{x \in D} \langle \nabla f(x^k), x \rangle$.

Step 3. If $\langle \nabla f(x^k), u^k - x^k \rangle = 0$, terminate.

Step 4. Choose $\alpha_k \in (0,1]$ using backtracking to satisfy

$$f(x^k + \alpha_k(u^k - x^k)) \le f(x^k) + c\alpha_k \langle \nabla f(x^k), u^k - x^k \rangle.$$

Step 5. Set $x^{k+1} = x^k + \alpha_k(u^k - x^k)$, $k \leftarrow k+1$. Go to Step 2.

Frank-Wolfe method cont.

<u>**Frank-Wolfe method**</u> for convex D: (Frank, Wolfe '56)

Step 1. Choose $x^0 \in D$. Pick any $c \in (0,1)$ and set $k = 0$.

Step 2. Compute $u^k \in \operatorname{Arg\,min}_{x \in D} \langle \nabla f(x^k), x \rangle$.

Step 3. If $\langle \nabla f(x^k), u^k - x^k \rangle = 0$, terminate.

Step 4. Choose $\alpha_k \in (0,1]$ using backtracking to satisfy

$$f(x^k + \alpha_k(u^k - x^k)) \le f(x^k) + c\alpha_k \langle \nabla f(x^k), u^k - x^k \rangle.$$

Step 5. Set $x^{k+1} = x^k + \alpha_k(u^k - x^k)$, $k \leftarrow k+1$. Go to Step 2.

Remarks:

- The algorithm either terminates finitely at a stationary point $x^{\bar{k}}$, or every accumulation point of $\{x^k\}$ is stationary.

Frank-Wolfe method cont.

Frank-Wolfe method for convex D: (Frank, Wolfe '56)

Step 1. Choose $x^0 \in D$. Pick any $c \in (0,1)$ and set $k = 0$.

Step 2. Compute $u^k \in \operatorname{Arg\,min}_{x \in D} \langle \nabla f(x^k), x \rangle$.

Step 3. If $\langle \nabla f(x^k), u^k - x^k \rangle = 0$, terminate.

Step 4. Choose $\alpha_k \in (0,1]$ using backtracking to satisfy

$$f(x^k + \alpha_k(u^k - x^k)) \le f(x^k) + c\alpha_k \langle \nabla f(x^k), u^k - x^k \rangle.$$

Step 5. Set $x^{k+1} = x^k + \alpha_k(u^k - x^k)$, $k \leftarrow k+1$. Go to Step 2.

Remarks:

- The algorithm either terminates finitely at a stationary point $x^{\bar{k}}$, or every accumulation point of $\{x^k\}$ is stationary.
- When f is convex with Lipschitz gradient (modulus L_f), one can choose in Step 4 (Dunn, Harshbarger '78, Levitin, Polyak '66)

$$\alpha_k = \frac{2}{k+2} \quad \text{or} \quad \alpha_k = \min\left\{1, -\frac{\langle \nabla f(x^k), u^k - x^k \rangle}{L_f \|u^k - x^k\|^2}\right\}.$$

Extending FW?

Frank-Wolfe method for convex D (recapped):

Step 1. Choose $x^0 \in D$ and set $k = 0$.

Step 2. Compute $u^k \in \operatorname{Arg\,min}_{x \in D} \langle \nabla f(x^k), x \rangle$.

Step 3. If $\langle \nabla f(x^k), u^k - x^k \rangle = 0$, terminate.

Step 4. Choose $\alpha_k \in (0, 1]$ to satisfy Armijo rule via backtracking.

Step 5. Set $x^{k+1} = x^k + \alpha_k (u^k - x^k)$, $k \leftarrow k + 1$. Go to Step 2.

Extending FW?

Frank-Wolfe method for convex D (recapped):

Step 1. Choose $x^0 \in D$ and set $k = 0$.

Step 2. Compute $u^k \in \operatorname{Arg\,min}_{x \in D} \langle \nabla f(x^k), x \rangle$.

Step 3. If $\langle \nabla f(x^k), u^k - x^k \rangle = 0$, terminate.

Step 4. Choose $\alpha_k \in (0, 1]$ to satisfy Armijo rule via backtracking.

Step 5. Set $x^{k+1} = x^k + \alpha_k (u^k - x^k)$, $k \leftarrow k + 1$. Go to Step 2.

What if D is nonconvex?

Extending FW?

Frank-Wolfe method for convex D (recapped):

Step 1. Choose $x^0 \in D$ and set $k = 0$.

Step 2. Compute $u^k \in \operatorname{Arg\,min}_{x \in D} \langle \nabla f(x^k), x \rangle$.

Step 3. If $\langle \nabla f(x^k), u^k - x^k \rangle = 0$, terminate.

Step 4. Choose $\alpha_k \in (0, 1]$ to satisfy Armijo rule via backtracking.

Step 5. Set $x^{k+1} = x^k + \alpha_k (u^k - x^k)$, $k \leftarrow k + 1$. Go to Step 2.

What if D is nonconvex?

- Is $u^k \in \operatorname{Arg\,min}_{x \in D} \langle \nabla f(x^k), x \rangle$ in Step 2 easy to solve? (Oracle issue)

Extending FW?

Frank-Wolfe method for convex D (recapped):

Step 1. Choose $x^0 \in D$ and set $k = 0$.

Step 2. Compute $u^k \in \operatorname{Arg\,min}_{x \in D} \langle \nabla f(x^k), x \rangle$.

Step 3. If $\langle \nabla f(x^k), u^k - x^k \rangle = 0$, terminate.

Step 4. Choose $\alpha_k \in (0, 1]$ to satisfy Armijo rule via backtracking.

Step 5. Set $x^{k+1} = x^k + \alpha_k (u^k - x^k)$, $k \leftarrow k + 1$. Go to Step 2.

What if D is nonconvex?

- Is $u^k \in \operatorname{Arg\,min}_{x \in D} \langle \nabla f(x^k), x \rangle$ in Step 2 easy to solve? (Oracle issue)
- Is the termination in Step 3 correct? (Termination issue)

Extending FW?

Frank-Wolfe method for convex D (recapped):

Step 1. Choose $x^0 \in D$ and set $k = 0$.

Step 2. Compute $u^k \in \operatorname{Arg\,min}_{x \in D} \langle \nabla f(x^k), x \rangle$.

Step 3. If $\langle \nabla f(x^k), u^k - x^k \rangle = 0$, terminate.

Step 4. Choose $\alpha_k \in (0, 1]$ to satisfy Armijo rule via backtracking.

Step 5. Set $x^{k+1} = x^k + \alpha_k(u^k - x^k)$, $k \leftarrow k + 1$. Go to Step 2.

What if D is nonconvex?

- Is $u^k \in \operatorname{Arg\,min}_{x \in D} \langle \nabla f(x^k), x \rangle$ in Step 2 easy to solve? (Oracle issue)
- Is the termination in Step 3 correct? (Termination issue)
- The convex combination in Step 5 can make $x^{k+1} \notin D$! (Feas. issue)

Existing work

D as subset of sphere: (Luss, Teboulle '13, Balashov, Polyak, Tremba '20)

- Arises naturally from sparse PCA.
- Assumes concavity of f, so that

$$f(x^k + (u^k - x^k)) \leq f(x^k) + \langle \nabla f(x^k), u^k - x^k \rangle.$$

i.e., $\alpha_k \equiv 1$, which means $x^{k+1} = u^k \in D$.

Existing work

D as subset of sphere: (Luss, Teboulle '13, Balashov, Polyak, Tremba '20)

- Arises naturally from sparse PCA.
- Assumes concavity of f, so that

$$f(x^k + (u^k - x^k)) \leq f(x^k) + \langle \nabla f(x^k), u^k - x^k \rangle.$$

  i.e., $\alpha_k \equiv 1$, which means $x^{k+1} = u^k \in D$.
- Note that ∇f being Lipschitz with modulus L_f implies concavity of $x \mapsto f(x) - L_f\|x\|^2$.

Existing work

D as subset of sphere: (Luss, Teboulle '13, Balashov, Polyak, Tremba '20)

- Arises naturally from sparse PCA.
- Assumes concavity of f, so that

$$f(x^k + (u^k - x^k)) \leq f(x^k) + \langle \nabla f(x^k), u^k - x^k \rangle.$$

  i.e., $\alpha_k \equiv 1$, which means $x^{k+1} = u^k \in D$.
- Note that ∇f being Lipschitz with modulus L_f implies concavity of $x \mapsto f(x) - L_f\|x\|^2$.
  $\therefore$ Concavity can be assumed WLOG on spheres, but can be restrictive for other D.

Existing work

D as subset of sphere: (Luss, Teboulle '13, Balashov, Polyak, Tremba '20)

- Arises naturally from sparse PCA.
- Assumes concavity of f, so that

$$f(x^k + (u^k - x^k)) \leq f(x^k) + \langle \nabla f(x^k), u^k - x^k \rangle.$$

  i.e., $\alpha_k \equiv 1$, which means $x^{k+1} = u^k \in D$.
- Note that ∇f being Lipschitz with modulus L_f implies concavity of $x \mapsto f(x) - L_f\|x\|^2$.
  $\therefore$ Concavity can be assumed WLOG on spheres, but can be restrictive for other D.

Our approach:

- Restrict to a different class of nonconvex D.
- Construct new linear oracles.
- Study optimality conditions.

Generalized LO

Consider compact sets of the form

$$D := \{x \in \mathbb{X} : \; P_1(x) - P_2(x) \leq \sigma\},$$

where $P_1 : \mathbb{X} \to \mathbb{R}$ and $P_2 : \mathbb{X} \to \mathbb{R}$ are convex, $\sigma > 0$.

Generalized LO

Consider compact sets of the form

$$D := \{x \in \mathbb{X} : \ P_1(x) - P_2(x) \le \sigma\},$$

where $P_1 : \mathbb{X} \to \mathbb{R}$ and $P_2 : \mathbb{X} \to \mathbb{R}$ are convex, $\sigma > 0$.

Definition: For P_1, P_2 and σ as above, $y \in D$ and $\xi \in \partial P_2(y)$, define

$$D(y,\xi) := \{x \in \mathbb{X} : \ P_1(x) - P_2(y) - \langle \xi, x - y\rangle \le \sigma\}.$$

For any $v \in \mathbb{X}$, a linear-optimization oracle for (v, y, ξ) (denoted by $\mathcal{LO}(v, y, \xi)$) computes a solution of

$$\min_{x \in \mathbb{X}} \langle v, x\rangle \ \text{ subject to } \ x \in D(y,\xi).$$

Generalized LO

Consider compact sets of the form

$$D := \{x \in \mathbb{X} :\ P_1(x) - P_2(x) \le \sigma\},$$

where $P_1 : \mathbb{X} \to \mathbb{R}$ and $P_2 : \mathbb{X} \to \mathbb{R}$ are convex, $\sigma > 0$.

Definition: For P_1, P_2 and σ as above, $y \in D$ and $\xi \in \partial P_2(y)$, define

$$D(y, \xi) := \{x \in \mathbb{X} :\ P_1(x) - P_2(y) - \langle \xi, x - y \rangle \le \sigma\}.$$

For any $v \in \mathbb{X}$, a linear-optimization oracle for (v, y, ξ) (denoted by $\mathcal{LO}(v, y, \xi)$) computes a solution of

$$\min_{x \in \mathbb{X}} \langle v, x \rangle \ \text{ subject to } \ x \in D(y, \xi).$$

Remarks:

- It holds that $y \in D(y, \xi) \subseteq D$. Thus, $\mathcal{LO}(v, y, \xi)$ is well-defined.
- For any output u of $\mathcal{LO}(v, y, \xi)$ and any $\alpha \in (0, 1)$, we have

$$\alpha y + (1 - \alpha) u \in D(y, \xi)$$

Generalized LO: Example

Matrix completion: Let $\mathbb{X} = \mathbb{R}^{m\times n}$, $P_1(x) := \|x\|_*$, $P_2(x) := \mu\|x\|_F$ for some $\mu \in (0,1)$ and $\sigma > 0$ so that $D := \{x : \ \|x\|_* - \mu\|x\|_F \leq \sigma\}$.

Now, for any $v \in \mathbb{R}^{m\times n}$, $y \in D$ and $\xi \in \partial P_2(y)$, the $\mathcal{LO}(v, y, \xi)$ solves

$$\min_{x\in\mathbb{R}^{m\times n}} \ \langle v, x\rangle \ \text{ subject to } \ \|x\|_* - \langle \xi, x\rangle \leq \sigma,$$

where $\|\xi\|_F \leq \mu < 1$.

Generalized LO: Example

Matrix completion: Let $\mathbb{X} = \mathbb{R}^{m\times n}$, $P_1(x) := \|x\|_*$, $P_2(x) := \mu\|x\|_F$ for some $\mu \in (0,1)$ and $\sigma > 0$ so that $D := \{x : \|x\|_* - \mu\|x\|_F \le \sigma\}$. Now, for any $v \in \mathbb{R}^{m\times n}$, $y \in D$ and $\xi \in \partial P_2(y)$, the $\mathcal{LO}(v, y, \xi)$ solves

$$\min_{x\in\mathbb{R}^{m\times n}} \langle v, x\rangle \ \text{ subject to } \ \|x\|_* - \langle \xi, x\rangle \le \sigma,$$

where $\|\xi\|_F \le \mu < 1$.

Theorem 1. (Zeng, Zhang, Li, P. '21)
Suppose that $v \neq 0$. Let $z = \begin{bmatrix} z_1^T & z_2^T \end{bmatrix}^T$ with $z_1 \in \mathbb{R}^m$ and $z_2 \in \mathbb{R}^n$ be a generalized eigenvector of the smallest generalized eigenvalue of the matrix pencil $(\widetilde{v}, I - \widetilde{\xi})$, and satisfy $z^T(I - \widetilde{\xi})z = 1$, where

$$\widetilde{v} = \begin{bmatrix} 0 & v \\ v^T & 0 \end{bmatrix} \quad \text{and} \quad \widetilde{\xi} = \begin{bmatrix} 0 & \xi \\ \xi^T & 0 \end{bmatrix}.$$

Then $u^* = 2\sigma z_1 z_2^T$ is an output of $\mathcal{LO}(v, y, \xi)$.
Remark: Since $I - \widetilde{\xi} \succ 0$, the above z can be computed using eigifp.

CQ & Optimality conditions

Consider

$$\min_{x \in \mathbb{X}} \; f(x) \;\; \text{subject to} \;\; D := \{x \in \mathbb{X} : \; P_1(x) - P_2(x) \leq \sigma\}, \qquad (\spadesuit)$$

where

- D is compact, $P_1, \; P_2 : \mathbb{X} \to \mathbb{R}$ are convex, $\sigma > 0$; and

CQ & Optimality conditions

Consider

$$\min_{x \in \mathbb{X}} \ f(x) \ \text{ subject to } \ D := \{x \in \mathbb{X} : \ P_1(x) - P_2(x) \leq \sigma\}, \qquad (\spadesuit)$$

where

- D is compact, $P_1, \ P_2 : \mathbb{X} \to \mathbb{R}$ are convex, $\sigma > 0$; and
- the generalized Slater's condition holds: For any $y \in D$ and $\xi \in \partial P_2(y)$, there exists $\hat{x} \in \mathbb{X}$ such that

$$P_1(\hat{x}) - P_2(y) - \langle \xi, \hat{x} - y \rangle < \sigma.$$

CQ & Optimality conditions

Consider

$$\min_{x\in\mathbb{X}} \; f(x) \;\; \text{subject to} \;\; D := \{x \in \mathbb{X} : \; P_1(x) - P_2(x) \leq \sigma\}, \qquad (\spadesuit)$$

where

- D is compact, $P_1,\; P_2 : \mathbb{X} \to \mathbb{R}$ are convex, $\sigma > 0$; and
- the generalized Slater's condition holds: For any $y \in D$ and $\xi \in \partial P_2(y)$, there exists $\hat{x} \in \mathbb{X}$ such that
$$P_1(\hat{x}) - P_2(y) - \langle \xi, \hat{x} - y \rangle < \sigma.$$

Note: The generalized Slater's condition holds for the D in the matrix completion problem.

CQ & Optimality conditions

Consider

$$\min_{x\in\mathbb{X}} \; f(x) \;\text{ subject to }\; D := \{x \in \mathbb{X} : \; P_1(x) - P_2(x) \le \sigma\}, \qquad (\spadesuit)$$

where

- D is compact, $P_1, \; P_2 : \mathbb{X} \to \mathbb{R}$ are convex, $\sigma > 0$; and
- the generalized Slater's condition holds: For any $y \in D$ and $\xi \in \partial P_2(y)$, there exists $\hat{x} \in \mathbb{X}$ such that
$$P_1(\hat{x}) - P_2(y) - \langle \xi, \hat{x} - y\rangle < \sigma.$$

Note: The generalized Slater's condition holds for the D in the matrix completion problem.

Theorem 2. (Zeng, Zhang, Li, P. '21)

Assume the generalized Slater's condition. Then TFAE:

- x^* is a stationary point of ($\spadesuit$), i.e., $\exists \lambda \ge 0$ such that
$$0 \in \nabla f(x^*) + \lambda \partial P_1(x^*) - \lambda \partial P_2(x^*).$$
- $\exists\, \xi^* \in \partial P_2(x^*)$ and $u^* \in \operatorname{Arg\,min}_{x\in D(x^*,\xi^*)} \langle \nabla f(x^*), x\rangle$ such that
$$\langle \nabla f(x^*), u^* - x^*\rangle = 0.$$

Nonconvex FW method

FW$_{\text{ncvx}}$: Frank-Wolfe method for (♠)

Step 1. Choose $x^0 \in D$ and set $k = 0$.

Step 2. Pick any $\xi^k \in \partial P_2(x^k)$ and compute

$$u^k \in \operatorname*{Arg\,min}_{x \in D(x^k, \xi^k)} \langle \nabla f(x^k), x \rangle.$$

Step 3. If $\langle \nabla f(x^k), u^k - x^k \rangle = 0$, terminate.

Step 4. Choose $\alpha_k \in (0, 1]$ to satisfy Armijo rule via backtracking.

Step 5. Set $x^{k+1} = x^k + \alpha_k (u^k - x^k)$, $k \leftarrow k + 1$. Go to Step 2.

Nonconvex FW method

FW$_{\text{ncvx}}$: Frank-Wolfe method for (♠)

Step 1. Choose $x^0 \in D$ and set $k = 0$.

Step 2. Pick any $\xi^k \in \partial P_2(x^k)$ and compute

$$u^k \in \operatorname*{Arg\,min}_{x \in D(x^k, \xi^k)} \langle \nabla f(x^k), x \rangle.$$

Step 3. If $\langle \nabla f(x^k), u^k - x^k \rangle = 0$, terminate.

Step 4. Choose $\alpha_k \in (0, 1]$ to satisfy Armijo rule via backtracking.

Step 5. Set $x^{k+1} = x^k + \alpha_k (u^k - x^k)$, $k \leftarrow k + 1$. Go to Step 2.

Theorem 3. (Zeng, Zhang, Li, P. '21)
Assume the generalized Slater's condition. Then:

- Finite termination returns a stationary point $x^{\bar{k}}$.
- Line-search loop in Step 4 terminates finitely.
- $\{x^k\} \subseteq D$ and each accumulation point is stationary.

Away-step oracles

When D is convex:

- Classical method to "accelerate" FW method. (Wolfe '70, GuéLat, Marcotte '86, Lacoste-Julien, Jaggi '15, Beck, Shtern '17, ...)

Away-step oracles

When D is convex:

- Classical method to "accelerate" FW method. (Wolfe '70, GuéLat, Marcotte '86, Lacoste-Julien, Jaggi '15, Beck, Shtern '17, ...)
- Idea:
  - ⋆ Start with a set of "atoms" $\mathcal{A}_0 \subset D$.
  - ⋆ For each iteration, find

$$a^k \in \operatorname*{Arg\,max}_{a \in \mathcal{A}_k} \langle \nabla f(x^k), a \rangle.$$

  - ⋆ Consider the away-step direction $x^k - a^k$.

Away-step oracles

When D is convex:

- Classical method to "accelerate" FW method. (Wolfe '70, GuéLat, Marcotte '86, Lacoste-Julien, Jaggi '15, Beck, Shtern '17, ...)
- Idea:
  - ⋆ Start with a set of "atoms" $\mathcal{A}_0 \subset D$.
  - ⋆ For each iteration, find

$$a^k \in \operatorname*{Arg\,max}_{a \in \mathcal{A}_k} \langle \nabla f(x^k), a \rangle.$$

  - ⋆ Consider the away-step direction $x^k - a^k$.
  - ⋆ Construct $\mathcal{A}_{k+1} \subset D$ based on $\mathcal{A}_k$.

Away-step oracles

When D is convex:

- Classical method to "accelerate" FW method. (Wolfe '70, GuéLat, Marcotte '86, Lacoste-Julien, Jaggi '15, Beck, Shtern '17, ...)
- Idea:
  - ⋆ Start with a set of "atoms" $\mathcal{A}_0 \subset D$.
  - ⋆ For each iteration, find

$$a^k \in \operatorname*{Arg\,max}_{a \in \mathcal{A}_k} \langle \nabla f(x^k), a \rangle.$$

  - ⋆ Consider the away-step direction $x^k - a^k$.
  - ⋆ Construct $\mathcal{A}_{k+1} \subset D$ based on $\mathcal{A}_k$.

When D is nonconvex:

- Construct $\mathcal{A}_k \subset D(x^k, \xi^k)$.

Away-step oracles

When D is convex:

- Classical method to "accelerate" FW method. (Wolfe '70, GuéLat, Marcotte '86, Lacoste-Julien, Jaggi '15, Beck, Shtern '17, ...)
- Idea:
  - ⋆ Start with a set of "atoms" $\mathcal{A}_0 \subset D$.
  - ⋆ For each iteration, find

$$a^k \in \operatorname*{Arg\,max}_{a \in \mathcal{A}_k} \langle \nabla f(x^k), a \rangle.$$

  - ⋆ Consider the away-step direction $x^k - a^k$.
  - ⋆ Construct $\mathcal{A}_{k+1} \subset D$ based on $\mathcal{A}_k$.

When D is nonconvex:

- Construct $\mathcal{A}_k \subset D(x^k, \xi^k)$.
- Previous atoms may not be feasible for $\mathcal{LO}$ as $D(x^k, \xi^k)$ changes from iteration to iteration.

Away-step oracles

When D is convex:

- Classical method to "accelerate" FW method. (Wolfe '70, GuéLat, Marcotte '86, Lacoste-Julien, Jaggi '15, Beck, Shtern '17, ...)
- Idea:
  - ⋆ Start with a set of "atoms" $\mathcal{A}_0 \subset D$.
  - ⋆ For each iteration, find

$$a^k \in \operatorname*{Arg\,max}_{a \in \mathcal{A}_k} \langle \nabla f(x^k), a \rangle.$$

  - ⋆ Consider the away-step direction $x^k - a^k$.
  - ⋆ Construct $\mathcal{A}_{k+1} \subset D$ based on $\mathcal{A}_k$.

When D is nonconvex:

- Construct $\mathcal{A}_k \subset D(x^k, \xi^k)$.
- Previous atoms may not be feasible for $\mathcal{LO}$ as $D(x^k, \xi^k)$ changes from iteration to iteration.
- A primitive approach: construct $\mathcal{A}_k \subset D(x^k, \xi^k)$ solely based on the current iterate x^k.

$\mathrm{FW}_{\mathrm{ncvx}}$ with away-step

$\mathrm{FW}_{\mathrm{ncvx}}$ with away step for (♠):

Step 1. Choose $x^0 \in D$ and set $k = 0$.

Step 2. Pick any $\xi^k \in \partial P_2(x^k)$ and compute $u^k \in \operatorname*{Arg\,min}_{x \in D(x^k, \xi^k)} \langle \nabla f(x^k), x \rangle$.

Step 3. If $\langle \nabla f(x^k), u^k - x^k \rangle = 0$, terminate.

FW_{ncvx} with away-step

FW_{ncvx} with away step for ($\spadesuit$):

Step 1. Choose $x^0 \in D$ and set $k = 0$.

Step 2. Pick any $\xi^k \in \partial P_2(x^k)$ and compute $u^k \in \operatorname*{Arg\,min}_{x \in D(x^k, \xi^k)} \langle \nabla f(x^k), x \rangle$.

Step 3. If $\langle \nabla f(x^k), u^k - x^k \rangle = 0$, terminate.

Step 4. Construct $\mathcal{A}_k \subset D(x^k, \xi^k)$ with $x^k \in \text{conv}(\mathcal{A}_k)$ and set

$$a^k \in \operatorname*{Arg\,max}_{a \in \mathcal{A}_k} \langle \nabla f(x^k), a \rangle.$$

Pick $\alpha_{\text{aw}} \leq \max\{\alpha \geq 0 : \ x^k + \alpha(x^k - a^k) \in D(x^k, \xi^k)\}$

$\mathrm{FW}_{\mathrm{ncvx}}$ with away-step

$\mathrm{FW}_{\mathrm{ncvx}}$ with away step for (♠):

Step 1. Choose $x^0 \in D$ and set $k = 0$. Choose $\epsilon > 0$.

Step 2. Pick any $\xi^k \in \partial P_2(x^k)$ and compute $u^k \in \operatorname*{Arg\,min}_{x \in D(x^k, \xi^k)} \langle \nabla f(x^k), x \rangle$.

Step 3. If $\langle \nabla f(x^k), u^k - x^k \rangle = 0$, terminate.

Step 4. Construct $\mathcal{A}_k \subset D(x^k, \xi^k)$ with $x^k \in \operatorname{conv}(\mathcal{A}_k)$ and set

$$a^k \in \operatorname*{Arg\,max}_{a \in \mathcal{A}_k} \langle \nabla f(x^k), a \rangle.$$

Pick $\alpha_{\mathrm{aw}} \le \max\{\alpha \ge 0 : \; x^k + \alpha(x^k - a^k) \in D(x^k, \xi^k)\}$

Step 5. If $\alpha_{\mathrm{aw}} < \epsilon$, set $d^k = u^k - x^k$; else, choose d^k among $u^k - x^k$ and $x^k - a^k$ for a more negative $\langle \nabla f(x^k), d^k \rangle$.

Step 6. If $d^k = u^k - x^k$, set $\alpha_{\mathrm{init}} = 1$; else, set $\alpha_{\mathrm{init}} = \alpha_{\mathrm{aw}}$.

FW_{ncvx} with away-step

FW_{ncvx} with away step for (♠):

Step 1. Choose $x^0 \in D$ and set $k = 0$. Choose $\epsilon > 0$.

Step 2. Pick any $\xi^k \in \partial P_2(x^k)$ and compute $u^k \in \operatorname*{Arg\,min}_{x \in D(x^k, \xi^k)} \langle \nabla f(x^k), x \rangle$.

Step 3. If $\langle \nabla f(x^k), u^k - x^k \rangle = 0$, terminate.

Step 4. Construct $\mathcal{A}_k \subset D(x^k, \xi^k)$ with $x^k \in \text{conv}(\mathcal{A}_k)$ and set

$$a^k \in \operatorname*{Arg\,max}_{a \in \mathcal{A}_k} \langle \nabla f(x^k), a \rangle.$$

Pick $\alpha_{\text{aw}} \le \max\{\alpha \ge 0 : x^k + \alpha(x^k - a^k) \in D(x^k, \xi^k)\}$

Step 5. If $\alpha_{\text{aw}} < \epsilon$, set $d^k = u^k - x^k$; else, choose d^k among $u^k - x^k$ and $x^k - a^k$ for a more negative $\langle \nabla f(x^k), d^k \rangle$.

Step 6. If $d^k = u^k - x^k$, set $\alpha_{\text{init}} = 1$; else, set $\alpha_{\text{init}} = \alpha_{\text{aw}}$.

Step 7. Choose $\alpha_k \in (0, \alpha_{\text{init}}]$ to satisfy Armijo rule via backtracking.

Step 8. Set $x^{k+1} = x^k + \alpha_k d^k$, $k \leftarrow k + 1$. Go to Step 2.

FW_{ncvx} with away-step

FW_{ncvx} with away step for (♠):

Step 1. Choose $x^0 \in D$ and set $k = 0$. Choose $\epsilon > 0$.

Step 2. Pick any $\xi^k \in \partial P_2(x^k)$ and compute $u^k \in \underset{x \in D(x^k, \xi^k)}{\operatorname{Arg\,min}} \langle \nabla f(x^k), x \rangle$.

Step 3. If $\langle \nabla f(x^k), u^k - x^k \rangle = 0$, terminate.

Step 4. Construct $\mathcal{A}_k \subset D(x^k, \xi^k)$ with $x^k \in \operatorname{conv}(\mathcal{A}_k)$ and set

$$a^k \in \underset{a \in \mathcal{A}_k}{\operatorname{Arg\,max}} \langle \nabla f(x^k), a \rangle.$$

Pick $\alpha_{\text{aw}} \leq \max\{\alpha \geq 0 : \; x^k + \alpha(x^k - a^k) \in D(x^k, \xi^k)\}$

Step 5. If $\alpha_{\text{aw}} < \epsilon$, set $d^k = u^k - x^k$; else, choose d^k among $u^k - x^k$ and $x^k - a^k$ for a more negative $\langle \nabla f(x^k), d^k \rangle$.

Step 6. If $d^k = u^k - x^k$, set $\alpha_{\text{init}} = 1$; else, set $\alpha_{\text{init}} = \alpha_{\text{aw}}$.

Step 7. Choose $\alpha_k \in (0, \alpha_{\text{init}}]$ to satisfy Armijo rule via backtracking.

Step 8. Set $x^{k+1} = x^k + \alpha_k d^k$, $k \leftarrow k + 1$. Go to Step 2.

Same convergence guarantee as $\textbf{FW}_{\text{ncvx}}$ under generalized Slater's condition.

Convergence proof idea

Define a gap function $G : D \to \mathbb{R}$ by

$$G(x) = \inf_{\xi \in \partial P_2(x)} \max_{y \in D(x,\xi)} \langle \nabla f(x), x - y \rangle.$$

Theorem 4. (Zeng, Zhang, Li, P. '21)
Assume the generalized Slater's condition. Then $G(x) \geq 0$ for all $x \in D$. Moreover, if $\{w^k\} \subseteq D$ is such that

$$G(w^k) \to 0 \ \text{ and } \ w^k \to x^*$$

for some x^*, then $x^* \in D$ and is a stationary point of (♠).

Convergence proof idea

Define a gap function $G : D \to \mathbb{R}$ by

$$G(x) = \inf_{\xi \in \partial P_2(x)} \max_{y \in D(x,\xi)} \langle \nabla f(x), x - y \rangle.$$

Theorem 4. (Zeng, Zhang, Li, P. '21)
Assume the generalized Slater's condition. Then $G(x) \geq 0$ for all $x \in D$. Moreover, if $\{w^k\} \subseteq D$ is such that

$$G(w^k) \to 0 \ \text{ and } \ w^k \to x^*$$

for some x^*, then $x^* \in D$ and is a stationary point of (♠).

Convergence of FW_{ncvx}: Let $\{x^k\}$ be generated by FW_{ncvx}.

- Direct computation shows that $0 \leq G(x^k) \leq -\langle \nabla f(x^k), u^k - x^k \rangle$.
- Backtracking + Armijo rule give $\langle \nabla f(x^k), u^k - x^k \rangle \to 0$.
- Convergence follows from these and **Theorem 4**.

Convergence of FW_{ncvx} with away step can be proved similarly.

Numerical experiments

- Matrix completion:

$$\min_{x\in\mathbb{R}^{m\times n}} \sum_{(i,j)\in\Omega} (x_{ij} - \bar{x}_{ij})^2 \;\; \text{subject to} \;\; \|x\|_* - 0.5\|x\|_F \leq \sigma,$$

  where
  - ⋆ Ω collects the indices of observed entries;
  - ⋆ $\overline{x}$ comes from observation, $\sigma > 0$;
  - ⋆ $\|x\|_*$ and $\|x\|_F$ are resp. nuclear and Fröbenius norm.

Numerical experiments

- Matrix completion:

$$\min_{x \in \mathbb{R}^{m\times n}} \sum_{(i,j)\in\Omega} (x_{ij} - \bar{x}_{ij})^2 \ \text{ subject to } \ \|x\|_* - 0.5\|x\|_F \le \sigma,$$

  where
  - ⋆ Ω collects the indices of observed entries;
  - ⋆ $\bar{x}$ comes from observation, $\sigma > 0$;
  - ⋆ $\|x\|_*$ and $\|x\|_F$ are resp. nuclear and Fröbenius norm.

- Efficient implementation: Following (Freund, Grigas, Mazumder '17)

Numerical experiments

- Matrix completion:

$$\min_{x\in\mathbb{R}^{m\times n}} \sum_{(i,j)\in\Omega} (x_{ij}-\bar{x}_{ij})^2 \text{ subject to } \|x\|_* - 0.5\|x\|_F \le \sigma,$$

  where

  - ⋆ Ω collects the indices of observed entries;
  - ⋆ $\bar{x}$ comes from observation, $\sigma > 0$;
  - ⋆ $\|x\|_*$ and $\|x\|_F$ are resp. nuclear and Fröbenius norm.

- Efficient implementation: Following (Freund, Grigas, Mazumder '17)
  - ⋆ Maintain (R^k, Σ^k, T^k) (reduced SVD of x^k), never form x^k.

Numerical experiments

- Matrix completion:

$$\min_{x \in \mathbb{R}^{m \times n}} \sum_{(i,j) \in \Omega} (x_{ij} - \bar{x}_{ij})^2 \text{ subject to } \|x\|_* - 0.5\|x\|_F \leq \sigma,$$

  where

  - ⋆ Ω collects the indices of observed entries;
  - ⋆ $\bar{x}$ comes from observation, $\sigma > 0$;
  - ⋆ $\|x\|_*$ and $\|x\|_F$ are resp. nuclear and Fröbenius norm.

- Efficient implementation: Following (Freund, Grigas, Mazumder '17)
  - ⋆ Maintain (R^k, Σ^k, T^k) (reduced SVD of x^k), never form x^k.
  - ⋆ Compute x_{ij}^k for $(i, j) \in \Omega$ only to obtain the gradient.

Numerical experiments

- Matrix completion:

$$\min_{x\in\mathbb{R}^{m\times n}} \sum_{(i,j)\in\Omega} (x_{ij} - \bar{x}_{ij})^2 \ \text{ subject to } \ \|x\|_* - 0.5\|x\|_F \le \sigma,$$

  where

  - ⋆ Ω collects the indices of observed entries;
  - ⋆ $\overline{x}$ comes from observation, $\sigma > 0$;
  - ⋆ $\|x\|_*$ and $\|x\|_F$ are resp. nuclear and Fröbenius norm.

- Efficient implementation: Following (Freund, Grigas, Mazumder '17)
  - ⋆ Maintain (R^k, Σ^k, T^k) (reduced SVD of x^k), never form x^k.
  - ⋆ Compute x^k_{ij} for $(i,j) \in \Omega$ only to obtain the gradient.
  - ⋆ Compute u^k using eigifp, which has rank ONE.

Numerical experiments

- Matrix completion:

$$\min_{x \in \mathbb{R}^{m \times n}} \sum_{(i,j)\in\Omega} (x_{ij} - \bar{x}_{ij})^2 \text{ subject to } \|x\|_* - 0.5\|x\|_F \le \sigma,$$

  where

  ⋆ Ω collects the indices of observed entries;
  ⋆ $\bar{x}$ comes from observation, $\sigma > 0$;
  ⋆ $\|x\|_*$ and $\|x\|_F$ are resp. nuclear and Fröbenius norm.

- Efficient implementation: Following (Freund, Grigas, Mazumder '17)

  ⋆ Maintain (R^k, Σ^k, T^k) (reduced SVD of x^k), never form x^k.
  ⋆ Compute x_{ij}^k for $(i,j) \in \Omega$ only to obtain the gradient.
  ⋆ Compute u^k using eigifp, which has rank ONE.
  ⋆ **KEY**: Since

$$x^{k+1} = (1 - \alpha_k)x^k + \alpha_k u^k,$$

  one can obtain $(R^{k+1}, \Sigma^{k+1}, T^{k+1})$ using SVD rank-one update.

Numerical experiments

- Matrix completion:

$$\min_{x \in \mathbb{R}^{m \times n}} \sum_{(i,j) \in \Omega} (x_{ij} - \bar{x}_{ij})^2 \ \text{ subject to } \ \|x\|_* - 0.5\|x\|_F \le \sigma,$$

  where
  - ⋆ Ω collects the indices of observed entries;
  - ⋆ $\overline{x}$ comes from observation, $\sigma > 0$;
  - ⋆ $\|x\|_*$ and $\|x\|_F$ are resp. nuclear and Fröbenius norm.

- Efficient implementation: Following (Freund, Grigas, Mazumder '17)
  - ⋆ Maintain (R^k, Σ^k, T^k) (reduced SVD of x^k), never form x^k.
  - ⋆ Compute x_{ij}^k for $(i,j) \in \Omega$ only to obtain the gradient.
  - ⋆ Compute u^k using eigifp, which has rank ONE.
  - ⋆ **KEY**: Since

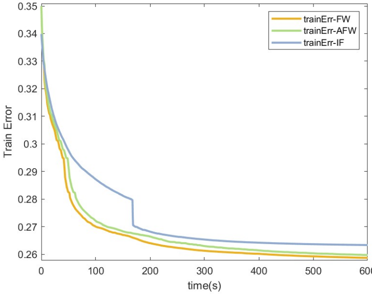
$$x^{k+1} = (1 - \alpha_k)x^k + \alpha_k u^k,$$

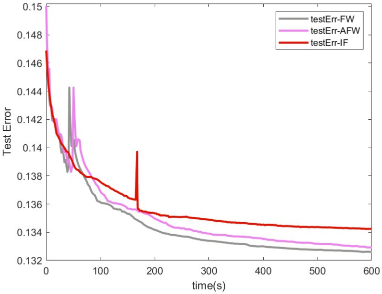
    one can obtain $(R^{k+1}, \Sigma^{k+1}, T^{k+1})$ using SVD rank-one update.

  In contrast, GP will need to form x^k, and perform full SVD (for projection).

Numerical experiments cont.

- MovieLens10M: $n = 10677$ movie ratings from $m = 69878$ users.
- Randomly choose 70% as training dataset (i.e.,Ω). Training and testing errors as the algorithm progresses are shown below.
- For simplicity, we used the same Ω and the same σ (determined via CV on nuc. norm model) as in (Freund, Grigas, Mazumder '17).





Matlab 2017b on a 64-bit PC with an Intel(R) Core(TM) i5-7200 CPU (2.50GHz) and 8GB of RAM

Numerical experiments cont.

Table: Relative optimality measure ($\widehat{\varepsilon}$), Rank and RMSE for **IF**, $\mathbf{FW_{ncvx}}$ and $\mathbf{AFW_{ncvx}}$ within different maximal computational time $T^{\max}$

	MovieLens10M Dataset								
	IF			$\mathbf{FW_{ncvx}}$			$\mathbf{AFW_{ncvx}}$		
$T^{\max}(s)$	$\widehat{\varepsilon}$	rank	RMSE	$\widehat{\varepsilon}$	rank	RMSE	$\widehat{\varepsilon}$	rank	RMSE
1000	8.0e-03	135	0.8086	5.8e-03	218	0.8036	8.4e-03	99	0.8044
1500	5.1e-03	144	0.8084	4.3e-03	274	0.8031	5.5e-03	114	0.8035
2000	3.9e-03	145	0.8082	3.8e-03	322	0.8029	8.4e-03	120	0.8032
2500	3.1e-03	147	0.8081	3.9e-03	365	0.8027	2.4e-03	129	0.8030
3000	2.8e-03	147	0.8081	2.5e-03	401	0.8028	2.1e-03	132	0.8029

Note:

$$\widehat{\varepsilon} := \frac{|\langle \nabla f(x^k), d^k|}{\max\{|f(x^k) + \langle \nabla f(x^k), d^k\rangle|, 1\}}, \quad \text{RMSE} := \sqrt{\frac{1}{mn}\sum_{(i,j)\in\Omega}(x_{ij}^k - x_{ij}^{\text{true}})^2}$$

Conclusion

Conclusion:

- Extended FW method for special nonconvex sets: Level set of DC functions satisfying some regularity conditions.
- Introduced generalized LO: Efficient implementation for applications such as matrix completion.
- Established subsequential convergence.

Reference:

- L. Zeng, Y. Zhang, G. Li and T. K. Pong.
  Frank-Wolfe-type methods for nonconvex inequality-constrained problems.
  Preprint. Available at https://arxiv.org/abs/2112.14404.

Thanks for coming! :)