Frank-Wolfe Methods for Minimizing Log-Homogeneous Self-Concordant Barriers

Renbo Zhao

Tippie College of Business
University of Iowa

Joint work with Robert M. Freund (MIT Sloan)

National University of Singapore
May, 2024

1. Introduction
2. Applications
3. Our Method: New Generalized Frank-Wolfe Method
4. Computational Experiments

Problem of Interest

$$F^* := \min_{x\in\mathbb{R}^n} \left[F(x) := f(\mathsf{A}x) + h(x)\right] \qquad \texttt{(P)}$$

Problem of Interest

$$F^* := \min_{x\in\mathbb{R}^n} \left[F(x) := f(\mathsf{A}x) + h(x)\right] \quad \texttt{(P)}$$

- $\mathsf{A} : \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator (not necessarily invertible)

Problem of Interest

$$F^* := \min_{x\in\mathbb{R}^n} \; [F(x) := f(\mathsf{A}x) + h(x)] \tag{P}$$

- ▷ $\mathsf{A} : \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator (not necessarily invertible)
- ▷ $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is a θ-log-homogeneous self-concordant barrier (θ-LHSCB) for some regular cone $\mathcal{K} \subseteq \mathbb{R}^m$

Problem of Interest

$$F^* := \min_{x\in\mathbb{R}^n} \left[F(x) := f(\mathsf{A}x) + h(x)\right] \quad \texttt{(P)}$$

- $\mathsf{A} : \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator (not necessarily invertible)
- $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is a θ-log-homogeneous self-concordant barrier (θ-LHSCB) for some regular cone $\mathcal{K} \subseteq \mathbb{R}^m$
- $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a closed and convex function, with compact domain $\mathcal{X} := \mathsf{dom}\, h$

Problem of Interest

$$F^* := \min_{x\in\mathbb{R}^n} \left[F(x) := f(\mathsf{A}x) + h(x)\right] \quad \text{(P)}$$

- $\mathsf{A} : \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator (not necessarily invertible)
- $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is a θ-log-homogeneous self-concordant barrier (θ-LHSCB) for some regular cone $\mathcal{K} \subseteq \mathbb{R}^m$
- $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a closed and convex function, with compact domain $\mathcal{X} := \mathsf{dom}\, h$
- Assume $\mathsf{dom}\, F \neq \emptyset$, so at least one minimizer $x^* \in \mathsf{dom}\, F$ exists, and define $F^* := F(x^*)$

Problem of Interest

$$F^* := \min_{x \in \mathbb{R}^n} \left[F(x) := f(\mathsf{A}x) + h(x) \right] \qquad \texttt{(P)}$$

- $\mathsf{A} : \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator (not necessarily invertible)
- $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is a θ-log-homogeneous self-concordant barrier (θ-LHSCB) for some regular cone $\mathcal{K} \subseteq \mathbb{R}^m$
- $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a closed and convex function, with compact domain $\mathcal{X} := \mathsf{dom}\, h$
- Assume $\mathsf{dom}\, F \neq \emptyset$, so at least one minimizer $x^* \in \mathsf{dom}\, F$ exists, and define $F^* := F(x^*)$
- Includes many applications (coming up later).

θ-**LHSCB** (logarithmically-homogeneous self-concordant barrier)

θ-**LHSCB** (logarithmically-homogeneous self-concordant barrier)

- Let $\mathcal{K} \subsetneq \mathbb{R}^m$ be a regular cone, i.e., $\mathcal{K}$ is closed, convex, pointed and has nonempty interior.

θ-**LHSCB** (logarithmically-homogeneous self-concordant barrier)

- Let $\mathcal{K} \subsetneqq \mathbb{R}^m$ be a regular cone, i.e., $\mathcal{K}$ is closed, convex, pointed and has nonempty interior.
- f is a θ-LHSCB on $\mathcal{K}$ with *complexity parameter* $\theta \geq 1$ if f is three-times differentiable and strictly convex on $\mathsf{int}\,\mathcal{K}$, and satisfies

θ-LHSCB (logarithmically-homogeneous self-concordant barrier)

- Let $\mathcal{K} \subsetneq \mathbb{R}^m$ be a regular cone, i.e., $\mathcal{K}$ is closed, convex, pointed and has nonempty interior.
- f is a θ-LHSCB on $\mathcal{K}$ with *complexity parameter* $\theta \geq 1$ if f is three-times differentiable and strictly convex on $\mathsf{int}\,\mathcal{K}$, and satisfies

  1. $\left|D^3 f(u)[w,w,w]\right| \leq 2\|w\|_u^3 \;\; \forall\, u \in \mathsf{int}\,\mathcal{K}, \forall\, w \in \mathbb{R}^m,$
  2. $f(u_k) \to \infty$ for any $\{u_k\}_{k\geq 1} \subseteq \mathsf{int}\,\mathcal{K}$ such that $u_k \to u \in \mathsf{bd}\,\mathcal{K}$,
  3. $f(tu) = f(u) - \theta \ln(t) \;\; \forall\, u \in \mathsf{int}\,\mathcal{K}, \forall\, t > 0.$

  where $\|w\|_u := \langle \nabla^2 f(u) w, w\rangle^{1/2}$ denotes the local norm of w at $u \in \mathsf{int}\,\mathcal{K}$.

θ-**LHSCB** (logarithmically-homogeneous self-concordant barrier)

- Let $\mathcal{K} \subsetneq \mathbb{R}^m$ be a regular cone, i.e., $\mathcal{K}$ is closed, convex, pointed and has nonempty interior.
- f is a θ-LHSCB on $\mathcal{K}$ with *complexity parameter* $\theta \geq 1$ if f is three-times differentiable and strictly convex on $\mathsf{int}\,\mathcal{K}$, and satisfies
  1. $\left| D^3 f(u)[w,w,w] \right| \leq 2\|w\|_u^3 \;\; \forall\, u \in \mathsf{int}\,\mathcal{K},\, \forall\, w \in \mathbb{R}^m$,
  2. $f(u_k) \to \infty$ for any $\{u_k\}_{k\geq 1} \subseteq \mathsf{int}\,\mathcal{K}$ such that $u_k \to u \in \mathsf{bd}\,\mathcal{K}$,
  3. $f(tu) = f(u) - \theta \ln(t) \;\; \forall\, u \in \mathsf{int}\,\mathcal{K},\, \forall\, t > 0$.

  where $\|w\|_u := \langle \nabla^2 f(u) w, w \rangle^{1/2}$ denotes the local norm of w at $u \in \mathsf{int}\,\mathcal{K}$.
- Two prototypical examples:

θ-**LHSCB** (logarithmically-homogeneous self-concordant barrier)

- Let $\mathcal{K} \subsetneq \mathbb{R}^m$ be a regular cone, i.e., $\mathcal{K}$ is closed, convex, pointed and has nonempty interior.
- f is a θ-LHSCB on $\mathcal{K}$ with *complexity parameter* $\theta \geq 1$ if f is three-times differentiable and strictly convex on $\mathsf{int}\,\mathcal{K}$, and satisfies
  1. $\left|D^3 f(u)[w,w,w]\right| \leq 2\|w\|_u^3 \;\; \forall\, u \in \mathsf{int}\,\mathcal{K}, \forall\, w \in \mathbb{R}^m$,
  2. $f(u_k) \to \infty$ for any $\{u_k\}_{k\geq 1} \subseteq \mathsf{int}\,\mathcal{K}$ such that $u_k \to u \in \mathsf{bd}\,\mathcal{K}$,
  3. $f(tu) = f(u) - \theta \ln(t) \;\; \forall\, u \in \mathsf{int}\,\mathcal{K}, \forall\, t > 0$.

  where $\|w\|_u := \langle \nabla^2 f(u) w, w\rangle^{1/2}$ denotes the local norm of w at $u \in \mathsf{int}\,\mathcal{K}$.
- Two prototypical examples:
  - $f(U) = -\ln\det(U)$ for $U \in \mathcal{K} := \mathbb{S}_+^k$ and $\theta = k$,

θ-LHSCB (logarithmically-homogeneous self-concordant barrier)

- Let $\mathcal{K} \subsetneq \mathbb{R}^m$ be a regular cone, i.e., $\mathcal{K}$ is closed, convex, pointed and has nonempty interior.
- f is a θ-LHSCB on $\mathcal{K}$ with *complexity parameter* $\theta \geq 1$ if f is three-times differentiable and strictly convex on $\mathsf{int}\,\mathcal{K}$, and satisfies
  1. $\left|D^3 f(u)[w,w,w]\right| \leq 2\|w\|_u^3 \;\; \forall\, u \in \mathsf{int}\,\mathcal{K}, \forall\, w \in \mathbb{R}^m$,
  2. $f(u_k) \to \infty$ for any $\{u_k\}_{k\geq 1} \subseteq \mathsf{int}\,\mathcal{K}$ such that $u_k \to u \in \mathsf{bd}\,\mathcal{K}$,
  3. $f(tu) = f(u) - \theta \ln(t) \;\; \forall\, u \in \mathsf{int}\,\mathcal{K}, \forall\, t > 0$.

  where $\|w\|_u := \langle \nabla^2 f(u) w, w\rangle^{1/2}$ denotes the local norm of w at $u \in \mathsf{int}\,\mathcal{K}$.
- Two prototypical examples:
  - $f(U) = -\ln\det(U)$ for $U \in \mathcal{K} := \mathbb{S}_+^k$ and $\theta = k$,
  - $f(u) = -\sum_{j=1}^m w_j \ln(u_j)$ for $u \in \mathcal{K} := \mathbb{R}_+^m$ and $\theta = \sum_{j=1}^m w_j$ where $w_1, \ldots, w_n \geq 1$.

A Motivating Example: D-optimal Design

$$\begin{array}{ll} \min_p & -\ln\det(\sum_{i=1}^m p_i a_i a_i^\top) \\ \text{s.t.} & \sum_{i=1}^m p_i = 1,\ p_i \geq 0,\ \forall\, i \in [m]. \end{array} \quad \texttt{(D-OPT)}$$

A Motivating Example: D-optimal Design

$$
\begin{aligned}
\min_p \quad & -\ln\det\left(\textstyle\sum_{i=1}^m p_i a_i a_i^\top\right) \\
\text{s.t.} \quad & \textstyle\sum_{i=1}^m p_i = 1,\ p_i \ge 0,\ \forall\, i \in [m].
\end{aligned}
\tag{D-OPT}
$$

▷ Problem data: $\{a_i\}_{i=1}^m \subseteq \mathbb{R}^n$ whose linear span is $\mathbb{R}^n$.

A Motivating Example: D-optimal Design

$$\begin{aligned} \min_p \quad & -\ln\det(\textstyle\sum_{i=1}^m p_i a_i a_i^\top) \\ \text{s.t.} \quad & \textstyle\sum_{i=1}^m p_i = 1,\ p_i \geq 0,\ \forall\, i \in [m]. \end{aligned} \qquad \texttt{(D-OPT)}$$

- Problem data: $\{a_i\}_{i=1}^m \subseteq \mathbb{R}^n$ whose linear span is $\mathbb{R}^n$.
- Arises in many places, including optimal experimental design, and as the dual problem of the minimum volume enclosing ellipsoid problem.

A Motivating Example: D-optimal Design

$$\begin{aligned} \min_p \quad & -\ln\det\left(\textstyle\sum_{i=1}^m p_i a_i a_i^\top\right) \\ \text{s.t.} \quad & \textstyle\sum_{i=1}^m p_i = 1,\ p_i \geq 0,\ \forall\, i \in [m]. \end{aligned} \tag{D-OPT}$$

- Problem data: $\{a_i\}_{i=1}^m \subseteq \mathbb{R}^n$ whose linear span is $\mathbb{R}^n$.
- Arises in many places, including optimal experimental design, and as the dual problem of the minimum volume enclosing ellipsoid problem.
- Khachiyan (1996) proposed a "barycentric coordinate ascent" method with exact line-search, which is actually FW with exact line-search. Method works remarkably well both in theory and practice: it computes an ε-optimal solution of (D-OPT) in (essentially) $O(n^2/\varepsilon)$ iterations.

A Motivating Example: D-optimal Design

$$\begin{aligned} \min_p \quad & -\ln\det\left(\sum_{i=1}^m p_i a_i a_i^\top\right) \\ \text{s.t.} \quad & \sum_{i=1}^m p_i = 1,\ p_i \ge 0,\ \forall\, i \in [m]. \end{aligned} \tag{D-OPT}$$

A Motivating Example: D-optimal Design

$$
\begin{aligned}
\min_p \quad & -\ln\det\left(\textstyle\sum_{i=1}^m p_i a_i a_i^\top\right) \\
\text{s. t.} \quad & \textstyle\sum_{i=1}^m p_i = 1,\ p_i \ge 0,\ \forall\, i \in [m].
\end{aligned}
\tag{D-OPT}
$$

- The success of this method has been a mysterious outlier for more than 20 years, since the objective function in (D-OPT) does not have Lipschitz gradient on the feasible region, which is a critical assumption for traditional Frank-Wolfe.

A Motivating Example: D-optimal Design

$$\begin{aligned} \min_p \quad & -\ln\det\left(\textstyle\sum_{i=1}^m p_i a_i a_i^\top\right) \\ \text{s.t.} \quad & \textstyle\sum_{i=1}^m p_i = 1,\ p_i \ge 0,\ \forall\, i \in [m]. \end{aligned} \qquad \texttt{(D-OPT)}$$

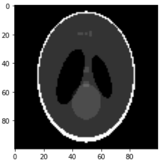
- The success of this method has been a mysterious outlier for more than 20 years, since the objective function in (D-OPT) does not have Lipschitz gradient on the feasible region, which is a critical assumption for traditional Frank-Wolfe.
- What problem structure actually drives the success of Khachiyan's method? And might such structure exist anywhere else?

A Motivating Example: D-optimal Design

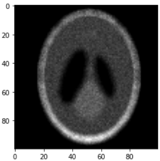
$$\begin{aligned} \min_p \quad & -\ln\det\left(\textstyle\sum_{i=1}^m p_i a_i a_i^\top\right) \\ \text{s. t.} \quad & \textstyle\sum_{i=1}^m p_i = 1,\ p_i \ge 0,\ \forall\, i \in [m]. \end{aligned} \qquad \texttt{(D-OPT)}$$

- The success of this method has been a mysterious outlier for more than 20 years, since the objective function in (D-OPT) does not have Lipschitz gradient on the feasible region, which is a critical assumption for traditional Frank-Wolfe.
- What problem structure actually drives the success of Khachiyan's method? And might such structure exist anywhere else?
- We resolve this mystery and generalize his method to the much broader class of problems in (P), even while relaxing the exact line-search requirement.

Poisson Image Deblurring with TV Regularization

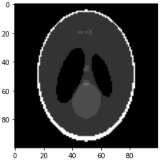


True image X

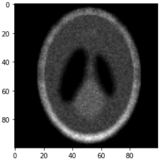


Noisy image Y

Poisson Image Deblurring with TV Regularization



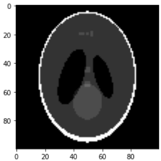
True image X



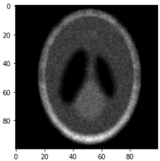
Noisy image Y

- Let an $m \times n$ matrix X denote the true representation of an image, such that $0 \leq X_{ij} \leq M$ denotes the pixel level at location (i, j).

Poisson Image Deblurring with TV Regularization



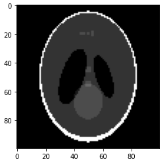
True image X



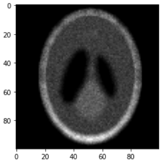
Noisy image Y

- Let an $m \times n$ matrix X denote the true representation of an image, such that $0 \leq X_{ij} \leq M$ denotes the pixel level at location (i, j).
- Let $\mathsf{A} : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ denote the 2D discrete convolutional (linear) operator, which is assumed to be known.

Poisson Image Deblurring with TV Regularization



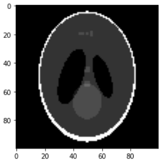
True image X



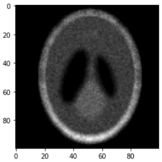
Noisy image Y

- Let an $m \times n$ matrix X denote the true representation of an image, such that $0 \leq X_{ij} \leq M$ denotes the pixel level at location (i, j).
- Let $\mathsf{A} : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ denote the 2D discrete convolutional (linear) operator, which is assumed to be known.
- The observed image Y is obtained by first passing X through A, and then subject to additive independent (entry-wise) Poisson noise.

Poisson Image Deblurring with TV Regularization



True image X



Noisy image Y

- Let an $m \times n$ matrix X denote the true representation of an image, such that $0 \le X_{ij} \le M$ denotes the pixel level at location (i, j).
- Let $\mathsf{A} : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ denote the 2D discrete convolutional (linear) operator, which is assumed to be known.
- The observed image Y is obtained by first passing X through A, and then subject to additive independent (entry-wise) Poisson noise.
- For convenience, we also represent A in its matrix form $A \in \mathbb{R}^{N \times N}$, where $N := mn$, and vectorize Y and X into $y \in \mathbb{R}^N$ and $x \in \mathbb{R}^N$, respectively. Notation: we write $x = \mathsf{vec}(X)$ and $X = \mathsf{mat}(x)$, etc.

Poisson Image Deblurring with TV Regularization

Poisson Image Deblurring with TV Regularization

- We seek to recover X from Y (equivalently x from y) using maximum-likelihood estimation on the TV-regularized problem:

$$\begin{aligned} \min_{x\in\mathbb{R}^N} \quad & \bar{F}(x) := -\textstyle\sum_{l=1}^N y_l \ln(a_l^\top x) + (\sum_{l=1}^N a_l)^\top x + \lambda \mathrm{TV}(x) \\ \text{s.t.} \quad & 0 \le x \le Me \end{aligned} \tag{\texttt{Deblur}}$$

Poisson Image Deblurring with TV Regularization

- We seek to recover X from Y (equivalently x from y) using maximum-likelihood estimation on the TV-regularized problem:

$$\begin{aligned} \min_{x\in\mathbb{R}^N} \quad & \bar{F}(x) := -\textstyle\sum_{l=1}^N y_l \ln(a_l^\top x) + (\sum_{l=1}^N a_l)^\top x + \lambda \mathrm{TV}(x) \\ \text{s.t.} \quad & 0 \le x \le Me \end{aligned} \qquad \texttt{(Deblur)}$$

- (`Deblur`) has a (standard) total-variation (TV) regularization term to recover a smooth image with sharp edges. The TV term is given by

$$\begin{aligned} \mathrm{TV}(x) := & \textstyle\sum_{i=1}^m \sum_{j=1}^{n-1} \left|[\mathsf{mat}(x)]_{i,j} - [\mathsf{mat}(x)]_{i,j+1}\right| \\ & + \textstyle\sum_{i=1}^{m-1} \sum_{j=1}^{n} \left|[\mathsf{mat}(x)]_{i,j} - [\mathsf{mat}(x)]_{i+1,j}\right|. \end{aligned}$$

Some Other Applications

Some Other Applications

- Analyzing social networks (learning of Multivariate Hawkes processes)

Some Other Applications

- Analyzing social networks (learning of Multivariate Hawkes processes)
- Medical imaging reconstruction (Positron emission tomography)

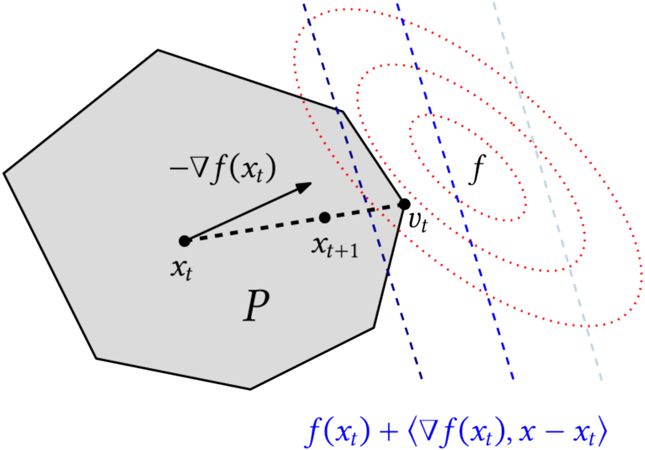
Some Other Applications

- Analyzing social networks (learning of Multivariate Hawkes processes)
- Medical imaging reconstruction (Positron emission tomography)
- Quantum physics (quantum state tomography)

Some Other Applications

- Analyzing social networks (learning of Multivariate Hawkes processes)
- Medical imaging reconstruction (Positron emission tomography)
- Quantum physics (quantum state tomography)
- Computational geometry (computing the analytic center of a polytope)

A Simple Illustration When $h = \iota_{\mathcal{P}}$



New Generalized Frank-Wolfe Method (gFW-LHSCB)

$$F^* := \min_{x\in\mathbb{R}^n} \left[F(x) := f(\mathsf{A}x) + h(x)\right] \qquad \texttt{(P)}$$

New Generalized Frank-Wolfe Method (gFW-LHSCB)

$$F^* := \min_{x\in\mathbb{R}^n} \left[F(x) := f(\mathsf{A}x) + h(x)\right] \tag{P}$$

- **Initialize**: $x^0 \in \mathsf{dom}\, F$, $k := 0$

New Generalized Frank-Wolfe Method (gFW-LHSCB)

$$F^* := \min_{x \in \mathbb{R}^n} \; [F(x) := f(\mathsf{A}x) + h(x)] \tag{P}$$

- **Initialize**: $x^0 \in \mathsf{dom}\, F$, $k := 0$
- **Repeat** (until some convergence criterion is met)

$$v^k \in \arg\min_{x \in \mathbb{R}^n} \langle \nabla f(\mathsf{A}x^k), \mathsf{A}x \rangle + h(x) \qquad \text{(Solve Lin. subproblem)}$$

New Generalized Frank-Wolfe Method (gFW-LHSCB)

$$F^* := \min_{x\in\mathbb{R}^n} \left[F(x) := f(\mathsf{A}x) + h(x)\right] \qquad \text{(P)}$$

- **Initialize**: $x^0 \in \mathsf{dom}\, F$, $k := 0$
- **Repeat** (until some convergence criterion is met)

$$v^k \in \arg\min_{x\in\mathbb{R}^n} \langle \nabla f(\mathsf{A}x^k), \mathsf{A}x\rangle + h(x) \qquad \text{(Solve Lin. subproblem)}$$

$$G_k := \langle \nabla f(\mathsf{A}x^k), \mathsf{A}(x^k - v^k)\rangle + h(x^k) - h(v^k) \qquad \text{(FW-Gap)}$$

New Generalized Frank-Wolfe Method (gFW-LHSCB)

$$F^* := \min_{x\in\mathbb{R}^n} \left[F(x) := f(\mathsf{A}x) + h(x)\right] \qquad \texttt{(P)}$$

- **Initialize**: $x^0 \in \mathsf{dom}\, F$, $k := 0$
- **Repeat** (until some convergence criterion is met)

$$v^k \in \arg\min_{x\in\mathbb{R}^n} \langle \nabla f(\mathsf{A}x^k), \mathsf{A}x\rangle + h(x) \qquad \text{(Solve Lin. subproblem)}$$

$$G_k := \langle \nabla f(\mathsf{A}x^k), \mathsf{A}(x^k - v^k)\rangle + h(x^k) - h(v^k) \qquad \text{(FW-Gap)}$$

$$D_k := \|\mathsf{A}(v^k - x^k)\|_{\mathsf{A}x^k} \qquad \text{(Local Distance)}$$

New Generalized Frank-Wolfe Method (gFW-LHSCB)

$$F^* := \min_{x\in\mathbb{R}^n} \left[F(x) := f(\mathsf{A}x) + h(x)\right] \tag{P}$$

- **Initialize**: $x^0 \in \mathsf{dom}\, F$, $k := 0$
- **Repeat** (until some convergence criterion is met)

$$v^k \in \arg\min_{x\in\mathbb{R}^n} \langle \nabla f(\mathsf{A}x^k), \mathsf{A}x\rangle + h(x) \qquad \text{(Solve Lin. subproblem)}$$

$$G_k := \langle \nabla f(\mathsf{A}x^k), \mathsf{A}(x^k - v^k)\rangle + h(x^k) - h(v^k) \qquad \text{(FW-Gap)}$$

$$D_k := \|\mathsf{A}(v^k - x^k)\|_{\mathsf{A}x^k} \qquad \text{(Local Distance)}$$

$$\alpha_k := \min\left\{ \frac{G_k}{D_k(G_k + D_k)}, 1\right\} \qquad \text{(Stepsize)}$$

New Generalized Frank-Wolfe Method (gFW-LHSCB)

$$F^* := \min_{x\in\mathbb{R}^n} \left[F(x) := f(\mathsf{A}x) + h(x)\right] \qquad \texttt{(P)}$$

- **Initialize**: $x^0 \in \mathsf{dom}\, F$, $k := 0$
- **Repeat** (until some convergence criterion is met)

$$v^k \in \arg\min_{x\in\mathbb{R}^n} \langle \nabla f(\mathsf{A}x^k), \mathsf{A}x\rangle + h(x) \qquad \text{(Solve Lin. subproblem)}$$

$$G_k := \langle \nabla f(\mathsf{A}x^k), \mathsf{A}(x^k - v^k)\rangle + h(x^k) - h(v^k) \qquad \text{(FW-Gap)}$$

$$D_k := \|\mathsf{A}(v^k - x^k)\|_{\mathsf{A}x^k} \qquad \text{(Local Distance)}$$

$$\alpha_k := \min\left\{\frac{G_k}{D_k(G_k + D_k)}, 1\right\} \qquad \text{(Stepsize)}$$

$$x^{k+1} := x^k + \alpha_k(v^k - x^k) \qquad \text{(Update)}$$

New Generalized Frank-Wolfe Method (gFW-LHSCB)

$$F^* := \min_{x\in\mathbb{R}^n} \left[F(x) := f(\mathsf{A}x) + h(x)\right] \tag{P}$$

- **Initialize**: $x^0 \in \mathsf{dom}\, F$, $k := 0$
- **Repeat** (until some convergence criterion is met)

$$v^k \in \arg\min_{x\in\mathbb{R}^n} \langle \nabla f(\mathsf{A}x^k), \mathsf{A}x\rangle + h(x) \qquad \text{(Solve Lin. subproblem)}$$

$$G_k := \langle \nabla f(\mathsf{A}x^k), \mathsf{A}(x^k - v^k)\rangle + h(x^k) - h(v^k) \qquad \text{(FW-Gap)}$$

$$D_k := \|\mathsf{A}(v^k - x^k)\|_{\mathsf{A}x^k} \qquad \text{(Local Distance)}$$

$$\alpha_k := \min\left\{\frac{G_k}{D_k(G_k + D_k)}, 1\right\} \qquad \text{(Stepsize)}$$

$$x^{k+1} := x^k + \alpha_k(v^k - x^k) \qquad \text{(Update)}$$

$$k := k + 1$$

Remarks on gFW-LHSCB

Remarks on gFW-LHSCB

- For most applications (including all of the applications mentioned previously), D_k in (Local Distance) can be computed in $O(n)$ time.

Remarks on gFW-LHSCB

- For most applications (including all of the applications mentioned previously), D_k in (Local Distance) can be computed in $O(n)$ time.
- The FW-gap G_k provides an effective stopping criterion:
$$G_k \geq [\delta_k := F(x^k) - F^*], \quad \text{for all } k \geq 0.$$

Remarks on gFW-LHSCB

- For most applications (including all of the applications mentioned previously), D_k in (Local Distance) can be computed in $O(n)$ time.
- The FW-gap G_k provides an effective stopping criterion:
$$G_k \geq [\delta_k := F(x^k) - F^*], \quad \text{for all } k \geq 0.$$
- The step-size rule in (Stepsize) is derived from the "curvature property" of a (standard strongly non-degenerate) self-concordant function:
$$f(x^k + \alpha(v^k - x^k)) \leq f(x^k) - \alpha G_k + \omega(\alpha D_k), \qquad \texttt{(Curvature)}$$
where $\omega(t) := -t - \ln(1-t)$ for $t < 1$.

Remarks on gFW-LHSCB

- ▷ For most applications (including all of the applications mentioned previously), D_k in (Local Distance) can be computed in $O(n)$ time.
- ▷ The FW-gap G_k provides an effective stopping criterion:
$$G_k \geq [\delta_k := F(x^k) - F^*], \quad \text{for all } k \geq 0.$$
- ▷ The step-size rule in (Stepsize) is derived from the "curvature property" of a (standard strongly non-degenerate) self-concordant function:
$$f(x^k + \alpha(v^k - x^k)) \leq f(x^k) - \alpha G_k + \omega(\alpha D_k), \qquad \texttt{(Curvature)}$$
where $\omega(t) := -t - \ln(1-t)$ for $t < 1$.
- ▷ For some applications (e.g., PET and D-optimal design), the step-size can also be efficiently computed via exact line-search.

Computational Guarantees

Define $R_h := \max_{x,y\in \mathsf{dom}\, h} \ |h(x) - h(y)|$ (the variation of h on its domain)

Recall that δ_0 is the initial optimality gap

Theorem:

Computational Guarantees

Define $R_h := \max_{x,y \in \mathsf{dom}\, h} \; |h(x) - h(y)|$ (the variation of h on its domain)

Recall that δ_0 is the initial optimality gap

Theorem:

- (Iteration complexity for ε-optimality gap) Let K_ε denote the number of iterations required by gFW-LHSCB to obtain $\delta_k \le \varepsilon$. Then:

$$K_\varepsilon \le \lceil 5.3(\delta_0 + \theta + R_h) \ln(10.6\delta_0) \rceil + \left\lceil 12(\theta + R_h)^2 \max\left\{ \frac{1}{\varepsilon} - \frac{1}{\delta_0} \; , 0 \right\} \right\rceil \; .$$

Computational Guarantees

Define $R_h := \max_{x,y \in \mathsf{dom}\, h} \; |h(x) - h(y)|$ (the variation of h on its domain)

Recall that δ_0 is the initial optimality gap

Theorem:

- (Iteration complexity for ε-optimality gap) Let K_ε denote the number of iterations required by gFW-LHSCB to obtain $\delta_k \le \varepsilon$. Then:

$$K_\varepsilon \le \lceil 5.3(\delta_0 + \theta + R_h) \ln(10.6\delta_0) \rceil + \left\lceil 12(\theta + R_h)^2 \max\left\{ \frac{1}{\varepsilon} - \frac{1}{\delta_0} \, , 0 \right\} \right\rceil \, .$$

- (Iteration complexity for ε-FW gap) Let $\mathrm{FWGAP}_\varepsilon$ denote the number of iterations required by gFW-LHSCB to obtain $G_k \le \varepsilon$. Then:

$$\mathrm{FWGAP}_\varepsilon \le \lceil 5.3(\delta_0 + \theta + R_h) \ln(10.6\delta_0) \rceil + \left\lceil \frac{24(\theta + R_h)^2}{\varepsilon} \right\rceil \, .$$

Remarks on the Computational Guarantees

Our computational guarantees only depend on three (natural) quantities:

Remarks on the Computational Guarantees

Our computational guarantees only depend on three (natural) quantities:

- the initial optimality gap δ_0,

Remarks on the Computational Guarantees

Our computational guarantees only depend on three (natural) quantities:

- the initial optimality gap δ_0,
- the complexity parameter θ of the barrier f,

Remarks on the Computational Guarantees

Our computational guarantees only depend on three (natural) quantities:

- the initial optimality gap δ_0,
- the complexity parameter θ of the barrier f,
- the variation of h on its domain $\mathsf{dom}\, h$ ($= 0$ if $h = \iota_{\mathcal{X}}$).

Remarks on the Computational Guarantees

Our computational guarantees only depend on three (natural) quantities:

- the initial optimality gap δ_0,
- the complexity parameter θ of the barrier f,
- the variation of h on its domain $\mathsf{dom}\, h$ ($= 0$ if $h = \iota_{\mathcal{X}}$).

For many applications, all of the three quantities can be easily estimated, and hence the computational guarantees are known before running the algorithm.

Computational Experiments on Poisson Image Deblurring with TV Regularization

$$\min_{x\in\mathbb{R}^N} \;\; \bar{F}(x) := \underbrace{-\textstyle\sum_{l=1}^N y_l \ln(a_l^\top x)}_{=f(\mathsf{A}x)} + \underbrace{\langle \textstyle\sum_{l=1}^N a_l, x\rangle + \lambda \mathrm{TV}(x)}_{=h(x)} \qquad \texttt{(Deblur)}$$

$$\text{s.t.} \quad 0 \le x \le Me \; ,$$

Computational Experiments on Poisson Image Deblurring with TV Regularization

$$\min_{x\in\mathbb{R}^N} \quad \bar{F}(x) := \underbrace{-\sum_{l=1}^N y_l \ln(a_l^\top x)}_{=f(\mathsf{A}x)} + \underbrace{\langle \sum_{l=1}^N a_l, x\rangle + \lambda \mathrm{TV}(x)}_{=h(x)} \qquad \texttt{(Deblur)}$$
$$\text{s.t.} \quad 0 \le x \le Me\ ,$$

▷ Since $\mathrm{TV}(\cdot)$ is polyhedral, and the linear-optimization sub-problem

$$v^k \in \arg\min_{0\le x\le Me} \langle \nabla f(\mathsf{A}x^k), \mathsf{A}x\rangle + \langle \sum_{l=1}^N a_l, x\rangle + \lambda \mathrm{TV}(x)$$

can be formulated as a relatively simple LP and solved easily using a standard LP solver such as Gurobi.

Implementation Details/Issues

Implementation Details/Issues

- We evaluate the numerical performance of our FW method gFW-LHSCB (with adaptive stepsize) which we call `FW-Adapt`.

Implementation Details/Issues

- We evaluate the numerical performance of our FW method gFW-LHSCB (with adaptive stepsize) which we call `FW-Adapt`.
- It turns out that an exact line-search step-size for gFW-LHSCB can be computed for this particular problem, which we call `FW-Exact`.

Implementation Details/Issues

- We evaluate the numerical performance of our FW method gFW-LHSCB (with adaptive stepsize) which we call `FW-Adapt`.
- It turns out that an exact line-search step-size for gFW-LHSCB can be computed for this particular problem, which we call `FW-Exact`.
- We tested `FW-Adapt` and `FW-Exact` on the Shepp-Logan phantom image of size 100×100 (hence $N = 10,000$).

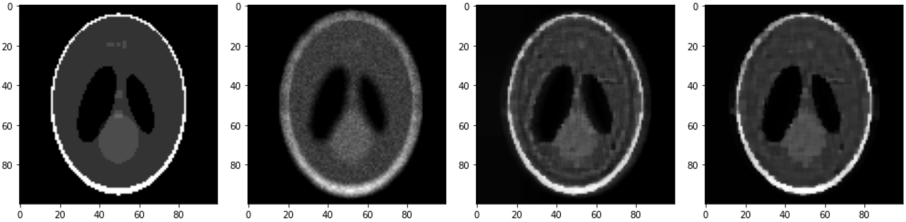
Implementation Details/Issues

- We evaluate the numerical performance of our FW method gFW-LHSCB (with adaptive stepsize) which we call `FW-Adapt`.
- It turns out that an exact line-search step-size for gFW-LHSCB can be computed for this particular problem, which we call `FW-Exact`.
- We tested `FW-Adapt` and `FW-Exact` on the Shepp-Logan phantom image of size 100×100 (hence $N = 10,000$).
- We chose the starting point $x^0 = \mathsf{vec}(Y)$ (the vectorized noisy image), and we set $\lambda = 0.01$.

Implementation Details/Issues

- We evaluate the numerical performance of our FW method gFW-LHSCB (with adaptive stepsize) which we call `FW-Adapt`.
- It turns out that an exact line-search step-size for gFW-LHSCB can be computed for this particular problem, which we call `FW-Exact`.
- We tested `FW-Adapt` and `FW-Exact` on the Shepp-Logan phantom image of size 100×100 (hence $N = 10,000$).
- We chose the starting point $x^0 = \mathsf{vec}(Y)$ (the vectorized noisy image), and we set $\lambda = 0.01$.
- We used CVXPY to (approximately) compute the optimal objective value $\bar{F}^*$ of (`Deblur`) in order to compute optimality gaps.

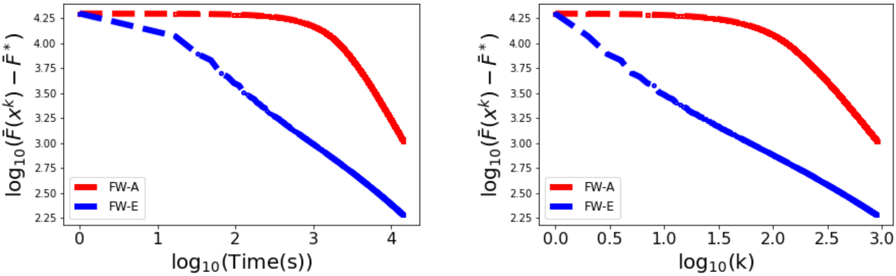
Results: Recovered Images



(k) True image X (l) Noisy image Y (m) `FW-Adapt` (n) `FW-Exact`

Figure 1: True, noisy and recovered Shepp-Logan phantom images.

Results: Optimality Gaps versus Time and Iterations



(a) Optimality gap versus time (in seconds)

(b) Optimality gap versus iterations

Figure 2: Comparison of optimality gaps of `FW-Adapt` (`FW-A`) and `FW-Exact` (`FW-E`) for image recovery of the Shepp-Logan phantom image.

Motivating Example: D-Optimal Design

$$\begin{aligned} \min \;\; & f(x) := -\ln\det\left(\textstyle\sum_{i=1}^m x_i a_i a_i^\top\right) \\ \text{s.t.} \;\; & x \in \Delta_m := \{\textstyle\sum_{i=1}^m x_i = 1,\ x_i \ge 0,\ \forall\, i \in [m]\}. \end{aligned} \tag{D-OPT}$$

Motivating Example: D-Optimal Design

$$\begin{aligned} \min \;\; & f(x) := -\ln\det\left(\textstyle\sum_{i=1}^m x_i a_i a_i^\top\right) \\ \text{s.t.} \;\; & x \in \Delta_m := \{\textstyle\sum_{i=1}^m x_i = 1,\ x_i \ge 0,\ \forall\, i \in [m]\}. \end{aligned} \tag{D-OPT}$$

- Problem data: m points $\{a_i\}_{i=1}^m$ that span $\mathbb{R}^n$.

Motivating Example: D-Optimal Design

$$\begin{aligned} \min \quad & f(x) := -\ln\det\left(\textstyle\sum_{i=1}^m x_i a_i a_i^\top\right) \\ \text{s.t.} \quad & x \in \Delta_m := \{\textstyle\sum_{i=1}^m x_i = 1,\ x_i \ge 0,\ \forall\, i \in [m]\}. \end{aligned} \qquad \texttt{(D-OPT)}$$

- Problem data: m points $\{a_i\}_{i=1}^m$ that span $\mathbb{R}^n$.
- In statistics, (D-OPT) is the continuous relaxation of the (discrete) D-optimal experimental design problem; in computational geometry, it is the dual problem of the minimum volume enclosing ellipsoid (MVEE) problem.

Motivating Example: D-Optimal Design

$$\begin{aligned} \min \;\; & f(x) := -\ln\det\left(\textstyle\sum_{i=1}^m x_i a_i a_i^\top\right) \\ \text{s.t.} \;\; & x \in \Delta_m := \{\textstyle\sum_{i=1}^m x_i = 1,\; x_i \geq 0,\; \forall\, i \in [m]\}. \end{aligned} \tag{D-OPT}$$

- Problem data: m points $\{a_i\}_{i=1}^m$ that span $\mathbb{R}^n$.
- In statistics, (D-OPT) is the continuous relaxation of the (discrete) D-optimal experimental design problem; in computational geometry, it is the dual problem of the minimum volume enclosing ellipsoid (MVEE) problem.
- Despite its seemingly simple structure, (D-OPT) is not quite amenable to (traditional) first-order methods (since f blows up on part of $\partial\Delta_m$, and has no L-smoothness property on Δ_m).

Motivating Example: D-Optimal Design

$$\begin{aligned} \min \ & f(x) := -\ln\det\left(\textstyle\sum_{i=1}^m x_i a_i a_i^\top\right) \\ \text{s.t.} \ & x \in \Delta_m := \{\textstyle\sum_{i=1}^m x_i = 1,\ x_i \geq 0,\ \forall\, i \in [m]\}. \end{aligned} \tag{D-OPT}$$

- Problem data: m points $\{a_i\}_{i=1}^m$ that span $\mathbb{R}^n$.
- In statistics, (D-OPT) is the continuous relaxation of the (discrete) D-optimal experimental design problem; in computational geometry, it is the dual problem of the minimum volume enclosing ellipsoid (MVEE) problem.
- Despite its seemingly simple structure, (D-OPT) is not quite amenable to (traditional) first-order methods (since f blows up on part of $\partial\Delta_m$, and has no L-smoothness property on Δ_m).
- Atwood (1973) proposed the following algorithm for solving (D-OPT):

Motivating Example: D-Optimal Design

$$\begin{aligned} &\min \;\; f(x) := -\ln\det\left(\textstyle\sum_{i=1}^m x_i a_i a_i^\top\right) \\ &\text{s.t.} \;\; x \in \Delta_m := \{\textstyle\sum_{i=1}^m x_i = 1,\; x_i \ge 0,\; \forall\, i \in [m]\}. \end{aligned} \qquad \text{(D-OPT)}$$

- Problem data: m points $\{a_i\}_{i=1}^m$ that span $\mathbb{R}^n$.
- In statistics, (D-OPT) is the continuous relaxation of the (discrete) D-optimal experimental design problem; in computational geometry, it is the dual problem of the minimum volume enclosing ellipsoid (MVEE) problem.
- Despite its seemingly simple structure, (D-OPT) is not quite amenable to (traditional) first-order methods (since f blows up on part of $\partial\Delta_m$, and has no L-smoothness property on Δ_m).
- Atwood (1973) proposed the following algorithm for solving (D-OPT):

$$i_k \in \arg\min_{i\in[m]} \nabla_i f(x^k), \qquad G_k := -\nabla_{i_k} f(x^k) - n,$$

$$j_k \in \arg\max_{j: x_j^k > 0} \nabla_i f(x^k), \quad \tilde{G}_k := \nabla_{j_k} f(x^k) + n,$$

$$d^k = \begin{cases} e_{i_k} - x^k & \text{if } G_k > \tilde{G}_k \\ x^k - e_{j_k} & \text{otherwise} \end{cases}, \qquad x^{k+1} := x^k + \alpha_k d^k,$$

where the stepsize $\alpha_k \ge 0$ is given by exact line-search.

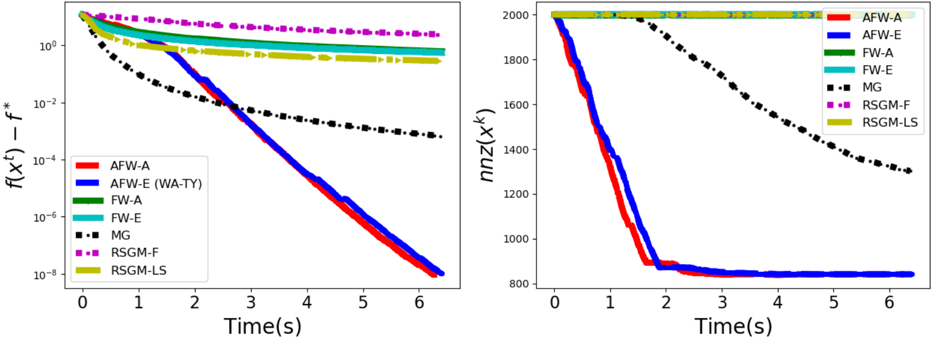
The WA-TY Method

The WA-TY Method

- Structurally, this method coincides with the Frank-Wolfe method with Wolfe's away-step (1970), and it was rediscovered by Todd and Yıldırım (2005) — therefore, it is referred to as the WA-TY method.

The WA-TY Method

- ▷ Structurally, this method coincides with the Frank-Wolfe method with Wolfe's away-step (1970), and it was rediscovered by Todd and Yıldırım (2005) — therefore, it is referred to as the WA-TY method.
- ▷ Excellent numerical performance:



ASFW-A & ASFW-E (this work): Away-step FW methods for LHB
FW-A & FW-E [Fed72; Kha96; ZFce]: Generalized FW methods for LHB
RSGM-F & RSGM-LS [BBT17; LFN18]: Relatively smooth gradient method
MG [STT78]: Multiplicative gradient method

Mystery of the WA-TY Method

Mystery of the WA-TY Method

- The excellent numerical performance of the WA-TY method has attracted some research interests — Ahipasaoglu, Sun and Todd (2008) showed *local* linear convergence of this method, but the *global* linear convergence remains open.

Mystery of the WA-TY Method

- The excellent numerical performance of the WA-TY method has attracted some research interests — Ahipasaoglu, Sun and Todd (2008) showed *local* linear convergence of this method, but the *global* linear convergence remains open.

- The authors pointed out two structural difficulties of (`D-OPT`): i) f is not L-smooth on Δ_m and ii) f is degenerate on the feasible region.

Mystery of the WA-TY Method

- The excellent numerical performance of the WA-TY method has attracted some research interests — Ahipasaoglu, Sun and Todd (2008) showed *local* linear convergence of this method, but the *global* linear convergence remains open.
- The authors pointed out two structural difficulties of (D-OPT): i) f is not L-smooth on Δ_m and ii) f is degenerate on the feasible region.
- This difficulty prevents the recent analyses of the away-step FW (AFW) methods for L-smooth functions [LJJ15; BS17; PR19], as well as for *non-degenerate* generalized self-concordant function [Dvu+23] being applied to (D-OPT).

Mystery of the WA-TY Method

- The excellent numerical performance of the WA-TY method has attracted some research interests — Ahipasaoglu, Sun and Todd (2008) showed *local* linear convergence of this method, but the *global* linear convergence remains open.
- The authors pointed out two structural difficulties of (D-OPT): i) f is not L-smooth on Δ_m and ii) f is degenerate on the feasible region.
- This difficulty prevents the recent analyses of the away-step FW (AFW) methods for L-smooth functions [LJJ15; BS17; PR19], as well as for *non-degenerate* generalized self-concordant function [Dvu+23] being applied to (D-OPT).
- Some deeper questions:
  - What is the essential structure of (D-OPT) that drives the linear convergence of the WA-TY method (or the AFW method)?
  - Can it help us develop and analyze a new type of AFW methods for an "unconventional" class of problems?

Mystery of the WA-TY Method

- The excellent numerical performance of the WA-TY method has attracted some research interests — Ahipasaoglu, Sun and Todd (2008) showed *local* linear convergence of this method, but the *global* linear convergence remains open.
- The authors pointed out two structural difficulties of (D-OPT): i) f is not L-smooth on Δ_m and ii) f is degenerate on the feasible region.
- This difficulty prevents the recent analyses of the away-step FW (AFW) methods for L-smooth functions [LJJ15; BS17; PR19], as well as for *non-degenerate* generalized self-concordant function [Dvu+23] being applied to (D-OPT).
- Some deeper questions:
  - What is the essential structure of (D-OPT) that drives the linear convergence of the WA-TY method (or the AFW method)?
  - Can it help us develop and analyze a new type of AFW methods for an "unconventional" class of problems?
- In this work, we will provide affirmative answers to the questions above.

Problem of Interest

$$F^* := \min_{x \in \mathcal{X}} \left[F(x) := f(\mathsf{A}x) + \langle c, x \rangle \right] \tag{P}$$

Problem of Interest

$$F^* := \min_{x \in \mathcal{X}} \left[F(x) := f(\mathsf{A}x) + \langle c, x \rangle \right] \qquad \texttt{(P)}$$

- $\mathbb{X}$ and $\mathbb{Y}$ are finite-dimensional vector spaces

Problem of Interest

$$F^* := \min_{x\in\mathcal{X}} \left[F(x) := f(\mathsf{A}x) + \langle c, x\rangle\right] \qquad \texttt{(P)}$$

- $\mathbb{X}$ and $\mathbb{Y}$ are finite-dimensional vector spaces
- $\mathcal{X} \subseteq \mathbb{X}$ is a polytope such that $\mathcal{X} = \mathsf{conv}(\mathcal{V})$, where $\mathcal{V}$ is a finite set of *atoms*

Problem of Interest

$$F^* := \min_{x\in\mathcal{X}} \left[F(x) := f(\mathsf{A}x) + \langle c, x\rangle\right] \qquad \texttt{(P)}$$

- $\mathbb{X}$ and $\mathbb{Y}$ are finite-dimensional vector spaces
- $\mathcal{X} \subseteq \mathbb{X}$ is a polytope such that $\mathcal{X} = \mathsf{conv}(\mathcal{V})$, where $\mathcal{V}$ is a finite set of *atoms*
- $f : \mathbb{Y} \to \mathbb{R} \cup \{+\infty\}$ is a θ-log-homogeneous self-concordant barrier (θ-LHSCB) for some regular cone $\mathcal{K} \subseteq \mathbb{Y}$

Problem of Interest

$$F^* := \min_{x \in \mathcal{X}} \left[F(x) := f(\mathsf{A}x) + \langle c, x \rangle \right] \qquad \texttt{(P)}$$

- $\mathbb{X}$ and $\mathbb{Y}$ are finite-dimensional vector spaces
- $\mathcal{X} \subseteq \mathbb{X}$ is a polytope such that $\mathcal{X} = \mathsf{conv}(\mathcal{V})$, where $\mathcal{V}$ is a finite set of *atoms*
- $f : \mathbb{Y} \to \mathbb{R} \cup \{+\infty\}$ is a θ-log-homogeneous self-concordant barrier (θ-LHSCB) for some regular cone $\mathcal{K} \subseteq \mathbb{Y}$
- $\mathsf{A} : \mathbb{X} \to \mathbb{Y}$ is a linear operator such that $\mathsf{A}(\mathcal{X}) \subseteq \mathcal{K}$ and $\mathsf{A}(\mathcal{X}) \cap \mathsf{int}\,\mathcal{K} \neq \emptyset$

Problem of Interest

$$F^* := \min_{x \in \mathcal{X}} \left[F(x) := f(\mathsf{A}x) + \langle c, x \rangle \right] \qquad \texttt{(P)}$$

- $\mathbb{X}$ and $\mathbb{Y}$ are finite-dimensional vector spaces
- $\mathcal{X} \subseteq \mathbb{X}$ is a polytope such that $\mathcal{X} = \mathsf{conv}(\mathcal{V})$, where $\mathcal{V}$ is a finite set of *atoms*
- $f : \mathbb{Y} \to \mathbb{R} \cup \{+\infty\}$ is a θ-log-homogeneous self-concordant barrier (θ-LHSCB) for some regular cone $\mathcal{K} \subseteq \mathbb{Y}$
- $\mathsf{A} : \mathbb{X} \to \mathbb{Y}$ is a linear operator such that $\mathsf{A}(\mathcal{X}) \subseteq \mathcal{K}$ and $\mathsf{A}(\mathcal{X}) \cap \mathsf{int}\,\mathcal{K} \neq \emptyset$
- $\langle c, \cdot \rangle : \mathbb{X} \to \mathbb{R}$ is a linear function

Problem of Interest

$$F^* := \min_{x \in \mathcal{X}} \left[F(x) := f(\mathsf{A}x) + \langle c, x \rangle \right] \qquad \texttt{(P)}$$

- $\mathbb{X}$ and $\mathbb{Y}$ are finite-dimensional vector spaces
- $\mathcal{X} \subseteq \mathbb{X}$ is a polytope such that $\mathcal{X} = \mathsf{conv}(\mathcal{V})$, where $\mathcal{V}$ is a finite set of *atoms*
- $f : \mathbb{Y} \to \mathbb{R} \cup \{+\infty\}$ is a θ-log-homogeneous self-concordant barrier (θ-LHSCB) for some regular cone $\mathcal{K} \subseteq \mathbb{Y}$
- $\mathsf{A} : \mathbb{X} \to \mathbb{Y}$ is a linear operator such that $\mathsf{A}(\mathcal{X}) \subseteq \mathcal{K}$ and $\mathsf{A}(\mathcal{X}) \cap \mathsf{int}\, \mathcal{K} \neq \emptyset$
- $\langle c, \cdot \rangle : \mathbb{X} \to \mathbb{R}$ is a linear function
- Besides D-optimal design, other applications include
  - Budget-constrained D-optimal design
  - Positron emission tomography
  - (Reformulated) Poisson image deblurring with TV-regularization

Away-step Frank-Wolfe Method for solving (P)

Away-step Frank-Wolfe Method for solving (P)

- **Input**: $x^0 \in \mathcal{X} \cap \mathsf{dom}\, F$, $\beta^0 \in \Delta_{|\mathcal{V}|}$ such that $x^0 = \sum_{v \in \mathcal{V}} \beta_v^0 v$, $\mathcal{S}_0 := \mathrm{supp}(\beta^0)$.

Away-step Frank-Wolfe Method for solving (P)

- **Input**: $x^0 \in \mathcal{X} \cap \mathsf{dom}\, F$, $\beta^0 \in \Delta_{|\mathcal{V}|}$ such that $x^0 = \sum_{v \in \mathcal{V}} \beta_v^0 v$, $\mathcal{S}_0 := \mathrm{supp}(\beta^0)$.
- **At iteration** $k \geq 0$:

Away-step Frank-Wolfe Method for solving (P)

- **Input**: $x^0 \in \mathcal{X} \cap \mathsf{dom}\, F$, $\beta^0 \in \Delta_{|\mathcal{V}|}$ such that $x^0 = \sum_{v \in \mathcal{V}} \beta^0_v v$, $\mathcal{S}_0 := \mathrm{supp}(\beta^0)$.
- **At iteration** $k \geq 0$:
  - (FW direction) Compute $v^k \in \arg\min_{x \in \mathcal{V}} \langle \nabla F(x^k), x \rangle$, $d^k_{\mathrm{F}} := v^k - x^k$ and $G_k := \langle -\nabla F(x^k), d^k_{\mathrm{F}} \rangle$. If $G_k = 0$, then STOP.

Away-step Frank-Wolfe Method for solving (P)

- **Input**: $x^0 \in \mathcal{X} \cap \mathsf{dom}\, F$, $\beta^0 \in \Delta_{|\mathcal{V}|}$ such that $x^0 = \sum_{v\in\mathcal{V}} \beta_v^0 v$, $\mathcal{S}_0 := \mathrm{supp}(\beta^0)$.
- **At iteration** $k \geq 0$:
  - (FW direction) Compute $v^k \in \arg\min_{x\in\mathcal{V}} \langle \nabla F(x^k), x\rangle$, $d_{\mathrm{F}}^k := v^k - x^k$ and $G_k := \langle -\nabla F(x^k), d_{\mathrm{F}}^k\rangle$. If $G_k = 0$, then STOP.
  - (Away direction) If $|\mathcal{S}_k| > 1$, compute $a^k \in \arg\max_{x\in\mathcal{S}_k} \langle \nabla F(x^k), x\rangle$, $d_{\mathrm{A}}^k := x^k - a^k$ and $\tilde{G}_k := \langle -\nabla F(x^k), d_{\mathrm{A}}^k\rangle$.

Away-step Frank-Wolfe Method for solving (P)

- **Input**: $x^0 \in \mathcal{X} \cap \mathsf{dom}\, F$, $\beta^0 \in \Delta_{|\mathcal{V}|}$ such that $x^0 = \sum_{v \in \mathcal{V}} \beta_v^0 v$, $\mathcal{S}_0 := \mathrm{supp}(\beta^0)$.
- **At iteration** $k \ge 0$:
  - (FW direction) Compute $v^k \in \arg\min_{x \in \mathcal{V}} \langle \nabla F(x^k), x \rangle$, $d_{\mathrm{F}}^k := v^k - x^k$ and $G_k := \langle -\nabla F(x^k), d_{\mathrm{F}}^k \rangle$. If $G_k = 0$, then STOP.
  - (Away direction) If $|\mathcal{S}_k| > 1$, compute $a^k \in \arg\max_{x \in \mathcal{S}_k} \langle \nabla F(x^k), x \rangle$, $d_{\mathrm{A}}^k := x^k - a^k$ and $\tilde{G}_k := \langle -\nabla F(x^k), d_{\mathrm{A}}^k \rangle$.
  - (Choose direction) If $|\mathcal{S}_k| = 1$ or $G_k > \tilde{G}_k$, let $d^k := d_{\mathrm{F}}^k$ and $\bar{\alpha}_k := 1$; otherwise, let $d^k := d_{\mathrm{A}}^k$ and $\bar{\alpha}_k := \beta_{a^k}^k / (1 - \beta_{a^k}^k)$.

Away-step Frank-Wolfe Method for solving (P)

- **Input**: $x^0 \in \mathcal{X} \cap \mathsf{dom}\, F$, $\beta^0 \in \Delta_{|\mathcal{V}|}$ such that $x^0 = \sum_{v\in\mathcal{V}} \beta_v^0 v$, $\mathcal{S}_0 := \mathrm{supp}(\beta^0)$.
- **At iteration** $k \geq 0$:
  - (FW direction) Compute $v^k \in \arg\min_{x\in\mathcal{V}} \langle \nabla F(x^k), x\rangle$, $d_{\mathrm{F}}^k := v^k - x^k$ and $G_k := \langle -\nabla F(x^k), d_{\mathrm{F}}^k\rangle$. If $G_k = 0$, then STOP.
  - (Away direction) If $|\mathcal{S}_k| > 1$, compute $a^k \in \arg\max_{x\in\mathcal{S}_k} \langle \nabla F(x^k), x\rangle$, $d_{\mathrm{A}}^k := x^k - a^k$ and $\tilde{G}_k := \langle -\nabla F(x^k), d_{\mathrm{A}}^k\rangle$.
  - (Choose direction) If $|\mathcal{S}_k| = 1$ or $G_k > \tilde{G}_k$, let $d^k := d_{\mathrm{F}}^k$ and $\bar{\alpha}_k := 1$; otherwise, let $d^k := d_{\mathrm{A}}^k$ and $\bar{\alpha}_k := \beta_{a^k}^k/(1-\beta_{a^k}^k)$.
  - (Choose stepsize) Choose $\alpha_k \in (0, \bar{\alpha}_k]$ in one of the following two ways:
    - Adaptive stepsize: Compute $r_k := -\langle \nabla F(x^k), d^k\rangle$ and $D_k := \|\mathsf{A}d^k\|_{y^k}$. If $D_k = 0$, then $\alpha_k := \bar{\alpha}_k$; otherwise, $\alpha_k := \min\{b_k, \bar{\alpha}_k\}$, where $b_k := r_k/(D_k(r_k + D_k))$.
    - Exact line-search: $\alpha_k \in \arg\min_{\alpha_k\in(0,\bar{\alpha}_k]} F(x^k + \alpha d^k)$.

Away-step Frank-Wolfe Method for solving (P)

- **Input**: $x^0 \in \mathcal{X} \cap \mathsf{dom}\, F$, $\beta^0 \in \Delta_{|\mathcal{V}|}$ such that $x^0 = \sum_{v\in\mathcal{V}} \beta_v^0 v$, $\mathcal{S}_0 := \mathrm{supp}(\beta^0)$.
- **At iteration** $k \ge 0$:
  - (FW direction) Compute $v^k \in \arg\min_{x\in\mathcal{V}} \langle \nabla F(x^k), x\rangle$, $d_{\mathrm{F}}^k := v^k - x^k$ and $G_k := \langle -\nabla F(x^k), d_{\mathrm{F}}^k\rangle$. If $G_k = 0$, then STOP.
  - (Away direction) If $|\mathcal{S}_k| > 1$, compute $a^k \in \arg\max_{x\in\mathcal{S}_k} \langle \nabla F(x^k), x\rangle$, $d_{\mathrm{A}}^k := x^k - a^k$ and $\tilde{G}_k := \langle -\nabla F(x^k), d_{\mathrm{A}}^k\rangle$.
  - (Choose direction) If $|\mathcal{S}_k| = 1$ or $G_k > \tilde{G}_k$, let $d^k := d_{\mathrm{F}}^k$ and $\bar{\alpha}_k := 1$; otherwise, let $d^k := d_{\mathrm{A}}^k$ and $\bar{\alpha}_k := \beta_{a^k}^k/(1-\beta_{a^k}^k)$.
  - (Choose stepsize) Choose $\alpha_k \in (0, \bar{\alpha}_k]$ in one of the following two ways:
    - Adaptive stepsize: Compute $r_k := -\langle \nabla F(x^k), d^k\rangle$ and $D_k := \|\mathsf{A}d^k\|_{y^k}$. If $D_k = 0$, then $\alpha_k := \bar{\alpha}_k$; otherwise, $\alpha_k := \min\{b_k, \bar{\alpha}_k\}$, where $b_k := r_k/(D_k(r_k + D_k))$.
    - Exact line-search: $\alpha_k \in \arg\min_{\alpha_k\in(0,\bar{\alpha}_k]} F(x^k + \alpha d^k)$.
  - (Update iterates) Update $x^{k+1} := x^k + \alpha_k d^k$ and $\beta^{k+1} \in \Delta_{|\mathcal{V}|}$ such that $x^{k+1} = \sum_{v\in\mathcal{V}} \beta_v^{k+1} v$, and let $\mathcal{S}_{k+1} := \mathrm{supp}(\beta^{k+1})$.

Some Remarks

Denote $\dim \mathbb{X} = n$.

Some Remarks

Denote $\dim \mathbb{X} = n$.

- Depending on $\mathcal{X}$, we may prefer to solve $\min_{x\in\mathcal{V}} \langle \nabla F(x^k), x\rangle$ either by either minimizing over $\mathcal{X}$ (e.g., $\mathcal{X} = \prod_{i=1}^n [a_i, b_i]$) or $\mathcal{V}$ (e.g., $\mathcal{X} = \Delta_n$).

Some Remarks

Denote $\dim \mathbb{X} = n$.

- Depending on $\mathcal{X}$, we may prefer to solve $\min_{x \in \mathcal{V}} \langle \nabla F(x^k), x \rangle$ either by either minimizing over $\mathcal{X}$ (e.g., $\mathcal{X} = \prod_{i=1}^n [a_i, b_i]$) or $\mathcal{V}$ (e.g., $\mathcal{X} = \Delta_n$).

- The FW-gap $G_k = \langle \nabla F(x^k), x^k - v^k \rangle$ provides an effective stopping criterion:
$$G_k \geq [\delta_k := F(x^k) - F^*] \quad \text{for } k \geq 0.$$

Some Remarks

Denote $\dim \mathbb{X} = n$.

- Depending on $\mathcal{X}$, we may prefer to solve $\min_{x\in\mathcal{V}} \langle \nabla F(x^k), x\rangle$ either by either minimizing over $\mathcal{X}$ (e.g., $\mathcal{X} = \prod_{i=1}^n [a_i, b_i]$) or $\mathcal{V}$ (e.g., $\mathcal{X} = \Delta_n$).

- The FW-gap $G_k = \langle \nabla F(x^k), x^k - v^k\rangle$ provides an effective stopping criterion:

$$G_k \geq [\delta_k := F(x^k) - F^*] \quad \text{for } k \geq 0.$$

- If $|\mathcal{V}| = \omega(n)$, we may prefer to maintain a compact representation of $\mathcal{S}_k$ such that $|\mathcal{S}_k| = O(n)$ for $k \geq 0$, at computational cost of $O(n^2)$ per iteration [BS17].

Some Remarks

Denote $\dim \mathbb{X} = n$.

- ▷ Depending on $\mathcal{X}$, we may prefer to solve $\min_{x\in\mathcal{V}} \langle \nabla F(x^k), x\rangle$ either by either minimizing over $\mathcal{X}$ (e.g., $\mathcal{X} = \prod_{i=1}^n [a_i, b_i]$) or $\mathcal{V}$ (e.g., $\mathcal{X} = \Delta_n$).
- ▷ The FW-gap $G_k = \langle \nabla F(x^k), x^k - v^k\rangle$ provides an effective stopping criterion:
$$G_k \ge [\delta_k := F(x^k) - F^*] \quad \text{for } k \ge 0.$$
- ▷ If $|\mathcal{V}| = \omega(n)$, we may prefer to maintain a compact representation of $\mathcal{S}_k$ such that $|\mathcal{S}_k| = O(n)$ for $k \ge 0$, at computational cost of $O(n^2)$ per iteration [BS17].
- ▷ For all applications of interest, computing $D_k = \|\mathsf{A}d^k\|_{y^k} = \langle \nabla^2 F(x^k)d^k, d^k\rangle^{1/2}$ takes $O(n)$ times, instead of $O(n^2)$ time.

Computational Guarantees

$$F^* := \min_{x\in\mathbb{R}^n}\ \left[F(x) := f(\mathsf{A}x) + \langle c, x\rangle\right]$$

Computational Guarantees

$$F^* := \min_{x \in \mathbb{R}^n} \left[F(x) := f(\mathsf{A}x) + \langle c, x \rangle \right]$$

▷ Define $B := \max_{x, x' \in \mathcal{X}} \langle c, x - x' \rangle$ (the variation of $\langle c, \cdot \rangle$ on $\mathcal{X}$).

Computational Guarantees

$$F^* := \min_{x\in\mathbb{R}^n} \; [F(x) := f(\mathsf{A}x) + \langle c, x\rangle]$$

- Define $B := \max_{x,x'\in\mathcal{X}} \langle c, x - x'\rangle$ (the variation of $\langle c, \cdot\rangle$ on $\mathcal{X}$).
- Define $q := \min\{|\mathcal{W}| : \; \mathcal{W} \subseteq \mathcal{V} \;\text{ such that }\; \mathsf{conv}\mathcal{W} \cap \mathsf{dom}\, F \neq \emptyset\}$.

Computational Guarantees

$$F^* := \min_{x\in\mathbb{R}^n}\ [F(x) := f(\mathsf{A}x) + \langle c, x\rangle]$$

- Define $B := \max_{x,x'\in\mathcal{X}} \langle c, x - x'\rangle$ (the variation of $\langle c,\cdot\rangle$ on $\mathcal{X}$).
- Define $q := \min\{|\mathcal{W}| :\ \mathcal{W} \subseteq \mathcal{V}\ \text{ such that }\ \mathsf{conv}\mathcal{W} \cap \mathsf{dom}\, F \neq \emptyset\}$.
- Define $\mathcal{Y} := \mathsf{A}(\mathcal{X})$ and $R_{\mathcal{Y}}(y^*) := \sup_{y\in\mathsf{A}(\mathcal{X})} \|y - y^*\|_{y^*} < +\infty$.

Computational Guarantees

$$F^* := \min_{x\in\mathbb{R}^n} \left[F(x) := f(\mathsf{A}x) + \langle c, x\rangle\right]$$

- Define $B := \max_{x,x'\in\mathcal{X}} \langle c, x - x'\rangle$ (the variation of $\langle c,\cdot\rangle$ on $\mathcal{X}$).
- Define $q := \min\{|\mathcal{W}| : \mathcal{W} \subseteq \mathcal{V}$ such that $\mathsf{conv}\mathcal{W} \cap \mathsf{dom}\, F \neq \emptyset\}$.
- Define $\mathcal{Y} := \mathsf{A}(\mathcal{X})$ and $R_{\mathcal{Y}}(y^*) := \sup_{y\in\mathsf{A}(\mathcal{X})} \|y - y^*\|_{y^*} < +\infty$.

Global linear convergence of $\{\delta_k\}_{k\geq 0}$:

Computational Guarantees

$$F^* := \min_{x\in\mathbb{R}^n} \left[F(x) := f(\mathsf{A}x) + \langle c, x\rangle\right]$$

- ▷ Define $B := \max_{x,x'\in\mathcal{X}} \langle c, x - x'\rangle$ (the variation of $\langle c,\cdot\rangle$ on $\mathcal{X}$).
- ▷ Define $q := \min\{|\mathcal{W}| :\ \mathcal{W} \subseteq \mathcal{V}$ such that $\mathsf{conv}\mathcal{W} \cap \mathsf{dom}\, F \neq \emptyset\}$.
- ▷ Define $\mathcal{Y} := \mathsf{A}(\mathcal{X})$ and $R_{\mathcal{Y}}(y^*) := \sup_{y\in\mathsf{A}(\mathcal{X})} \|y - y^*\|_{y^*} < +\infty$.

Global linear convergence of $\{\delta_k\}_{k\geq 0}$:

- ▷ $\{\delta_k\}_{k\geq 0}$ is strictly decreasing (until termination).

Computational Guarantees

$$F^* := \min_{x\in\mathbb{R}^n} \left[F(x) := f(\mathsf{A}x) + \langle c, x\rangle\right]$$

- ▷ Define $B := \max_{x,x'\in\mathcal{X}} \langle c, x - x'\rangle$ (the variation of $\langle c, \cdot\rangle$ on $\mathcal{X}$).
- ▷ Define $q := \min\{|\mathcal{W}| : \mathcal{W} \subseteq \mathcal{V} \text{ such that } \mathsf{conv}\mathcal{W} \cap \mathsf{dom}\, F \neq \emptyset\}$.
- ▷ Define $\mathcal{Y} := \mathsf{A}(\mathcal{X})$ and $R_{\mathcal{Y}}(y^*) := \sup_{y\in\mathsf{A}(\mathcal{X})} \|y - y^*\|_{y^*} < +\infty$.

Global linear convergence of $\{\delta_k\}_{k\geq 0}$:

- ▷ $\{\delta_k\}_{k\geq 0}$ is strictly decreasing (until termination).
- ▷ For all $k \geq 0$, define $k_{\text{eff}} := \lceil \max\{(k - |\mathcal{S}_0| + q)/2, 0\}\rceil \approx k/2$, and then

$$\delta_k \leq (1-\rho)^{k_{\text{eff}}}\delta_0, \quad \text{where} \quad \rho := \min\left\{\frac{1}{5.3(\delta_0 + \theta + B)}, \frac{\mu\Phi(\mathcal{X},\mathcal{X}^*)^2}{42.4(\theta + B)^2}\right\},$$

  where

  - μ is the quadratic-growth constant of f on $\mathcal{Y}$ that only depends on $R_{\mathcal{Y}}(y^*)$
  - $\Phi(\mathcal{X},\mathcal{X}^*) > 0$ is a geometric constant about $\mathcal{X}^*$ and $\mathcal{X}$.

Computational Guarantees

$$F^* := \min_{x\in\mathbb{R}^n}\ [F(x) := f(\mathsf{A}x) + \langle c, x\rangle]$$

- ▷ Define $B := \max_{x,x'\in\mathcal{X}} \langle c, x - x'\rangle$ (the variation of $\langle c, \cdot\rangle$ on $\mathcal{X}$).
- ▷ Define $q := \min\{|\mathcal{W}| :\ \mathcal{W} \subseteq \mathcal{V}\ \text{such that}\ \mathsf{conv}\mathcal{W} \cap \mathsf{dom}\,F \neq \emptyset\}$.
- ▷ Define $\mathcal{Y} := \mathsf{A}(\mathcal{X})$ and $R_{\mathcal{Y}}(y^*) := \sup_{y\in\mathsf{A}(\mathcal{X})} \|y - y^*\|_{y^*} < +\infty$.

<u>Global linear convergence of $\{\delta_k\}_{k\geq 0}$:</u>

- ▷ $\{\delta_k\}_{k\geq 0}$ is strictly decreasing (until termination).
- ▷ For all $k \geq 0$, define $k_{\mathrm{eff}} := \lceil \max\{(k - |\mathcal{S}_0| + q)/2, 0\}\rceil \approx k/2$, and then

$$\delta_k \leq (1-\rho)^{k_{\mathrm{eff}}}\delta_0, \quad \text{where} \quad \rho := \min\left\{\frac{1}{5.3(\delta_0 + \theta + B)}, \frac{\mu\Phi(\mathcal{X},\mathcal{X}^*)^2}{42.4(\theta + B)^2}\right\},$$

  where
  - μ is the quadratic-growth constant of f on $\mathcal{Y}$ that only depends on $R_{\mathcal{Y}}(y^*)$
  - $\Phi(\mathcal{X},\mathcal{X}^*) > 0$ is a geometric constant about $\mathcal{X}^*$ and $\mathcal{X}$.
- ▷ All the quantities defining ρ are *affine-invariant* and *norm-independent*.

Computational Guarantees

Global linear convergence of $\{G_k\}_{k\geq 0}$:

For some (affine-invariant) $\bar{D} < +\infty$ and all $k \geq 0$, we have

$$G_k \leq \begin{cases} 4(1-\rho)^{k_{\text{eff}}}\delta_0 \max\{\bar{D}, 1\}, & \text{if } \delta_k > 1 \\ 4\sqrt{1-\rho}^{k_{\text{eff}}}\sqrt{\delta_0}\max\{\bar{D}, 1\}, & \text{if } \delta_k \leq 1 \end{cases}.$$

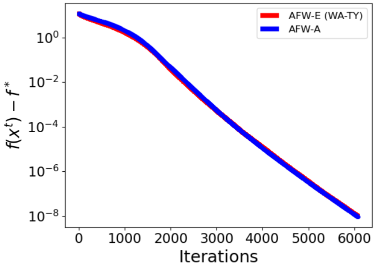
Essentially, this means $\{G_k\}_{k\geq 0}$ converges at the linear rate $\sqrt{1-\rho}$, which is worse than the rate of $\{\delta_k\}_{k\geq 0}$, namely $(1-\rho)$.

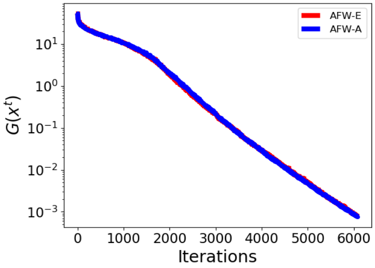
Global linear convergence of $\{G_k\}_{k\geq 0}$:

For some (affine-invariant) $\bar{D} < +\infty$ and all $k \geq 0$, we have

$$G_k \leq \begin{cases} 4(1-\rho)^{k_{\rm eff}}\delta_0 \max\{\bar{D},1\}, & \text{if } \delta_k > 1 \\ 4\sqrt{1-\rho}^{\,k_{\rm eff}}\sqrt{\delta_0}\max\{\bar{D},1\}, & \text{if } \delta_k \leq 1 \end{cases}.$$

Essentially, this means $\{G_k\}_{k\geq 0}$ converges at the linear rate $\sqrt{1-\rho}$, which is worse than the rate of $\{\delta_k\}_{k\geq 0}$, namely $(1-\rho)$.





Improved local linear rate

Improved local linear rate

▷ Let $\mathcal{X}^* \neq \emptyset$ denote the set of optimal solutions of (P)

Improved local linear rate

- Let $\mathcal{X}^* \neq \emptyset$ denote the set of optimal solutions of (P)
- There exists a face of $\mathcal{X}$, denoted by $\mathcal{F}$, such that for any $x^* \in \mathcal{X}^*$, if $x \in \mathcal{X}$, then

$$\langle \nabla F(x^*), x - x^* \rangle = 0 \quad \Longleftrightarrow \quad x \in \mathcal{F}.$$

Improved local linear rate

- Let $\mathcal{X}^* \neq \emptyset$ denote the set of optimal solutions of (P)
- There exists a face of $\mathcal{X}$, denoted by $\mathcal{F}$, such that for any $x^* \in \mathcal{X}^*$, if $x \in \mathcal{X}$, then
$$\langle \nabla F(x^*), x - x^* \rangle = 0 \quad \Longleftrightarrow \quad x \in \mathcal{F}.$$
- Define $\Delta_{\mathcal{F}} := \max_{x^* \in \mathcal{X}^*} \min_{v \in \mathcal{V} \setminus \mathcal{F}} \langle \nabla F(x^*), v - x^* \rangle > 0$.

Improved local linear rate

- Let $\mathcal{X}^* \neq \emptyset$ denote the set of optimal solutions of (P)
- There exists a face of $\mathcal{X}$, denoted by $\mathcal{F}$, such that for any $x^* \in \mathcal{X}^*$, if $x \in \mathcal{X}$, then
$$\langle \nabla F(x^*), x - x^* \rangle = 0 \quad \iff \quad x \in \mathcal{F}.$$
- Define $\Delta_{\mathcal{F}} := \max_{x^* \in \mathcal{X}^*} \min_{v \in \mathcal{V} \setminus \mathcal{F}} \langle \nabla F(x^*), v - x^* \rangle > 0$.

<u>Land on $\mathcal{F}$ in finite iterations:</u>

Let $\bar{k} \geq 0$ satisfy that
$$\delta_{\bar{k}} < \min\{V(\Delta_{\mathcal{F}}, R_{\mathcal{Y}}(y^*)), \min_{v \in \mathcal{V} \setminus \mathcal{F}} F(v) - F^*\}.$$

Improved local linear rate

- Let $\mathcal{X}^* \neq \emptyset$ denote the set of optimal solutions of (P)
- There exists a face of $\mathcal{X}$, denoted by $\mathcal{F}$, such that for any $x^* \in \mathcal{X}^*$, if $x \in \mathcal{X}$, then
$$\langle \nabla F(x^*), x - x^* \rangle = 0 \quad \Longleftrightarrow \quad x \in \mathcal{F}.$$
- Define $\Delta_{\mathcal{F}} := \max_{x^* \in \mathcal{X}^*} \min_{v \in \mathcal{V} \setminus \mathcal{F}} \langle \nabla F(x^*), v - x^* \rangle > 0.$

Land on $\mathcal{F}$ in finite iterations:

Let $\bar{k} \geq 0$ satisfy that
$$\delta_{\bar{k}} < \min\{V(\Delta_{\mathcal{F}}, R_{\mathcal{Y}}(y^*)), \min_{v \in \mathcal{V} \setminus \mathcal{F}} F(v) - F^*\}.$$
For all $k \geq \bar{k}$, if $x^k \notin \mathcal{F}$, then

- $\mathcal{S}_{k+1} \subseteq \mathcal{S}_k$, when either exact line-search or adaptive stepsize is used in Step 26,
- $\mathcal{S}_{k+1} = \mathcal{S}_k \setminus \{a^k\}$ for some $a^k \in \mathcal{S}_k \cap \bar{\mathcal{V}}_{\mathcal{F}}$, when exact line-search is used in Step 26;

otherwise, if $x^k \in \mathcal{F}$, then $x^l \in \mathcal{F}$ for all $l \geq k$.

Another Example: Positron Emission Tomography

$$\max_{x\in\Delta_n}\ \left\{F(x) := \sum_{j=1}^m p_j \ln(a_j^\top x)\right\} \tag{PET}$$

Another Example: Positron Emission Tomography

$$\max_{x\in\Delta_n} \left\{F(x) := \sum_{j=1}^{m} p_j \ln(a_j^\top x)\right\} \qquad \texttt{(PET)}$$

- Known as Positron Emission Tomography (PET) in medical imaging, but has many other applications, e.g., inference of multi-dimensional Hawkes processes [ZZS13] and log-optimal investment [Cov84].

Another Example: Positron Emission Tomography

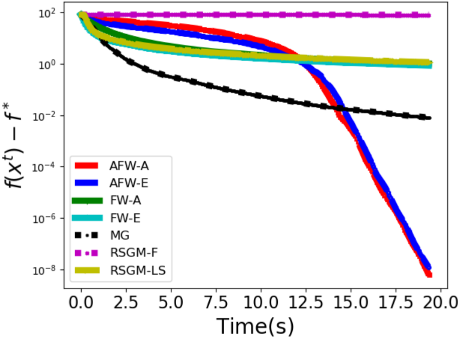
$$\max_{x \in \Delta_n} \left\{ F(x) := \sum_{j=1}^m p_j \ln(a_j^\top x) \right\} \qquad \texttt{(PET)}$$

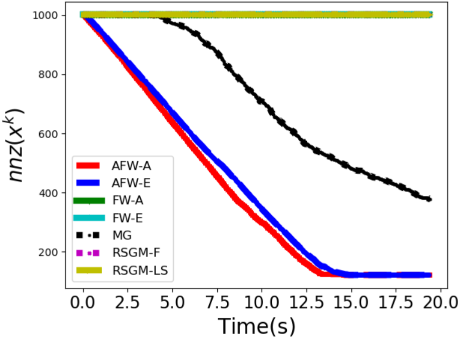
- Known as Positron Emission Tomography (PET) in medical imaging, but has many other applications, e.g., inference of multi-dimensional Hawkes processes [ZZS13] and log-optimal investment [Cov84].
- For all $j \in [m]$, let $p_j > 0$, $a_j \in \mathbb{R}_+^n$, $a_j \neq 0$ and $\sum_{j=1}^m p_j = 1$.

Another Example: Positron Emission Tomography

$$\max_{x\in\Delta_n} \left\{F(x) := \sum_{j=1}^m p_j \ln(a_j^\top x)\right\} \qquad \texttt{(PET)}$$

- ▷ Known as Positron Emission Tomography (PET) in medical imaging, but has many other applications, e.g., inference of multi-dimensional Hawkes processes [ZZS13] and log-optimal investment [Cov84].
- ▷ For all $j \in [m]$, let $p_j > 0$, $a_j \in \mathbb{R}^n_+$, $a_j \neq 0$ and $\sum_{j=1}^m p_j = 1$.





Thank you!

References

[BBT17] Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. "A Descent Lemma Beyond Lipschitz Gradient Continuity: First-Order Methods Revisited and Applications". In: *Math. Oper. Res.* 42.2 (2017), pp. 330–348.

[BS17] A. Beck and S. Shtern. "Linearly convergent away-step conditional gradient for non-strongly convex functions". In: *Math. Program.* 164 (2017), 1–27.

[Cov84] T. Cover. "An algorithm for maximizing expected log investment return". In: *IEEE Trans. Inf. Theory* 30.2 (1984), pp. 369–373.

[Dvu+23] P. Dvurechensky et al. "Generalized self-concordant analysis of Frank–Wolfe algorithms". In: *Math. Program.* 198 (2023), 255—323.

[Fed72] V. V. Fedorov. *Theory of Optimal Experiments*. Academic Press, 1972.

[Kha96] Leonid G. Khachiyan. "Rounding of Polytopes in the Real Number Model of Computation". In: *Math. Oper. Res.* 21.2 (1996), pp. 307–320.

[LFN18] Haihao. Lu, Robert M. Freund, and Yurii. Nesterov. "Relatively Smooth Convex Optimization by First-Order Methods, and Applications". In: *SIAM J. Optim.* 28.1 (2018), pp. 333–354.

[LJJ15] Simon Lacoste-Julien and Martin Jaggi. "On the Global Linear Convergence of Frank-Wolfe Optimization Variants". In: *Proc. NeurIPS*. Montreal, Canada, 2015, 496—504.

[PR19] Javier Peña and Daniel Rodríguez. "Polytope Conditioning and Linear Convergence of the Frank–Wolfe Algorithm". In: *Math. Oper. Res.* 44.1 (2019), pp. 1–18.

[STT78] S.D. Silvey, D.H. Titterington, and B. Torsney. "An algorithm for optimal designs on a design space". In: *Commun. Stat. Theory Methods* 7.14 (1978), pp. 1379–1389.

References

[ZFce] Renbo Zhao and Robert M. Freund. "Analysis of the Frank-Wolfe Method for Convex Composite Optimization involving a Logarithmically-Homogeneous Barrier". In: *Math. Program.* (accepted, 2022).

[ZZS13] Ke Zhou, Hongyuan Zha, and Le Song. "Learning Social Infectivity in Sparse Low-rank Networks Using Multi-dimensional Hawkes Processes". In: *Proc. AISTATS*. 2013, pp. 641–649.