Away-step Frank-Wolfe Method for Convex Optimization Involving a Log-Homogeneous Barrier

Renbo Zhao

MIT Operations Research Center

SIAM Conference on Optimization

Seattle, WA

June, 2023

Motivating Example: D-Optimal Design

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

Motivating Example: D-Optimal Design

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

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

Motivating Example: D-Optimal Design

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

- 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.

Motivating Example: D-Optimal Design

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

- 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

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

- 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

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

- 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):

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

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

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

Motivating Example: D-Optimal Design

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

- 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):

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

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

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

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

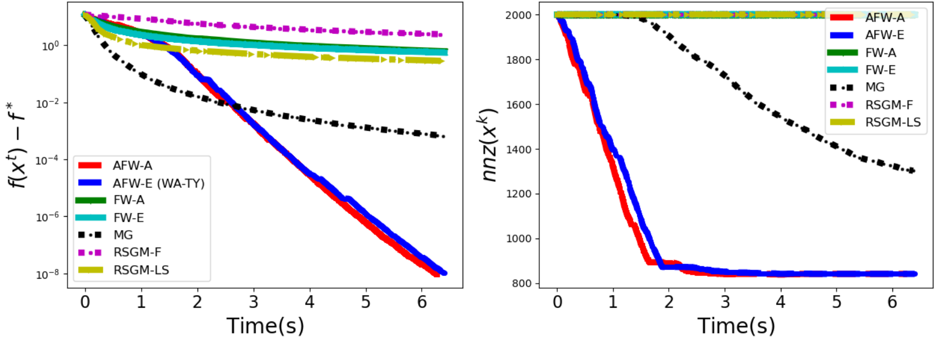
The WA-TY Method

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 [Dvu23] 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 [Dvu23] 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 [Dvu23] 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

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

Problem of Interest

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

▷ $\mathbb{X}$ 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

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

- $\mathbb{X}$ 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

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

- $\mathbb{X}$ 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

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

- $\mathbb{X}$ 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

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

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

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

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

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

- Two prototypical examples:

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

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

- Two prototypical examples:
  - $f(Y) = -\ln\det(Y)$ for $\mathcal{K} := \mathbb{S}^n_+$ and $\theta = n$,

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

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

- Two prototypical examples:
  - $f(Y) = -\ln\det(Y)$ for $\mathcal{K} := \mathbb{S}^n_+$ and $\theta = n$,
  - $f(y) = -\sum_{j=1}^m w_j \ln(y_j)$ for $\mathcal{K} := \mathbb{R}^m_+$ and $\theta = \sum_{j=1}^m w_j$ (where $w_1, \ldots, w_n \geq 1$).

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

- Let $\mathcal{K} \subsetneq \mathbb{Y}$ be a regular cone, i.e., $\mathcal{K}$ is closed, convex, pointed and has nonempty interior.
- Two prototypical examples:
  - $f(Y) = -\ln\det(Y)$ for $\mathcal{K} := \mathbb{S}_+^n$ and $\theta = n$,
  - $f(y) = -\sum_{j=1}^m w_j \ln(y_j)$ for $\mathcal{K} := \mathbb{R}_+^m$ and $\theta = \sum_{j=1}^m w_j$ (where $w_1, \ldots, w_n \geq 1$).
- f is a θ-LHSCB on $\mathcal{K}$ with *complexity parameter* $\theta \geq 1$, if f is three-times continuously differentiable and non-degenerate on $\mathsf{int}\,\mathcal{K}$, and satisfies

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

- Let $\mathcal{K} \subsetneq \mathbb{Y}$ be a regular cone, i.e., $\mathcal{K}$ is closed, convex, pointed and has nonempty interior.
- Two prototypical examples:
  - $f(Y) = -\ln\det(Y)$ for $\mathcal{K} := \mathbb{S}^n_+$ and $\theta = n$,
  - $f(y) = -\sum_{j=1}^m w_j \ln(y_j)$ for $\mathcal{K} := \mathbb{R}^m_+$ and $\theta = \sum_{j=1}^m w_j$ (where $w_1, \ldots, w_n \geq 1$).
- f is a θ-LHSCB on $\mathcal{K}$ with *complexity parameter* $\theta \geq 1$, if f is three-times continuously differentiable and non-degenerate on $\mathsf{int}\,\mathcal{K}$, and satisfies

  1. $\left|D^3 f(y)[w,w,w]\right| \leq 2\|w\|_y^3 \quad \forall\, y \in \mathsf{int}\,\mathcal{K}, \forall\, w \in \mathbb{Y}$,
  2. $f(y_k) \to +\infty$ for any $\{y_k\}_{k\geq 1} \subseteq \mathsf{int}\,\mathcal{K}$ such that $y_k \to u \in \mathsf{bd}\,\mathcal{K}$,
  3. $f(ty) = f(y) - \theta\ln(t) \quad \forall\, y \in \mathsf{int}\,\mathcal{K}, \forall\, t > 0$.

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

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 $x^0 = \sum_{v \in \mathcal{V}} \beta^0_v v$, $\mathcal{S}_0 := \mathrm{supp}(\beta^0)$.

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 $x^0 = \sum_{v\in\mathcal{V}} \beta_v^0 v$, $\mathcal{S}_0 := \mathrm{supp}(\beta^0)$.
- **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-step Frank-Wolfe Method for solving (P)

- **Input**: $x^0 \in \mathcal{X} \cap \operatorname{\mathsf{dom}} F$, $\beta^0 \in \Delta_{|\mathcal{V}|}$ such that $x^0 = \sum_{v \in \mathcal{V}} \beta^0_v v$, $\mathcal{S}_0 := \operatorname{supp}(\beta^0)$.
- **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 direction) If $|\mathcal{S}_k| > 1$, compute $a^k \in \arg\max_{x \in \mathcal{S}_k} \langle \nabla F(x^k), x \rangle$, $d^k_{\mathrm{A}} := x^k - a^k$ and $\widetilde{G}_k := \langle -\nabla F(x^k), d^k_{\mathrm{A}} \rangle$.

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** $k \geq 0$:
  - (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 $\widetilde{G}_k := \langle -\nabla F(x^k), d_{\mathrm{A}}^k\rangle$.
  - (Choose direction) If $|\mathcal{S}_k| = 1$ or $G_k > \widetilde{G}_k$, 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 $\widetilde{G}_k := \langle -\nabla F(x^k), d_{\mathrm{A}}^k\rangle$.
  - (Choose direction) If $|\mathcal{S}_k| = 1$ or $G_k > \widetilde{G}_k$, 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** $k \geq 0$:
  - (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 $\widetilde{G}_k := \langle -\nabla F(x^k), d_{\mathrm{A}}^k\rangle$.
  - (Choose direction) If $|\mathcal{S}_k| = 1$ or $G_k > \widetilde{G}_k$, 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:
$$G_k \ge [\delta_k := F(x^k) - F^*] \quad \text{for } k \ge 0.$$
- 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].

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].
- 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

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

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

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

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

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 $R_{\mathcal{Y}}(y^*) := \sup_{y\in\mathsf{A}(\mathcal{X})} \|y - y^*\|_{y^*} < +\infty$.

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 $\{\delta_k\}_{k\geq 0}$:

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 $R_{\mathcal{Y}}(y^*) := \sup_{y\in\mathsf{A}(\mathcal{X})}\|y - y^*\|_{y^*} < +\infty$.

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

- $\{\delta_k\}_{k\ge 0}$ 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 $q := \min\{|\mathcal{W}| :\ \mathcal{W} \subseteq \mathcal{V}$ 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

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

- 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 $\mathcal{X}$.
- 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 $(1-\rho)$.

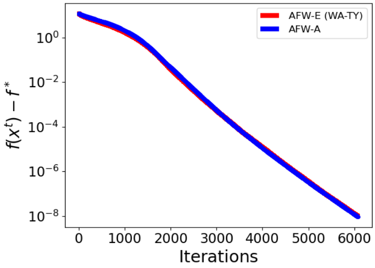
Computational Guarantees

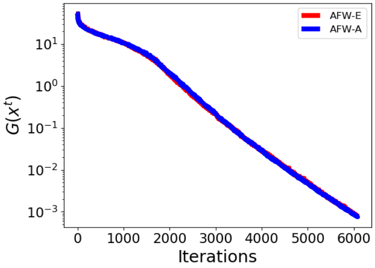
Global linear convergence of $\{G_k\}_{k\geq 0}$:

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

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

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
$$\langle \nabla F(x^*), x - x^* \rangle = 0 \quad \Longleftrightarrow \quad x \in \mathcal{F}.$$
- 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
$$\delta_{\bar{k}} < \min\{V(\Delta_{\mathcal{F}}, R_{\mathcal{Y}}(y^*)), \min_{v \in \mathcal{V} \setminus \mathcal{F}} F(v) - F^*\}.$$

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 $\Delta_{\mathcal{F}} := \max_{x^* \in \mathcal{X}^*} \min_{v \in \mathcal{V} \setminus \mathcal{F}} \langle \nabla F(x^*), v - x^* \rangle > 0$.

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

Let $\bar{k} \geq 0$ satisfy that
$$\delta_{\bar{k}} < \min\{V(\Delta_{\mathcal{F}}, R_{\mathcal{Y}}(y^*)), \min_{v \in \mathcal{V} \setminus \mathcal{F}} F(v) - F^*\}.$$
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 7,
- $\mathcal{S}_{k+1} = \mathcal{S}_k \setminus \{a^k\}$ for some $a^k \in \mathcal{S}_k \cap \bar{\mathcal{V}}_{\mathcal{F}}$, when exact line-search is used in Step 7;

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

Another Example: Positron Emission Tomography

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

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

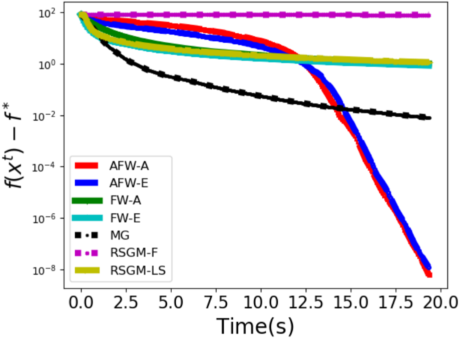
$$\max_{x\in\Delta_n} \left\{F(x) := \textstyle\sum_{j=1}^m p_j \ln(a_j^\top x)\right\} \qquad \texttt{(PET)}$$

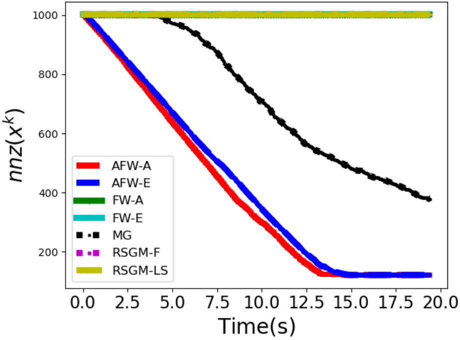
- Known as Positron Emission Tomography (PET) in medical imaging, but has many other applications, e.g., inference of multi-dimensional Hawkes processes [ZZS13] and log-optimal investment [Cov84].
- For all $j \in [m]$, let $p_j > 0$, $a_j \in \mathbb{R}_+^n$, $a_j \neq 0$ and $\sum_{j=1}^m p_j = 1$.

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.

[Dvu23] 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.