Weierstrass Institute for
Applied Analysis and Stochastics

Leibniz Association

Some results on minimization involving self-concordant functions and barriers

Pavel Dvurechensky

Based on joint works with Yurii Nesterov (UCLouvain), Petr Ostroukhov (MBZUAI), Kamil Safin (MIPT), Shimrit Shtern (Technion), Mathias Staudigl (Mannheim University)

ALGOPT2024 workshop on Algorithmic Optimization: Tools for AI and Data Science

Happy 50th anniversary of research career in Optimization, Prof. Yurii Nesterov!

Content

1 Barrier algorithms for non-convex optimization

- Problem statement
- Self-concordant barriers
- Approximate optimality conditions
- First-order algorithm
- Second-order algorithm

2 Minimizing self-concordant functions

E – finite dimensional vector space with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$.

We consider the problem:

$$\min_x f(x) \quad \text{s.t.: } \mathbf{A}x = b, \; x \in \bar{\mathsf{K}}. \tag{P}$$

E – finite dimensional vector space with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$.

We consider the problem:

$$\min_x f(x) \quad \text{s.t.: } \mathbf{A}x = b, \; x \in \bar{\mathsf{K}}. \tag{P}$$

Denote: $\mathsf{L} = \{x \in \mathsf{E} | \mathbf{A}x = b\}$, $\bar{\mathsf{X}} = \bar{\mathsf{K}} \cap \mathsf{L}$, $\mathsf{X} = \mathsf{K} \cap \mathsf{L}$.

E – finite dimensional vector space with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$.

We consider the problem:

$$\min_x f(x) \quad \text{s.t.: } \mathbf{A}x = b,\ x \in \bar{\mathsf{K}}. \tag{P}$$

Denote: $\mathsf{L} = \{x \in \mathsf{E} | \mathbf{A}x = b\}$, $\bar{\mathsf{X}} = \bar{\mathsf{K}} \cap \mathsf{L}$, $\mathsf{X} = \mathsf{K} \cap \mathsf{L}$.

Assumptions:

1. $f : \mathsf{E} \to \mathbb{R}$ is possibly non-convex, continuous on $\bar{\mathsf{X}}$ and continuously differentiable on X;

E – finite dimensional vector space with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$.

We consider the problem:

$$\min_x f(x) \quad \text{s.t.: } \mathbf{A}x = b,\ x \in \bar{\mathsf{K}}. \tag{P}$$

Denote: $\mathsf{L} = \{x \in \mathsf{E} | \mathbf{A}x = b\}$, $\bar{\mathsf{X}} = \bar{\mathsf{K}} \cap \mathsf{L}$, $\mathsf{X} = \mathsf{K} \cap \mathsf{L}$.

Assumptions:

1. $f : \mathsf{E} \to \mathbb{R}$ is possibly non-convex, continuous on $\bar{\mathsf{X}}$ and continuously differentiable on X;
2. $\bar{\mathsf{K}} \subset \mathsf{E}$ is closed convex either set or pointed one (i.e., $\bar{\mathsf{K}} \cap (-\bar{\mathsf{K}}) = \{0\}$);

E – finite dimensional vector space with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$.

We consider the problem:

$$\min_{x} f(x) \quad \text{s.t.: } \mathbf{A}x = b,\ x \in \bar{\mathsf{K}}. \tag{P}$$

Denote: $\mathsf{L} = \{x \in \mathsf{E} | \mathbf{A}x = b\}$, $\bar{\mathsf{X}} = \bar{\mathsf{K}} \cap \mathsf{L}$, $\mathsf{X} = \mathsf{K} \cap \mathsf{L}$.

Assumptions:

1. $f : \mathsf{E} \to \mathbb{R}$ is possibly non-convex, continuous on $\bar{\mathsf{X}}$ and continuously differentiable on X;
2. $\bar{\mathsf{K}} \subset \mathsf{E}$ is closed convex either set or pointed one (i.e., $\bar{\mathsf{K}} \cap (-\bar{\mathsf{K}}) = \{0\}$);
3. Linear operator $\mathbf{A} : \mathsf{E} \to \mathbb{R}^m$ has full rank, i.e., $\mathrm{im}(\mathbf{A}) = \mathbb{R}^m$, $b \in \mathbb{R}^m$;
4. Problem (P) admits a global solution. We let $f_{\min}(\mathsf{X}) = \min\{f(x) | x \in \bar{\mathsf{X}}\}$.

- Unconstrained or “Projection”-based, treating $\bar{X}$ as a simple set.
  [Nesterov, Polyak, ’06], [Agarwal et al., ’17], [Carmon et al., ’17], [Cartis, Gould, Toint, ’12, ’18, ’19], [Ghadimi, Lan, ’16], [Birgin, Martinez, ’18], [Curtis et al., ’17].

- Unconstrained or “Projection”-based, treating $\bar{X}$ as a simple set.
  [Nesterov, Polyak, ’06], [Agarwal et al., ’17], [Carmon et al., ’17], [Cartis, Gould, Toint, ’12, ’18, ’19], [Ghadimi, Lan, ’16], [Birgin, Martinez, ’18], [Curtis et al., ’17].
- Augmented Lagrangian algorithms.
  [Bolte et al., ’18], [Andreani et al., ’19, ’21], [Birgin, Martinez, ’20], [Grapiglia, Yuan, ’20], [Khanh, Mordukhovich, Tran, ’23].

- Unconstrained or "Projection"-based, treating $\bar{X}$ as a simple set.
  [Nesterov, Polyak, '06], [Agarwal et al., '17], [Carmon et al., '17], [Cartis, Gould, Toint, '12, '18, '19], [Ghadimi, Lan, '16], [Birgin, Martinez, '18], [Curtis et al., '17].
- Augmented Lagrangian algorithms.
  [Bolte et al., '18], [Andreani et al., '19, '21], [Birgin, Martinez, '20], [Grapiglia, Yuan, '20], [Khanh, Mordukhovich, Tran, '23].
- Barrier methods for non-negative orthant and/or quadratic programming
  [Ye, '92], [Faybusovich, Lu, '06], [Lu, Yuan, '07], [Tseng et al., '11], [Bian et al., '15], [Bomze et al., '19], **[Haeser, Liu, Ye, '19], [O'Neill, Wright, '20]**.

- Unconstrained or “Projection”-based, treating $\bar{X}$ as a simple set.
  [Nesterov, Polyak, ’06], [Agarwal et al., ’17], [Carmon et al., ’17], [Cartis, Gould, Toint, ’12, ’18, ’19], [Ghadimi, Lan, ’16], [Birgin, Martinez, ’18], [Curtis et al., ’17].
- Augmented Lagrangian algorithms.
  [Bolte et al., ’18], [Andreani et al., ’19, ’21], [Birgin, Martinez, ’20], [Grapiglia, Yuan, ’20], [Khanh, Mordukhovich, Tran, ’23].
- Barrier methods for non-negative orthant and/or quadratic programming
  [Ye, ’92], [Faybusovich, Lu, ’06], [Lu, Yuan, ’07], [Tseng et al., ’11], [Bian et al., ’15], [Bomze et al., ’19], **[Haeser, Liu, Ye, ’19], [O’Neill, Wright, ’20]**.

Our goals:

- Feasible iterates $\Rightarrow$ Interior-point algorithms.
- General sets or cones $\Rightarrow$ (Logarithmically homogeneous) self-concordant barriers.
- Favorable global complexity guarantees $\Rightarrow$ Quadratic/cubic regularization.

Content

A function $h : \bar{\mathsf{K}} \to (-\infty, \infty]$ with $\operatorname{dom} h = \mathsf{K}$ is called a ν-self-concordant barrier (SCB) [Nesterov, Nemirovski, 1994] for the set $\bar{\mathsf{K}}$ if:

(a) h is a standard *self-concordant function*:

$$|D^3 h(x)[u, u, u]| \leq 2 D^2 h(x)[u, u]^{3/2};$$

A function $h : \bar{\mathsf{K}} \to (-\infty, \infty]$ with $\operatorname{dom} h = \mathsf{K}$ is called a ν-self-concordant barrier (SCB) [Nesterov, Nemirovski, 1994] for the set $\bar{\mathsf{K}}$ if:

(a) h is a standard *self-concordant function*:

$$|D^3 h(x)[u,u,u]| \leq 2 D^2 h(x)[u,u]^{3/2};$$

(b) h is a ν*-self-concordant barrier* for $\bar{\mathsf{K}}$:

$$\sup_{u \in \mathbb{R}^n} \{2Dh(x)[u] - D^2 h(x)[u,u]\} \leq \nu; \quad (\langle \nabla h(x), (\nabla^2 h(x))^{-1} \nabla h(x) \rangle \leq \nu)$$

A function $h : \bar{\mathsf{K}} \to (-\infty, \infty]$ with $\operatorname{dom} h = \mathsf{K}$ is called a ν-self-concordant barrier (SCB) [Nesterov, Nemirovski, 1994] for the set $\bar{\mathsf{K}}$ if:

(a) h is a standard *self-concordant function*:

$$|D^3 h(x)[u,u,u]| \leq 2D^2 h(x)[u,u]^{3/2};$$

(b) h is a ν*-self-concordant barrier* for $\bar{\mathsf{K}}$:

$$\sup_{u\in\mathbb{R}^n} \{2Dh(x)[u] - D^2 h(x)[u,u]\} \leq \nu; \quad (\langle \nabla h(x), (\nabla^2 h(x))^{-1} \nabla h(x) \rangle \leq \nu)$$

If additionally $\bar{\mathsf{K}}$ is a regular cone: closed convex, solid, contains no lines, $\mathsf{K} \neq \emptyset$ and

(c) h is *logarithmically homogeneous:*

$$h(tx) = h(x) - \nu \ln(t) \qquad \forall x \in \mathsf{K}, t > 0.$$

Then h is called a logarithmically homogeneous ν-self-concordant barrier (LHSCB).

A function $h : \bar{\mathsf{K}} \to (-\infty, \infty]$ with $\operatorname{dom} h = \mathsf{K}$ is called a ν-self-concordant barrier (SCB) [Nesterov, Nemirovski, 1994] for the set $\bar{\mathsf{K}}$ if:

(a) h is a standard *self-concordant function*:

$$|D^3 h(x)[u,u,u]| \leq 2 D^2 h(x)[u,u]^{3/2};$$

(b) h is a ν*-self-concordant barrier* for $\bar{\mathsf{K}}$:

$$\sup_{u \in \mathbb{R}^n} \{2Dh(x)[u] - D^2 h(x)[u,u]\} \leq \nu; \quad (\langle \nabla h(x), (\nabla^2 h(x))^{-1} \nabla h(x) \rangle \leq \nu)$$

If additionally $\bar{\mathsf{K}}$ is a regular cone: closed convex, solid, contains no lines, $\mathsf{K} \neq \emptyset$ and

(c) h is *logarithmically homogeneous:*

$$h(tx) = h(x) - \nu \ln(t) \qquad \forall x \in \mathsf{K}, t > 0.$$

Then h is called a logarithmically homogeneous ν-self-concordant barrier (LHSCB).

Example: $h(x) = -\ln(x)$. Indeed $|-2/x^3| \leq 2(1/x^2)^{3/2}$,
$-1/x \cdot (1/x^2)^{-1}(-1/x) = 1$, $-\ln(tx) = -\ln(x) - \ln t$.

The Hessian $H(x) \triangleq \nabla^2 h(x)$: E $\to$ E* gives rise to a local norm and its dual

$$\|u\|_x \triangleq \langle H(x)u, u\rangle^{1/2}, \qquad \|s\|_x^* \triangleq \langle [H(x)]^{-1}s, s\rangle^{1/2}. \tag{1}$$

The Hessian $H(x) \triangleq \nabla^2 h(x)$: E $\to$ E* gives rise to a local norm and its dual

$$\|u\|_x \triangleq \langle H(x)u, u\rangle^{1/2}, \qquad \|s\|_x^* \triangleq \langle [H(x)]^{-1}s, s\rangle^{1/2}. \tag{1}$$

Let $d \in$ E. For all $t \in [0, \frac{1}{\|d\|_x})$, we have

$$x + td \in \mathsf{K} \tag{2}$$

$$h(x + td) \leq h(x) + t\langle \nabla h(x), d\rangle + t^2\|d\|_x^2 \omega(t\|d\|_x), \tag{3}$$

where $\omega(t) \triangleq \frac{-t-\ln(1-t)}{t^2}$, $t \in [0, 1)$.

Content

1. **Barrier algorithms for non-convex optimization**
   - Problem statement
   - Self-concordant barriers
   - Approximate optimality conditions
   - First-order algorithm
   - Second-order algorithm
2. **Minimizing self-concordant functions**
   - Unconstrained minimization by path-following methods
   - Composite minimization by gradient regularization of Newton method
   - Projection-free constrained minimization of self-concordant functions

Definition 1

Given $\varepsilon \geq 0$, a point $\bar{x} \in \mathsf{E}$ is an ε-KKT point for problem (P) if there exists $\bar{y} \in \mathbb{R}^m$ such that $\mathbf{A}\bar{x} = b, \bar{x} \in \mathsf{K}$

Definition 1

Given $\varepsilon \geq 0$, a point $\bar{x} \in \mathsf{E}$ is an ε-KKT point for problem (P) if there exists $\bar{y} \in \mathbb{R}^m$ such that $\mathbf{A}\bar{x} = b, \bar{x} \in \mathsf{K}$ and

- Option A: $\bar{\mathsf{K}}$ be a convex set: $\langle \nabla f(\bar{x}) - \mathbf{A}^* \bar{y}, x - \bar{x} \rangle \geq -\varepsilon \qquad \forall x \in \bar{\mathsf{K}}$.

Definition 1

Given $\varepsilon \geq 0$, a point $\bar{x} \in \mathsf{E}$ is an ε-KKT point for problem (P) if there exists $\bar{y} \in \mathbb{R}^m$ such that $\mathbf{A}\bar{x} = b, \bar{x} \in \mathsf{K}$ and

- Option A: $\bar{\mathsf{K}}$ be a convex set: $\langle \nabla f(\bar{x}) - \mathbf{A}^* \bar{y}, x - \bar{x} \rangle \geq -\varepsilon \qquad \forall x \in \bar{\mathsf{K}}$.
- Option B: $\bar{\mathsf{K}}$ be a convex cone:

$$\bar{s} = \nabla f(\bar{x}) - \mathbf{A}^* \bar{y} \in \bar{\mathsf{K}}^*,$$

$$(0 \leq) \langle \bar{s}, \bar{x} \rangle \leq \varepsilon.$$

Definition 1

Given $\varepsilon \geq 0$, a point $\bar{x} \in \mathsf{E}$ is an ε-KKT point for problem (P) if there exists $\bar{y} \in \mathbb{R}^m$ such that $\mathbf{A}\bar{x} = b, \bar{x} \in \mathsf{K}$ and

- Option A: $\bar{\mathsf{K}}$ be a convex set: $\langle \nabla f(\bar{x}) - \mathbf{A}^*\bar{y}, x - \bar{x} \rangle \geq -\varepsilon \qquad \forall x \in \bar{\mathsf{K}}$.
- Option B: $\bar{\mathsf{K}}$ be a convex cone:

$$\bar{s} = \nabla f(\bar{x}) - \mathbf{A}^*\bar{y} \in \bar{\mathsf{K}}^*,$$
$$(0 \leq)\langle \bar{s}, \bar{x} \rangle \leq \varepsilon.$$

Motivation: ε-perturbation of the standard first-order stationarity condition $\langle \nabla f(\bar{x}) - \mathbf{A}^*\bar{y}, x - \bar{x} \rangle \geq 0, \quad \forall x \in \bar{\mathsf{K}}$.

Definition 2

Given $\varepsilon_1, \varepsilon_2 \geq 0$, a point $\bar{x} \in \mathsf{E}$ is an $(\varepsilon_1, \varepsilon_2)$-2KKT point for problem (P) if there exists $\bar{y} \in \mathbb{R}^m$ such that $\mathbf{A}\bar{x} = b, \bar{x} \in \mathsf{K}$ and

- - Option A: $\bar{\mathsf{K}}$ be a convex set: $\langle \nabla f(\bar{x}) - \mathbf{A}^*\bar{y}, x - \bar{x} \rangle \geq -\varepsilon_1 \qquad \forall x \in \bar{\mathsf{K}}$.
  - Option B: $\bar{\mathsf{K}}$ be a convex cone:

$$\bar{s} = \nabla f(\bar{x}) - \mathbf{A}^*\bar{y} \in \bar{\mathsf{K}}^*,$$

$$(0 \leq)\langle \bar{s}, \bar{x} \rangle \leq \varepsilon_1.$$

Definition 2

Given $\varepsilon_1, \varepsilon_2 \geq 0$, a point $\bar{x} \in \mathsf{E}$ is an $(\varepsilon_1, \varepsilon_2)$-2KKT point for problem (P) if there exists $\bar{y} \in \mathbb{R}^m$ such that $\mathbf{A}\bar{x} = b, \bar{x} \in \mathsf{K}$ and

- - Option A: $\bar{\mathsf{K}}$ be a convex set: $\langle \nabla f(\bar{x}) - \mathbf{A}^* \bar{y}, x - \bar{x} \rangle \geq -\varepsilon_1 \qquad \forall x \in \bar{\mathsf{K}}$.
  - Option B: $\bar{\mathsf{K}}$ be a convex cone:

$$\bar{s} = \nabla f(\bar{x}) - \mathbf{A}^* \bar{y} \in \bar{\mathsf{K}}^*,$$

$$(0 \leq) \langle \bar{s}, \bar{x} \rangle \leq \varepsilon_1.$$

- $\nabla^2 f(\bar{x}) + \sqrt{\varepsilon_2} H(\bar{x}) \succeq 0$ on $\mathsf{L}_0 = \{v \in \mathsf{E} | \mathbf{A} v = 0\}$.

Content

1 Barrier algorithms for non-convex optimization

- Problem statement
- Self-concordant barriers
- Approximate optimality conditions
- First-order algorithm
- Second-order algorithm

2 Minimizing self-concordant functions

- Unconstrained minimization by path-following methods
- Composite minimization by gradient regularization of Newton method
- Projection-free constrained minimization of self-concordant functions

Potential function:

$$F_\mu(x) \triangleq f(x) + \mu h(x) \qquad \forall x \in \mathsf{K}, \mu > 0. \tag{4}$$

Potential function:

$$F_\mu(x) \triangleq f(x) + \mu h(x) \qquad \forall x \in \mathsf{K}, \mu > 0. \tag{4}$$

Define the set of feasible directions $\mathcal{T}_x \triangleq \{v \in \mathsf{E} | \mathbf{A}v = 0, \|v\|_x < 1\}$.

Local smoothness assumption

$f : \mathsf{E} \to \mathbb{R} \cup \{+\infty\}$ is continuously differentiable on X and there exists a constant $M > 0$ such that for all $x \in \mathsf{X}$ and $v \in \mathcal{T}_x$ we have

$$f(x+v) - f(x) - \langle \nabla f(x), v \rangle \leq \frac{M}{2} \|v\|_x^2. \tag{5}$$

Potential function:

$$F_\mu(x) \triangleq f(x) + \mu h(x) \qquad \forall x \in \mathsf{K}, \mu > 0. \tag{4}$$

Define the set of feasible directions $\mathcal{T}_x \triangleq \{v \in \mathsf{E} | \mathbf{A}v = 0, \|v\|_x < 1\}$.

Local smoothness assumption

$f : \mathsf{E} \to \mathbb{R} \cup \{+\infty\}$ is continuously differentiable on X and there exists a constant $M > 0$ such that for all $x \in \mathsf{X}$ and $v \in \mathcal{T}_x$ we have

$$f(x+v) - f(x) - \langle \nabla f(x), v \rangle \leq \frac{M}{2} \|v\|_x^2. \tag{5}$$

If the set X is bounded, we have $\lambda_{\min}(H(x)) \geq \sigma$ for some $\sigma > 0$. If f has a M-Lipschitz continuous gradient, then our assumption holds. Indeed,

$$f(x+v) - f(x) - \langle \nabla f(x), v \rangle \leq \frac{M}{2} \|v\|^2 \leq \frac{M}{2\sigma} \|v\|_x^2.$$

Step direction: $$v_\mu(x) \triangleq \operatorname*{argmin}_{v \in \mathsf{E}: \mathbf{A}v = 0} \{F_\mu(x) + \langle \nabla F_\mu(x), v\rangle + \frac{1}{2}\|v\|_x^2\}. \tag{6}$$

Step direction: $v_\mu(x) \triangleq \underset{v \in \mathsf{E}:\mathbf{A}v=0}{\operatorname{argmin}} \{F_\mu(x) + \langle \nabla F_\mu(x), v\rangle + \frac{1}{2}\|v\|_x^2\}.$ (6)

Optimality conditions ($y_\mu(x)$ is a Lagrange multiplier):

$$\nabla F_\mu(x) + H(x)v_\mu(x) - \mathbf{A}^* y_\mu(x) = 0, \tag{7}$$

$$-\mathbf{A}v_\mu(x) = 0. \tag{8}$$

Step direction: $v_\mu(x) \triangleq \underset{v \in \mathsf{E}:\mathbf{A}v=0}{\operatorname{argmin}} \{F_\mu(x) + \langle \nabla F_\mu(x), v\rangle + \frac{1}{2}\|v\|_x^2\}.$ (6)

Optimality conditions ($y_\mu(x)$ is a Lagrange multiplier):

$$\nabla F_\mu(x) + H(x)v_\mu(x) - \mathbf{A}^* y_\mu(x) = 0, \tag{7}$$

$$-\mathbf{A}v_\mu(x) = 0. \tag{8}$$

Parameterized arcs $x^+(t) \triangleq x + tv_\mu(x) \in \mathsf{X}$ for $t \in I_{x,\mu} \triangleq [0, \frac{1}{\|v_\mu(x)\|_x})$

Step direction: $v_\mu(x) \triangleq \underset{v \in \mathsf{E}:\mathbf{A}v=0}{\operatorname{argmin}} \{F_\mu(x) + \langle \nabla F_\mu(x), v\rangle + \frac{1}{2}\|v\|_x^2\}.$ (6)

Optimality conditions ($y_\mu(x)$ is a Lagrange multiplier):

$$\nabla F_\mu(x) + H(x)v_\mu(x) - \mathbf{A}^* y_\mu(x) = 0, \tag{7}$$

$$-\mathbf{A}v_\mu(x) = 0. \tag{8}$$

Parameterized arcs $x^+(t) \triangleq x + tv_\mu(x) \in \mathsf{X}$ for $t \in I_{x,\mu} \triangleq [0, \frac{1}{\|v_\mu(x)\|_x})$
If $t\|v_\mu(x)\|_x \leq 1/2$:

$$F_\mu(x^+(t)) - F_\mu(x) \leq -t\|v_\mu(x)\|_x^2 \left(1 - \frac{M+2\mu}{2}t\right) \triangleq -\eta_x(t). \tag{9}$$

$$\text{Step direction: } v_\mu(x) \triangleq \operatorname*{argmin}_{v \in \mathsf{E}: \mathbf{A}v = 0} \{F_\mu(x) + \langle \nabla F_\mu(x), v\rangle + \frac{1}{2}\|v\|_x^2\}. \tag{6}$$

Optimality conditions ($y_\mu(x)$ is a Lagrange multiplier):

$$\nabla F_\mu(x) + H(x)v_\mu(x) - \mathbf{A}^* y_\mu(x) = 0, \tag{7}$$

$$-\mathbf{A}v_\mu(x) = 0. \tag{8}$$

Parameterized arcs $x^+(t) \triangleq x + tv_\mu(x) \in \mathsf{X}$ for $t \in I_{x,\mu} \triangleq [0, \frac{1}{\|v_\mu(x)\|_x})$

If $t\|v_\mu(x)\|_x \leq 1/2$:

$$F_\mu(x^+(t)) - F_\mu(x) \leq -t\|v_\mu(x)\|_x^2 \left(1 - \frac{M+2\mu}{2}t\right) \triangleq -\eta_x(t). \tag{9}$$

Minimizing w.r.t. $t \in [0, \frac{1}{2\|v_\mu(x)\|_x}]$, we obtain stepsize:

$$\mathtt{t}_{\mu,M}(x) \triangleq \frac{1}{\max\{M + 2\mu, 2\|v_\mu(x)\|_x\}} = \min\left\{\frac{1}{M+2\mu}, \frac{1}{2\|v_\mu(x)\|_x}\right\}.$$

Result: Point x^k, dual variables y^k, $s^k = \nabla f(x^k) - \mathbf{A}^* y^k$.

repeat

Set $i_k = 0$. Find $v^k \triangleq v_\mu(x^k)$ and $y^k \triangleq y_\mu(x^k)$ from $\min_{v \in \mathsf{E}: \mathbf{A}v=0} \{F_\mu(x^k) + \langle \nabla F_\mu(x^k), v\rangle + \frac{1}{2}\|v\|^2_{x^k}\}$.

repeat

Set $\alpha_k \triangleq \min\left\{\frac{1}{2^{i_k} L_k + 2\mu}, \frac{1}{2\|v^k\|_{x^k}}\right\}$;

Set $z^k = x^k + \alpha_k v^k$, $i_k = i_k + 1$;

until

$$f(z^k) \leq f(x^k) + \langle \nabla f(x^k), z^k - x^k\rangle + 2^{i_k - 1} L_k \|z^k - x^k\|^2_{x^k}. \tag{10}$$

;

Set $L_{k+1} = 2^{i_k - 1} L_k$, $x^{k+1} = z^k$, $k = k + 1$;

until $\|v^k\|_{x^k} < \frac{\varepsilon}{3\nu}$;

Complexity theorem for FAHBA [D., Staudigl, 2021, 2024]

Let our assumptions hold. Set h - SCB if $\bar{\mathsf{K}}$ is a convex set or h - LHSCB if $\bar{\mathsf{K}}$ is a convex cone.

Complexity theorem for FAHBA [D., Staudigl, 2021, 2024]

Let our assumptions hold. Set h - SCB if $\bar{\mathsf{K}}$ is a convex set or h - LHSCB if $\bar{\mathsf{K}}$ is a convex cone. Fix $\varepsilon > 0$, some initial guess $L_0 > 0$ for the Lip. const. in (5),

Complexity theorem for FAHBA [D., Staudigl, 2021, 2024]

Let our assumptions hold. Set h - SCB if $\bar{\mathsf{K}}$ is a convex set or h - LHSCB if $\bar{\mathsf{K}}$ is a convex cone. Fix $\varepsilon > 0$, some initial guess $L_0 > 0$ for the Lip. const. in (5), the regularization parameter $\mu = \frac{\varepsilon}{\nu}$,

Complexity theorem for FAHBA [D., Staudigl, 2021, 2024]

Let our assumptions hold. Set h - SCB if $\bar{\mathsf{K}}$ is a convex set or h - LHSCB if $\bar{\mathsf{K}}$ is a convex cone. Fix $\varepsilon > 0$, some initial guess $L_0 > 0$ for the Lip. const. in (5), the regularization parameter $\mu = \frac{\varepsilon}{\nu}$,
and x^0 to be a ν-analytic center: $h(x) \geq h(x^0) - \nu \qquad \forall x \in \mathsf{X}$.

Complexity theorem for FAHBA [D., Staudigl, 2021, 2024]

Let our assumptions hold. Set h - SCB if $\bar{\mathsf{K}}$ is a convex set or h - LHSCB if $\bar{\mathsf{K}}$ is a convex cone. Fix $\varepsilon > 0$, some initial guess $L_0 > 0$ for the Lip. const. in (5), the regularization parameter $\mu = \frac{\varepsilon}{\nu}$,
and x^0 to be a ν-analytic center: $h(x) \geq h(x^0) - \nu \qquad \forall x \in \mathsf{X}$.
Let $(x^k)_{k \geq 0}$ be the trajectory generated by FAHBA.
Then the algorithm stops in no more than

$$K_I(\varepsilon, x^0) = \left\lceil 40(f(x^0) - f_{\min}(\mathsf{X}) + \varepsilon) \frac{\nu^2(\max\{M, L_0\} + \varepsilon/\nu)}{\varepsilon^2} \right\rceil = O\left(\frac{1}{\varepsilon^2}\right)$$

outer iterations, and the number of inner iterations is no more than
$2(K_I(\varepsilon, x^0) + 1) + \max\{\log_2(M/L_0), 0\}$.

Complexity theorem for FAHBA [D., Staudigl, 2021, 2024]

Let our assumptions hold. Set h - SCB if $\bar{\mathsf{K}}$ is a convex set or h - LHSCB if $\bar{\mathsf{K}}$ is a convex cone. Fix $\varepsilon > 0$, some initial guess $L_0 > 0$ for the Lip. const. in (5), the regularization parameter $\mu = \frac{\varepsilon}{\nu}$,
and x^0 to be a ν-analytic center: $h(x) \geq h(x^0) - \nu \qquad \forall x \in \mathsf{X}$.
Let $(x^k)_{k\geq 0}$ be the trajectory generated by FAHBA.
Then the algorithm stops in no more than

$$K_I(\varepsilon, x^0) = \left\lceil 40(f(x^0) - f_{\min}(\mathsf{X}) + \varepsilon)\frac{\nu^2(\max\{M, L_0\} + \varepsilon/\nu)}{\varepsilon^2}\right\rceil = O\left(\frac{1}{\varepsilon^2}\right)$$

outer iterations, and the number of inner iterations is no more than
$2(K_I(\varepsilon, x^0) + 1) + \max\{\log_2(M/L_0), 0\}$.
Moreover, the last iterate obtained by FAHBA constitutes a 2ε-KKT point for problem (P) in the sense of definition on slide 12.

Anytime convergence by restarting the procedure, i.e., “path-following” method.

Define $\varepsilon_i = 2^{-i}\varepsilon_0$ for $i \geq 0$.

i-th restart/epoch: run FAHBA with the accuracy ε_i as an input and starting point x_i^0 that is the output of the previous restart.

Anytime convergence by restarting the procedure, i.e., “path-following” method.

Define $\varepsilon_i = 2^{-i}\varepsilon_0$ for $i \geq 0$.

i-th restart/epoch: run FAHBA with the accuracy ε_i as an input and starting point x_i^0 that is the output of the previous restart.

$p = \left\lceil \log_2 \frac{\varepsilon_0}{\varepsilon} \right\rceil$ restarts to achieve any $\varepsilon \in (0, \varepsilon_0]$.

Anytime convergence by restarting the procedure, i.e., “path-following” method.

Define $\varepsilon_i = 2^{-i}\varepsilon_0$ for $i \geq 0$.

i-th restart/epoch: run FAHBA with the accuracy ε_i as an input and starting point x_i^0 that is the output of the previous restart.

$p = \left\lceil \log_2 \frac{\varepsilon_0}{\varepsilon} \right\rceil$ restarts to achieve any $\varepsilon \in (0, \varepsilon_0]$.

The total complexity is $\sum_{i=0}^{p} O(\varepsilon_i^{-2}) = O(\varepsilon^{-2})$.

Anytime convergence by restarting the procedure, i.e., “path-following” method.

Define $\varepsilon_i = 2^{-i}\varepsilon_0$ for $i \geq 0$.

i-th restart/epoch: run FAHBA with the accuracy ε_i as an input and starting point x_i^0 that is the output of the previous restart.

$p = \left\lceil \log_2 \frac{\varepsilon_0}{\varepsilon} \right\rceil$ restarts to achieve any $\varepsilon \in (0, \varepsilon_0]$.

The total complexity is $\sum_{i=0}^{p} O(\varepsilon_i^{-2}) = O(\varepsilon^{-2})$.

Discussion:

- Same complexity $O(\varepsilon^{-2})$ as for unconstrained setting.

Anytime convergence by restarting the procedure, i.e., “path-following” method.

Define $\varepsilon_i = 2^{-i}\varepsilon_0$ for $i \geq 0$.

i-th restart/epoch: run FAHBA with the accuracy ε_i as an input and starting point x_i^0 that is the output of the previous restart.

$p = \left\lceil \log_2 \frac{\varepsilon_0}{\varepsilon} \right\rceil$ restarts to achieve any $\varepsilon \in (0, \varepsilon_0]$.

The total complexity is $\sum_{i=0}^{p} O(\varepsilon_i^{-2}) = O(\varepsilon^{-2})$.

Discussion:

- Same complexity $O(\varepsilon^{-2})$ as for unconstrained setting.
- Previous works consider particular case $\bar{\mathsf{K}} = \mathbb{R}^n_+$.

Anytime convergence by restarting the procedure, i.e., “path-following” method.

Define $\varepsilon_i = 2^{-i}\varepsilon_0$ for $i \geq 0$.

i-th restart/epoch: run FAHBA with the accuracy ε_i as an input and starting point x_i^0 that is the output of the previous restart.

$p = \left\lceil \log_2 \frac{\varepsilon_0}{\varepsilon} \right\rceil$ restarts to achieve any $\varepsilon \in (0, \varepsilon_0]$.

The total complexity is $\sum_{i=0}^{p} O(\varepsilon_i^{-2}) = O(\varepsilon^{-2})$.

Discussion:

- Same complexity $O(\varepsilon^{-2})$ as for unconstrained setting.
- Previous works consider particular case $\bar{\mathsf{K}} = \mathbb{R}_+^n$.
- The closest to ours result [Haeser, Liu, Ye, 2019] is $O(\varepsilon^{-2})$ complexity under similar assumptions, but only for $\bar{\mathsf{K}} = \mathbb{R}_+^n$. (see detailed discussion in the paper).

Local second-order smoothness assumption

$f : \mathsf{E} \to \mathbb{R} \cup \{+\infty\}$ is twice continuously differentiable on X and there exists a constant $M > 0$ such that, for all $x \in \mathsf{X}$ and $v \in \mathcal{T}_x$, we have

$$\|\nabla f(x+v) - \nabla f(x) - \nabla^2 f(x)v\|_x^* \leq \frac{M}{2}\|v\|_x^2. \tag{11}$$

Local second-order smoothness assumption

f : E $\to \mathbb{R} \cup \{+\infty\}$ is twice continuously differentiable on X and there exists a constant $M > 0$ such that, for all $x \in$ X and $v \in \mathcal{T}_x$, we have

$$\|\nabla f(x+v) - \nabla f(x) - \nabla^2 f(x)v\|_x^* \leq \frac{M}{2}\|v\|_x^2. \tag{11}$$

Then: $$f(x+v) - \left[f(x) + \langle \nabla f(x), v\rangle + \frac{1}{2}\langle \nabla^2 f(x)v, v\rangle\right] \leq \frac{M}{6}\|v\|_x^3. \tag{12}$$

Local second-order smoothness assumption

f : E $\to \mathbb{R} \cup \{+\infty\}$ is twice continuously differentiable on X and there exists a constant $M > 0$ such that, for all $x \in$ X and $v \in \mathcal{T}_x$, we have

$$\|\nabla f(x+v) - \nabla f(x) - \nabla^2 f(x)v\|_x^* \leq \frac{M}{2}\|v\|_x^2. \tag{11}$$

$$\text{Then: } f(x+v) - \left[f(x) + \langle \nabla f(x), v\rangle + \frac{1}{2}\langle \nabla^2 f(x)v, v\rangle\right] \leq \frac{M}{6}\|v\|_x^3. \tag{12}$$

The above assumption subsumes the standard Lipschitz Hessian setting if X is bounded.

Local second-order smoothness assumption

$f : \mathsf{E} \to \mathbb{R} \cup \{+\infty\}$ is twice continuously differentiable on X and there exists a constant $M > 0$ such that, for all $x \in \mathsf{X}$ and $v \in \mathcal{T}_x$, we have

$$\|\nabla f(x+v) - \nabla f(x) - \nabla^2 f(x)v\|_x^* \leq \frac{M}{2}\|v\|_x^2. \tag{11}$$

Then: $$f(x+v) - \left[f(x) + \langle \nabla f(x), v\rangle + \frac{1}{2}\langle \nabla^2 f(x)v, v\rangle\right] \leq \frac{M}{6}\|v\|_x^3. \tag{12}$$

The above assumption subsumes the standard Lipschitz Hessian setting if X is bounded.

Step direction:

$$v_{\mu,L}(x) \in \underset{v \in \mathsf{E}:\mathbf{A}v=0}{\operatorname{Argmin}} \{Q_{\mu,L}^{(2)}(x,v) \triangleq F_\mu(x) + \langle \nabla F_\mu(x), v\rangle + \frac{1}{2}\langle \nabla^2 f(x)v, v\rangle + \frac{L}{6}\|v\|_x^3\}.$$

Result: Point x^k, dual variables y^{k-1}, $s^k = \nabla f(x^k) - \mathbf{A}^* y^{k-1}$.

Set $144\varepsilon \triangleq \underline{L} < M_0$ – guess for M, $\mu = \frac{\varepsilon}{4\nu}$, $k = 0$, $x^0 \in \mathsf{X}$ – 4ν-a.c.;

repeat

repeat

Set $L_k = 2^{i_k} M_k$. Find $v^k \triangleq v_{\mu, L_k}(x^k)$ and $y^k \triangleq y_{\mu, L_k}(x^k)$ from

$$\min_{v : \mathbf{A}v = 0} \left\{ F_\mu(x^k) + \langle \nabla F_\mu(x^k), v \rangle + \frac{1}{2} \langle \nabla^2 f(x^k) v, v \rangle + \frac{L_k}{6} \|v\|_{x^k}^3 \right\}.$$

Set $\alpha_k \triangleq \min \left\{ 1, \frac{1}{2\|v^k\|_{x^k}} \right\}$.

until

$$f(x^k + \alpha_k v^k) \le f(x^k) + \alpha_k \langle \nabla f(x^k), v^k \rangle + \frac{\alpha_k^2}{2} \langle \nabla^2 f(x^k) v^k, v^k \rangle + \frac{L_k \alpha_k^3}{6} \|v^k\|_{x^k}^3,$$

and $\|\nabla f(x^k + \alpha_k v^k) - \nabla f(x^k) - \alpha_k \nabla^2 f(x^k) v^k\|_{x^k}^* \le \frac{L_k \alpha_k^2}{2} \|v^k\|_{x^k}^2$.

Set $M_{k+1} = \max\{2^{i_k - 1} M_k, \underline{L}\}$, $x^{k+1} = x^k + \alpha_k v^k$, $k = k + 1$;

until ; $\|v^{k-1}\|_{x^{k-1}} < \Delta_{k-1} \triangleq \sqrt{\frac{\varepsilon}{12 L_{k-1} \nu}}$ *and* $\|v^k\|_{x^k} < \Delta_k \triangleq \sqrt{\frac{\varepsilon}{12 L_k \nu}}$;

Complexity theorem for SAHBA [D., Staudigl, 2021, 2024]

Let our assumptions hold. Set h - SCB if $\bar{\mathsf{K}}$ is a convex set or h - LHSCB if $\bar{\mathsf{K}}$ is a convex cone. Fix $\varepsilon > 0$, some initial guess $M_0 > 144\varepsilon$ for the Lip. const. in (11), the regularization parameter $\mu = \frac{\varepsilon}{4\nu}$,
and x^0 to be a 4ν-analytic center.
Let $(x^k)_{k\geq 0}$ be the trajectory generated by SAHBA.

Complexity theorem for SAHBA [D., Staudigl, 2021, 2024]

Let our assumptions hold. Set h - SCB if $\bar{\mathsf{K}}$ is a convex set or h - LHSCB if $\bar{\mathsf{K}}$ is a convex cone. Fix $\varepsilon > 0$, some initial guess $M_0 > 144\varepsilon$ for the Lip. const. in (11), the regularization parameter $\mu = \frac{\varepsilon}{4\nu}$,
and x^0 to be a 4ν-analytic center.
Let $(x^k)_{k\geq 0}$ be the trajectory generated by SAHBA.
Then the algorithm stops in no more than

$$K_{II}(\varepsilon, x^0) = \left\lceil \frac{576\nu^{3/2}\sqrt{2\max\{M, M_0\}}(f(x^0) - f_{\min}(\mathsf{X}) + \varepsilon)}{\varepsilon^{3/2}} \right\rceil = O\left(\frac{1}{\varepsilon^{\frac{3}{2}}}\right)$$

outer iterations, and the number of inner iterations is no more than
$2(K_{II}(\varepsilon, x^0) + 1) + 2\max\{\log_2(2M/M_0), 1\}$.

Complexity theorem for SAHBA [D., Staudigl, 2021, 2024]

Let our assumptions hold. Set h - SCB if $\bar{\mathsf{K}}$ is a convex set or h - LHSCB if $\bar{\mathsf{K}}$ is a convex cone. Fix $\varepsilon > 0$, some initial guess $M_0 > 144\varepsilon$ for the Lip. const. in (11), the regularization parameter $\mu = \frac{\varepsilon}{4\nu}$,
and x^0 to be a 4ν-analytic center.
Let $(x^k)_{k\geq 0}$ be the trajectory generated by SAHBA.
Then the algorithm stops in no more than

$$K_{II}(\varepsilon, x^0) = \left\lceil \frac{576\nu^{3/2}\sqrt{2\max\{M, M_0\}}(f(x^0) - f_{\min}(\mathsf{X}) + \varepsilon)}{\varepsilon^{3/2}} \right\rceil = O\left(\frac{1}{\varepsilon^{\frac{3}{2}}}\right)$$

outer iterations, and the number of inner iterations is no more than
$2(K_{II}(\varepsilon, x^0) + 1) + 2\max\{\log_2(2M/M_0), 1\}$.
Moreover, the output of SAHBA is an $(\varepsilon, \frac{\max\{M, M_0\}\varepsilon}{24\nu})$-2KKT point for problem (P) in the sense of definition on slide 13.

- Same restarting strategy can be applied to achieve any-time convergence via a "path-following" method with same complexity $O(\varepsilon^{-3/2})$ up to a constant factor.

Discussion

- Same restarting strategy can be applied to achieve any-time convergence via a "path-following" method with same complexity $O(\varepsilon^{-3/2})$ up to a constant factor.
- Similar bound $O(\varepsilon^{-3/2})$ as for unconstrained setting.

- Same restarting strategy can be applied to achieve any-time convergence via a "path-following" method with same complexity $O(\varepsilon^{-3/2})$ up to a constant factor.
- Similar bound $O(\varepsilon^{-3/2})$ as for unconstrained setting.
- The closest to ours result [Haeser, Liu, Ye, 2019] (trust-region method), [O'Neill, Wright, 2020] (Newton-CG method) is $O(\varepsilon^{-3/2})$ complexity under similar assumptions, but with $\bar{\mathsf{K}} = \mathbb{R}^n_+$. Later [He, Lu, 2022] obtained close results for convex cones.

- Same restarting strategy can be applied to achieve any-time convergence via a "path-following" method with same complexity $O(\varepsilon^{-3/2})$ up to a constant factor.
- Similar bound $O(\varepsilon^{-3/2})$ as for unconstrained setting.
- The closest to ours result [Haeser, Liu, Ye, 2019] (trust-region method), [O'Neill, Wright, 2020] (Newton-CG method) is $O(\varepsilon^{-3/2})$ complexity under similar assumptions, but with $\bar{\mathsf{K}} = \mathbb{R}^n_+$. Later [He, Lu, 2022] obtained close results for convex cones.

P. Dvurechensky, M. Staudigl, Hessian barrier algorithms for non-convex conic optimization, Mathematical Programming, 2024 (arXiv:2111.00100, 2021).

- Same restarting strategy can be applied to achieve any-time convergence via a "path-following" method with same complexity $O(\varepsilon^{-3/2})$ up to a constant factor.
- Similar bound $O(\varepsilon^{-3/2})$ as for unconstrained setting.
- The closest to ours result [Haeser, Liu, Ye, 2019] (trust-region method), [O'Neill, Wright, 2020] (Newton-CG method) is $O(\varepsilon^{-3/2})$ complexity under similar assumptions, but with $\bar{\mathsf{K}} = \mathbb{R}^n_+$. Later [He, Lu, 2022] obtained close results for convex cones.

P. Dvurechensky, M. Staudigl, Hessian barrier algorithms for non-convex conic optimization, Mathematical Programming, 2024 (arXiv:2111.00100, 2021).

P. Dvurechensky, M. Staudigl, Barrier Algorithms for Constrained Non-Convex Optimization, ICML 2024.

- Extensions for convex setting:

  If f is convex, level sets of F_μ are bounded (e.g., f coercive) or $\bar{\mathsf{K}}$ is compact, slightly modified algorithms guarantee $f(x_k) - f_{\min}(\mathsf{X}) \leq \varepsilon$ in
  - $O\left((f(x^0) - f_{\min}(\mathsf{X})) + \frac{1}{\varepsilon}\right)$ by the first-order method.
  - $O\left((f(x^0) - f_{\min}(\mathsf{X})) + \frac{1}{\sqrt{\varepsilon}}\right)$ by the second-order method.
- Inexact oracle information, inexact resolution of subproblems.
- Numerical implementation.

1 Barrier algorithms for non-convex optimization

2 Minimizing self-concordant functions

- Unconstrained minimization by path-following methods
- Composite minimization by gradient regularization of Newton method
- Projection-free constrained minimization of self-concordant functions

Content

$$f^* \;=\; \min_{x \in \mathsf{E}} f(x), \tag{13}$$

where f is a M_f-self-concordant function:

$$|D^3 f(x)[u,u,u]| \leq 2M_f D^2 f(x)[u,u]^{3/2}. \tag{14}$$

$$f^* \;=\; \min_{x \in \mathsf{E}} f(x), \tag{13}$$

where f is a M_f-self-concordant function:

$$|D^3 f(x)[u,u,u]| \le 2M_f D^2 f(x)[u,u]^{3/2}. \tag{14}$$

Standard approach (e.g., [Nesterov, 2004]): apply Damped Newton Method (DNM)

$$x_+ = x - \frac{[\nabla^2 f(x)]^{-1} \nabla f(x)}{1 + M_f \lambda_f(x)}, \tag{15}$$

where $\lambda_f(x) = \|\nabla f(x)\|_x^*$.

$$f^* = \min_{x \in \mathsf{E}} f(x), \tag{13}$$

where f is a M_f-self-concordant function:

$$|D^3 f(x)[u,u,u]| \leq 2 M_f D^2 f(x)[u,u]^{3/2}. \tag{14}$$

Standard approach (e.g., [Nesterov, 2004]): apply Damped Newton Method (DNM)

$$x_+ = x - \frac{[\nabla^2 f(x)]^{-1} \nabla f(x)}{1 + M_f \lambda_f(x)}, \tag{15}$$

where $\lambda_f(x) = \|\nabla f(x)\|_x^*$.
Local quadratic convergence if $x \in \mathsf{Q} \triangleq \left\{ x \in \mathsf{E} : \; \lambda_f(x) \leq \frac{1}{2M_f} \right\}$.

$$f^* = \min_{x \in \mathsf{E}} f(x), \tag{13}$$

where f is a M_f-self-concordant function:

$$|D^3 f(x)[u,u,u]| \leq 2M_f D^2 f(x)[u,u]^{3/2}. \tag{14}$$

Standard approach (e.g., [Nesterov, 2004]): apply Damped Newton Method (DNM)

$$x_+ = x - \frac{[\nabla^2 f(x)]^{-1} \nabla f(x)}{1 + M_f \lambda_f(x)}, \tag{15}$$

where $\lambda_f(x) = \|\nabla f(x)\|_x^*$.
Local quadratic convergence if $x \in \mathsf{Q} \triangleq \left\{ x \in \mathsf{E} : \ \lambda_f(x) \leq \frac{1}{2M_f} \right\}$.
Complexity to reach Q:

$$N \leq \frac{\Delta(x_0)}{\omega\left(\frac{1}{2}\right)} = O(\Delta(x_0)), \quad \Delta(x_0) \triangleq M_f^2(f(x_0) - f^*). \tag{16}$$

Start from some $x_0 \in \mathsf{E}$. Define the central path $x(t)$ for $0 \leq t \leq 1$:

$$\nabla f(x(t)) = t \nabla f(x_0). \tag{17}$$

Start from some $x_0 \in \mathsf{E}$. Define the central path $x(t)$ for $0 \leq t \leq 1$:

$$\nabla f(x(t)) = t \nabla f(x_0). \tag{17}$$

Clearly, $x(1) = x_0$ and $x(0) = x^*$

Start from some $x_0 \in \mathsf{E}$. Define the central path $x(t)$ for $0 \leq t \leq 1$:

$$\nabla f(x(t)) = t \nabla f(x_0). \tag{17}$$

Clearly, $x(1) = x_0$ and $x(0) = x^*$ and this is a trajectory of minimizers:

$$x(t) = \arg\min_{x \in \mathsf{E}} \left\{ f_t(x) \triangleq f(x) - t \langle \nabla f(x_0), x \rangle \right\}, \quad 0 \leq t \leq 1. \tag{18}$$

Start from some $x_0 \in \mathsf{E}$. Define the central path $x(t)$ for $0 \leq t \leq 1$:

$$\nabla f(x(t)) = t\nabla f(x_0). \tag{17}$$

Clearly, $x(1) = x_0$ and $x(0) = x^*$ and this is a trajectory of minimizers:

$$x(t) = \arg\min_{x \in \mathsf{E}} \left\{ f_t(x) \triangleq f(x) - t\langle \nabla f(x_0), x\rangle \right\}, \quad 0 \leq t \leq 1. \tag{18}$$

Define: $\beta = 0.026$, $\gamma = 0.1125$.

Our goal is to follow the central path approximately:

$$\lambda_{f_t}(x) \equiv \|\nabla f(x) - t\nabla f(x_0)\|_x^* \leq \frac{\beta}{M_f} \tag{19}$$

Start from some $x_0 \in \mathsf{E}$. Define the central path $x(t)$ for $0 \le t \le 1$:

$$\nabla f(x(t)) = t \nabla f(x_0). \tag{17}$$

Clearly, $x(1) = x_0$ and $x(0) = x^*$ and this is a trajectory of minimizers:

$$x(t) = \arg\min_{x \in \mathsf{E}} \left\{ f_t(x) \stackrel{\Delta}{=} f(x) - t\langle \nabla f(x_0), x \rangle \right\}, \quad 0 \le t \le 1. \tag{18}$$

Define: $\beta = 0.026$, $\gamma = 0.1125$.

Our goal is to follow the central path approximately:

$$\lambda_{f_t}(x) \equiv \|\nabla f(x) - t\nabla f(x_0)\|_x^* \le \frac{\beta}{M_f} \tag{19}$$

by the path-following (PF) scheme:

$$(t_+, x_+) \;=\; \mathcal{P}(t,x) \;\equiv\; \left\{ \begin{array}{rcl} t_+ & = & \max\left\{ t - \frac{\gamma}{M_f \|\nabla f(x_0)\|_x^*}, 0 \right\}, \\ x_+ & = & x - [\nabla^2 f(x)]^{-1}(\nabla f(x) - t_+ \nabla f(x_0)). \end{array} \right. \tag{20}$$

Start from some $x_0 \in \mathsf{E}$. Define the central path $x(t)$ for $0 \leq t \leq 1$:

$$\nabla f(x(t)) = t \nabla f(x_0). \tag{17}$$

Clearly, $x(1) = x_0$ and $x(0) = x^*$ and this is a trajectory of minimizers:

$$x(t) = \arg\min_{x \in \mathsf{E}} \left\{ f_t(x) \triangleq f(x) - t\langle \nabla f(x_0), x \rangle \right\}, \quad 0 \leq t \leq 1. \tag{18}$$

Define: $\beta = 0.026$, $\gamma = 0.1125$.

Our goal is to follow the central path approximately:

$$\lambda_{f_t}(x) \equiv \|\nabla f(x) - t\nabla f(x_0)\|_x^* \leq \frac{\beta}{M_f} \tag{19}$$

by the path-following (PF) scheme:

$$(t_+, x_+) = \mathcal{P}(t, x) \equiv \begin{cases} t_+ = \max\left\{ t - \frac{\gamma}{M_f \|\nabla f(x_0)\|_x^*}, 0 \right\}, \\ x_+ = x - [\nabla^2 f(x)]^{-1}(\nabla f(x) - t_+ \nabla f(x_0)). \end{cases} \tag{20}$$

Unlike the standard setting, f is only a SCF, not SCB.

Complexity theorem for the path-following scheme [D., Nesterov, 2018]

Let f be a M_f-self-concordant function.

Complexity theorem for the path-following scheme [D., Nesterov, 2018]

Let f be a M_f-self-concordant function. Consider the following process:

$$t_0 = 1,\ x_0 \in \mathbb{E}, \quad (t_{k+1}, x_{k+1}) = \mathcal{P}(t_k, x_k), \quad k \geq 0, \tag{21}$$

where $\mathcal{P}$ is defined in (20).

Complexity theorem for the path-following scheme [D., Nesterov, 2018]

Let f be a M_f-self-concordant function. Consider the following process:

$$t_0 = 1,\ x_0 \in \mathbb{E}, \quad (t_{k+1}, x_{k+1}) = \mathcal{P}(t_k, x_k), \quad k \geq 0, \tag{21}$$

where $\mathcal{P}$ is defined in (20). Assume that $\lambda_f(x_k) \geq \frac{1}{2M_f}$ for all $k = 0, \ldots, N$.

Complexity theorem for the path-following scheme [D., Nesterov, 2018]

Let f be a M_f-self-concordant function. Consider the following process:

$$t_0 = 1,\ x_0 \in \mathbb{E}, \quad (t_{k+1}, x_{k+1}) = \mathcal{P}(t_k, x_k), \quad k \geq 0, \tag{21}$$

where $\mathcal{P}$ is defined in (20). Assume that $\lambda_f(x_k) \geq \frac{1}{2M_f}$ for all $k = 0, \ldots, N$. Then

$$t_N \leq \left(1 - \frac{\gamma(\gamma - 2\beta)N}{2M_f^2(f(x_0) - f^*)}\right)^N \leq \exp\left\{-\frac{\gamma(\gamma - 2\beta)N^2}{2M_f^2(f(x_0) - f^*)}\right\}. \tag{22}$$

Complexity theorem for the path-following scheme [D., Nesterov, 2018]

Let f be a M_f-self-concordant function. Consider the following process:

$$t_0 = 1,\ x_0 \in \mathbb{E}, \quad (t_{k+1}, x_{k+1}) = \mathcal{P}(t_k, x_k), \quad k \geq 0, \tag{21}$$

where $\mathcal{P}$ is defined in (20). Assume that $\lambda_f(x_k) \geq \frac{1}{2M_f}$ for all $k = 0, \dots, N$. Then

$$t_N \leq \left(1 - \frac{\gamma(\gamma - 2\beta)N}{2M_f^2(f(x_0) - f^*)}\right)^N \leq \exp\left\{-\frac{\gamma(\gamma - 2\beta)N^2}{2M_f^2(f(x_0) - f^*)}\right\}. \tag{22}$$

Moreover, when $t_{k+1} = 0$, the scheme automatically switches to the quadratically-convergent Newton method.

Complexity theorem for the path-following scheme [D., Nesterov, 2018]

Let f be a M_f-self-concordant function. Consider the following process:

$$t_0 = 1,\ x_0 \in \mathbb{E}, \quad (t_{k+1}, x_{k+1}) = \mathcal{P}(t_k, x_k), \quad k \geq 0, \tag{21}$$

where $\mathcal{P}$ is defined in (20). Assume that $\lambda_f(x_k) \geq \frac{1}{2M_f}$ for all $k = 0, \ldots, N$. Then

$$t_N \leq \left(1 - \frac{\gamma(\gamma - 2\beta)N}{2M_f^2(f(x_0) - f^*)}\right)^N \leq \exp\left\{-\frac{\gamma(\gamma - 2\beta)N^2}{2M_f^2(f(x_0) - f^*)}\right\}. \tag{22}$$

Moreover, when $t_{k+1} = 0$, the scheme automatically switches to the quadratically-convergent Newton method.

Finally, the complexity to find $x_N \in$ Q is $\widetilde{O}(\sqrt{\Delta(x_0)})$.

Complexity theorem for the path-following scheme [D., Nesterov, 2018]

Let f be a M_f-self-concordant function. Consider the following process:

$$t_0 = 1,\ x_0 \in \mathbb{E}, \quad (t_{k+1}, x_{k+1}) = \mathcal{P}(t_k, x_k), \quad k \geq 0, \tag{21}$$

where $\mathcal{P}$ is defined in (20). Assume that $\lambda_f(x_k) \geq \frac{1}{2M_f}$ for all $k = 0, \ldots, N$. Then

$$t_N \leq \left(1 - \frac{\gamma(\gamma - 2\beta)N}{2M_f^2(f(x_0) - f^*)}\right)^N \leq \exp\left\{-\frac{\gamma(\gamma - 2\beta)N^2}{2M_f^2(f(x_0) - f^*)}\right\}. \tag{22}$$

Moreover, when $t_{k+1} = 0$, the scheme automatically switches to the quadratically-convergent Newton method.
Finally, the complexity to find $x_N \in$ Q is $\widetilde{O}(\sqrt{\Delta(x_0)})$.

- Global super linear convergence.

Complexity theorem for the path-following scheme [D., Nesterov, 2018]

Let f be a M_f-self-concordant function. Consider the following process:

$$t_0 = 1,\ x_0 \in \mathbb{E}, \quad (t_{k+1}, x_{k+1}) = \mathcal{P}(t_k, x_k), \quad k \geq 0, \tag{21}$$

where $\mathcal{P}$ is defined in (20). Assume that $\lambda_f(x_k) \geq \frac{1}{2M_f}$ for all $k = 0, \ldots, N$. Then

$$t_N \leq \left(1 - \frac{\gamma(\gamma - 2\beta)N}{2M_f^2(f(x_0) - f^*)}\right)^N \leq \exp\left\{-\frac{\gamma(\gamma - 2\beta)N^2}{2M_f^2(f(x_0) - f^*)}\right\}. \tag{22}$$

Moreover, when $t_{k+1} = 0$, the scheme automatically switches to the quadratically-convergent Newton method.

Finally, the complexity to find $x_N \in$ Q is $\widetilde{O}(\sqrt{\Delta(x_0)})$.

- Global super linear convergence.
- Improved, „accelerated“, complexity $\widetilde{O}(\sqrt{\Delta(x_0)})$ (cf. $\widetilde{O}(\Delta(x_0))$ for the DNM).

Complexity theorem for the path-following scheme [D., Nesterov, 2018]

Let f be a M_f-self-concordant function. Consider the following process:

$$t_0 = 1,\ x_0 \in \mathbb{E}, \quad (t_{k+1}, x_{k+1}) = \mathcal{P}(t_k, x_k), \quad k \geq 0, \tag{21}$$

where $\mathcal{P}$ is defined in (20). Assume that $\lambda_f(x_k) \geq \frac{1}{2M_f}$ for all $k = 0, \ldots, N$. Then

$$t_N \leq \left(1 - \frac{\gamma(\gamma - 2\beta)N}{2M_f^2(f(x_0) - f^*)}\right)^N \leq \exp\left\{-\frac{\gamma(\gamma - 2\beta)N^2}{2M_f^2(f(x_0) - f^*)}\right\}. \tag{22}$$

Moreover, when $t_{k+1} = 0$, the scheme automatically switches to the quadratically-convergent Newton method.

Finally, the complexity to find $x_N \in$ Q is $\widetilde{O}(\sqrt{\Delta(x_0)})$.

- Global super linear convergence.
- Improved, „accelerated“, complexity $\widetilde{O}(\sqrt{\Delta(x_0)})$ (cf. $\widetilde{O}(\Delta(x_0))$ for the DNM).
- Adaptive version: iteratively try step-sizes $\gamma_k = 2^{1-i_k}\gamma_{k-1}$.

Define: $\beta = 0.0015$, $\gamma = 0.1158$.

Predictor-corrector path-following (PCPF) scheme:

$$(t_+, x_+) = \widetilde{\mathcal{P}}(t, x) \equiv \begin{cases} t_+ &= \max\left\{t - \frac{\gamma}{M_f \|\nabla f(x_0)\|_x^*}, 0\right\} \\ y &= x - \frac{\gamma}{M_f \|\nabla f(x_0)\|_x^*} [\nabla^2 f(x)]^{-1} \nabla f(x_0) \\ x_+ &= y - [\nabla^2 f(y)]^{-1} (\nabla f(y) - t_+ \nabla f(x_0)). \end{cases} \tag{23}$$

Unlike the standard setting, f is only an SCF, not SCB.

Complexity theorem for PCPF scheme [D., Nesterov, 2022]

Let f be a M_f-self-concordant function. Consider the following process:

$$t_0 = 1,\ x_0 \in \mathbb{E}, \quad (t_{k+1}, x_{k+1}) = \widetilde{\mathcal{P}}(t_k, x_k), \quad k \geq 0, \tag{24}$$

where $\mathcal{P}$ is defined in (23). Assume that $\lambda_f(x_k) \geq \frac{1}{2M_f}$ for all $k = 0, \ldots, N$. Then

$$t_N \leq \left(1 - \frac{\kappa(\beta,\gamma)N}{2M_f^2(f(x_0) - f^*)}\right)^N \leq \exp\left\{-\frac{\kappa(\beta,\gamma)N^2}{M_f^2(f(x_0) - f^*)}\right\}. \tag{25}$$

Moreover, when $t_{k+1} = 0$, the scheme automatically switches to the quadratically-convergent Newton method.

Finally, the complexity to find $x_N \in$ Q is $\widetilde{O}(\sqrt{\Delta(x_0)})$.

- Global super linear convergence.
- Improved, „accelerated“, complexity $\widetilde{O}(\sqrt{\Delta(x_0)})$ (cf. $\widetilde{O}(\Delta(x_0))$ for the DNM).
- Adaptive version: iteratively try step-sizes $\gamma_k = 2^{1-i_k}\gamma_{k-1}$.

Complexity theorem for PCPF scheme [D., Nesterov, 2022]

Let f be a M_f-self-concordant function. Consider the following process:

$$t_0 = 1,\ x_0 \in \mathbb{E}, \quad (t_{k+1}, x_{k+1}) = \widetilde{\mathcal{P}}(t_k, x_k), \quad k \geq 0, \tag{24}$$

where $\mathcal{P}$ is defined in (23). Assume that $\lambda_f(x_k) \geq \frac{1}{2M_f}$ for all $k = 0, \ldots, N$. Then

$$t_N \leq \left(1 - \frac{\kappa(\beta,\gamma)N}{2M_f^2(f(x_0) - f^*)}\right)^N \leq \exp\left\{-\frac{\kappa(\beta,\gamma)N^2}{M_f^2(f(x_0) - f^*)}\right\}. \tag{25}$$

Moreover, when $t_{k+1} = 0$, the scheme automatically switches to the quadratically-convergent Newton method.

Finally, the complexity to find $x_N \in$ Q is $\widetilde{O}(\sqrt{\Delta(x_0)})$.

- Global super linear convergence.
- Improved, „accelerated“, complexity $\widetilde{O}(\sqrt{\Delta(x_0)})$ (cf. $\widetilde{O}(\Delta(x_0))$ for the DNM).
- Adaptive version: iteratively try step-sizes $\gamma_k = 2^{1-i_k}\gamma_{k-1}$.
- Improved constant factor compared to path-following scheme.

- Improved complexity for feasibility problems

$$\text{Find } \; x \; \text{ s.t. } \; x \in Q \subset \mathbb{R}^n \; \text{ and } \; Ax = b, \tag{26}$$

where $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$, Q – closed, convex with $0 \in \mathrm{int} Q$.

- Improved complexity for feasibility problems

$$\text{Find } x \text{ s.t. } x \in Q \subset \mathbb{R}^n \text{ and } Ax = b, \tag{26}$$

where $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$, Q – closed, convex with $0 \in \mathrm{int}Q$.

- Improved constants in the complexity for minimization with primal method

$$\min \langle c, x\rangle \quad \text{s.t.} \quad x \in Q \subset \mathbb{R}^n, \tag{27}$$

Q – convex compact with nonempty interior.

- Improved complexity for feasibility problems

$$\text{Find } \ x \ \text{ s.t. } \ x \in Q \subset \mathbb{R}^n \ \text{ and } \ Ax = b, \tag{26}$$

where $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$, Q – closed, convex with $0 \in \mathrm{int}Q$.

- Improved constants in the complexity for minimization with primal method

$$\min \langle c, x\rangle \quad \text{s.t.} \quad x \in Q \subset \mathbb{R}^n, \tag{27}$$

Q – convex compact with nonempty interior.

- Improved constants in the complexity for minimization with dual method

$$\min\langle c, x\rangle \quad \text{s.t.} \quad Bx = 0, \quad x \in Q \subset \mathbb{R}^n, \tag{28}$$

where $B \in \mathbb{R}^{m\times n}$ and $0 \in \mathrm{int}Q$.

- Improved complexity for feasibility problems

$$\text{Find } \; x \; \text{ s.t. } \; x \in Q \subset \mathbb{R}^n \; \text{ and } \; Ax = b, \tag{26}$$

where $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$, Q – closed, convex with $0 \in \mathrm{int}Q$.

- Improved constants in the complexity for minimization with primal method

$$\min \langle c, x\rangle \quad \text{s.t.} \quad x \in Q \subset \mathbb{R}^n, \tag{27}$$

Q – convex compact with nonempty interior.

- Improved constants in the complexity for minimization with dual method

$$\min\langle c, x\rangle \quad \text{s.t.} \quad Bx = 0, \quad x \in Q \subset \mathbb{R}^n, \tag{28}$$

where $B \in \mathbb{R}^{m\times n}$ and $0 \in \mathrm{int}Q$.

P. Dvurechensky, Y. Nesterov. Global performance guarantees of second-order methods for unconstrained convex minimization. CORE Discussion Paper 2018/32.
P. Dvurechensky, Y. Nesterov. Improved global performance guarantees of second-order methods in convex minimization. arXiv:2408.11022.

Content

1. **Barrier algorithms for non-convex optimization**
   - Problem statement
   - Self-concordant barriers
   - Approximate optimality conditions
   - First-order algorithm
   - Second-order algorithm
2. **Minimizing self-concordant functions**
   - Unconstrained minimization by path-following methods
   - Composite minimization by gradient regularization of Newton method
   - Projection-free constrained minimization of self-concordant functions

$$\min_{x \in \mathsf{E}} \{F(x) \triangleq f(x) + \psi(x)\}, \tag{29}$$

where f is a M_f-self-concordant function, ψ is a simple closed convex function.

$$\min_{x \in \mathsf{E}} \{F(x) \triangleq f(x) + \psi(x)\}, \tag{29}$$

where f is a M_f-self-concordant function, ψ is a simple closed convex function.

Related works

- Proximal DNM [Tran-Dinh, Kyrillidis, Cevher, 2015].

$$\min_{x \in \mathsf{E}} \{ F(x) \triangleq f(x) + \psi(x) \}, \tag{29}$$

where f is a M_f-self-concordant function, ψ is a simple closed convex function.

Related works

- Proximal DNM [Tran-Dinh, Kyrillidis, Cevher, 2015].
- Composite PF method [Tran-Dinh, Liang, Toh, 2022] with ψ Lipschitz.

$$\min_{x \in \mathsf{E}} \{F(x) \triangleq f(x) + \psi(x)\}, \tag{29}$$

where f is a M_f-self-concordant function, ψ is a simple closed convex function.

Related works

- Proximal DNM [Tran-Dinh, Kyrillidis, Cevher, 2015].
- Composite PF method [Tran-Dinh, Liang, Toh, 2022] with ψ Lipschitz.
- Cubic regularization [Hanzely et al., 2022] for $\psi = 0$ and semi-strongly self-concordant f, sublinear rate.

$$\min_{x \in \mathsf{E}} \{F(x) \triangleq f(x) + \psi(x)\}, \tag{29}$$

where f is a M_f-self-concordant function, ψ is a simple closed convex function.

Related works

- Proximal DNM [Tran-Dinh, Kyrillidis, Cevher, 2015].
- Composite PF method [Tran-Dinh, Liang, Toh, 2022] with ψ Lipschitz.
- Cubic regularization [Hanzely et al., 2022] for $\psi = 0$ and semi-strongly self-concordant f, sublinear rate.
- Newton algorithms with gradient norm regularization for f with Lipschitz Hessian [Mishchenko, 2021], [Doikov, Nesterov, 2021], [Doikov, Mishchenko, Nesterov, 2022] or quasi-self-concordant **[Doikov, 2023]**.

$$\min_{x \in \mathsf{E}} \{F(x) \triangleq f(x) + \psi(x)\}, \tag{29}$$

where f is a M_f-self-concordant function, ψ is a simple closed convex function.

Related works

- Proximal DNM [Tran-Dinh, Kyrillidis, Cevher, 2015].
- Composite PF method [Tran-Dinh, Liang, Toh, 2022] with ψ Lipschitz.
- Cubic regularization [Hanzely et al., 2022] for $\psi = 0$ and semi-strongly self-concordant f, sublinear rate.
- Newton algorithms with gradient norm regularization for f with Lipschitz Hessian [Mishchenko, 2021], [Doikov, Nesterov, 2021], [Doikov, Mishchenko, Nesterov, 2022] or quasi-self-concordant **[Doikov, 2023]**.

We analyze a Newton method with gradient norm regularization for self-concordant functions (GRN-SCF).

Gradient-regularized Newton method for self-concordant functions (GRN-SCF):

$$x^+ = \arg\min_{y\in\mathsf{E}}\Big\{\langle\nabla f(x), y-x\rangle + \frac{1}{2}\langle\nabla^2 f(x)(y-x), y-x\rangle \tag{30}$$

$$+ \frac{\sigma\|F'(x)\|_x}{2}\|y-x\|_x^2 + \psi(y)\Big\}, \tag{31}$$

where $\sigma \geq 0$ and $F'(x) \in \partial F(x)$, meaning that we use (sub)gradient regularization.

Gradient-regularized Newton method for self-concordant functions (GRN-SCF):

$$x^+ = \arg\min_{y \in \mathsf{E}} \Big\{ \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y-x), y - x \rangle \tag{30}$$

$$+ \frac{\sigma \|F'(x)\|_x}{2} \|y - x\|_x^2 + \psi(y) \Big\}, \tag{31}$$

where $\sigma \geq 0$ and $F'(x) \in \partial F(x)$, meaning that we use (sub)gradient regularization.

NB: if ψ is an indicator of a convex set, GRN-SCF requires projection.

Gradient-regularized Newton method for self-concordant functions (GRN-SCF):

$$x^+ = \arg\min_{y \in \mathsf{E}} \Big\{ \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y-x), y - x \rangle \tag{30}$$

$$+ \frac{\sigma \|F'(x)\|_x}{2} \|y - x\|_x^2 + \psi(y) \Big\}, \tag{31}$$

where $\sigma \geq 0$ and $F'(x) \in \partial F(x)$, meaning that we use (sub)gradient regularization.

NB: if ψ is an indicator of a convex set, GRN-SCF requires projection.

We show that the iterates stay on the sublevel set defined by the starting point $\mathcal{L}(x^0) \triangleq \{x \in \operatorname{dom} \psi : F(x) \leq F(x_0)\}$.

We assume that this sublevel set is bounded.

Gradient-regularized Newton method for self-concordant functions (GRN-SCF):

$$x^+ = \arg\min_{y\in\mathsf{E}} \Big\{ \langle \nabla f(x), y-x\rangle + \frac{1}{2}\langle \nabla^2 f(x)(y-x), y-x\rangle \tag{30}$$

$$+ \frac{\sigma\|F'(x)\|_x}{2}\|y-x\|_x^2 + \psi(y) \Big\}, \tag{31}$$

where $\sigma \geq 0$ and $F'(x) \in \partial F(x)$, meaning that we use (sub)gradient regularization.

NB: if ψ is an indicator of a convex set, GRN-SCF requires projection.

We show that the iterates stay on the sublevel set defined by the starting point $\mathcal{L}(x^0) \triangleq \{x \in \operatorname{dom}\psi : F(x) \leq F(x_0)\}$.

We assume that this sublevel set is bounded. This implies

$$D(x^0) \triangleq \sup_{x,y\in\mathcal{L}(x^0)} \|y-x\|_x < +\infty. \tag{32}$$

Complexity theorem for GRN-SCF [D., 2024]

Let in (29) f be a M_f-self-concordant function, sublevel set $\mathcal{L}(x^0)$ be bounded, $\sigma = 3M_f$.

Complexity theorem for GRN-SCF [D., 2024]

Let in (29) f be a M_f-self-concordant function, sublevel set $\mathcal{L}(x^0)$ be bounded, $\sigma = 3M_f$. Then, GRN-SCF has global linear convergence rate, i.e., for $k \geq 1$,

$$F(x^k) - F(x^*) \leq \exp\left(-\frac{k}{54 M_f D(x^0)}\right)\left(F(x^0) - F(x^*)\right) + \exp\left(-\frac{k}{4}\right) g_0 D(x^0).$$

Complexity theorem for GRN-SCF [D., 2024]

Let in (29) f be a M_f-self-concordant function, sublevel set $\mathcal{L}(x^0)$ be bounded, $\sigma = 3M_f$. Then, GRN-SCF has global linear convergence rate, i.e., for $k \geq 1$,

$$F(x^k) - F(x^*) \leq \exp\left(-\frac{k}{54M_f D(x^0)}\right)\left(F(x^0) - F(x^*)\right) + \exp\left(-\frac{k}{4}\right) g_0 D(x^0).$$

Moreover, if $\|F'(x^0)\|^*_{x^0} \leq \frac{4}{45M_f}$, GRN-SCF has local quadratic convergence

$$\|F'(x^{k+1})\|^*_{x^{k+1}} \leq \frac{45M_f}{4}\left(\|F'(x^k)\|^*_{x^k}\right)^2.$$

Complexity theorem for GRN-SCF [D., 2024]

Let in (29) f be a M_f-self-concordant function, sublevel set $\mathcal{L}(x^0)$ be bounded, $\sigma = 3M_f$. Then, GRN-SCF has global linear convergence rate, i.e., for $k \geq 1$,

$$F(x^k) - F(x^*) \leq \exp\left(-\frac{k}{54 M_f D(x^0)}\right)\left(F(x^0) - F(x^*)\right) + \exp\left(-\frac{k}{4}\right) g_0 D(x^0).$$

Moreover, if $\|F'(x^0)\|_{x^0}^* \leq \frac{4}{45 M_f}$, GRN-SCF has local quadratic convergence

$$\|F'(x^{k+1})\|_{x^{k+1}}^* \leq \frac{45 M_f}{4}\left(\|F'(x^k)\|_{x^k}^*\right)^2.$$

We propose also an adaptive version.

Complexity theorem for GRN-SCF [D., 2024]

Let in (29) f be a M_f-self-concordant function, sublevel set $\mathcal{L}(x^0)$ be bounded, $\sigma = 3M_f$. Then, GRN-SCF has global linear convergence rate, i.e., for $k \geq 1$,

$$F(x^k) - F(x^*) \leq \exp\left(-\frac{k}{54M_f D(x^0)}\right)\left(F(x^0) - F(x^*)\right) + \exp\left(-\frac{k}{4}\right) g_0 D(x^0).$$

Moreover, if $\|F'(x^0)\|_{x^0}^* \leq \frac{4}{45M_f}$, GRN-SCF has local quadratic convergence

$$\|F'(x^{k+1})\|_{x^{k+1}}^* \leq \frac{45M_f}{4}(\|F'(x^k)\|_{x^k}^*)^2.$$

We propose also an adaptive version.

Ours vs [Hanzely et al., 2022]: wider problem class and linear convergence.

Complexity theorem for GRN-SCF [D., 2024]

Let in (29) f be a M_f-self-concordant function, sublevel set $\mathcal{L}(x^0)$ be bounded, $\sigma = 3M_f$. Then, GRN-SCF has global linear convergence rate, i.e., for $k \geq 1$,

$$F(x^k) - F(x^*) \leq \exp\left(-\frac{k}{54 M_f D(x^0)}\right)\left(F(x^0) - F(x^*)\right) + \exp\left(-\frac{k}{4}\right) g_0 D(x^0).$$

Moreover, if $\|F'(x^0)\|_{x^0}^* \leq \frac{4}{45 M_f}$, GRN-SCF has local quadratic convergence

$$\|F'(x^{k+1})\|_{x^{k+1}}^* \leq \frac{45 M_f}{4}\left(\|F'(x^k)\|_{x^k}^*\right)^2.$$

We propose also an adaptive version.

Ours vs [Hanzely et al., 2022]: wider problem class and linear convergence.

Future work: combination of HBA and gradient regularization.

Complexity theorem for GRN-SCF [D., 2024]

Let in (29) f be a M_f-self-concordant function, sublevel set $\mathcal{L}(x^0)$ be bounded, $\sigma = 3M_f$. Then, GRN-SCF has global linear convergence rate, i.e., for $k \geq 1$,

$$F(x^k) - F(x^*) \leq \exp\left(-\frac{k}{54M_f D(x^0)}\right)\left(F(x^0) - F(x^*)\right) + \exp\left(-\frac{k}{4}\right) g_0 D(x^0).$$

Moreover, if $\|F'(x^0)\|^*_{x^0} \leq \frac{4}{45M_f}$, GRN-SCF has local quadratic convergence

$$\|F'(x^{k+1})\|^*_{x^{k+1}} \leq \frac{45M_f}{4}\left(\|F'(x^k)\|^*_{x^k}\right)^2.$$

We propose also an adaptive version.

Ours vs [Hanzely et al., 2022]: wider problem class and linear convergence.

Future work: combination of HBA and gradient regularization.

P. Dvurechensky. Newton method with gradient regularization for minimizing self-concordant functions. In preparation.

Content

$$\min_{x \in \mathcal{X} \subset \mathsf{E}} f(x), \tag{P}$$

where f is M_f-self-concordant function,

$$\min_{x \in \mathcal{X} \subset \mathsf{E}} f(x), \tag{P}$$

where f is M_f-self-concordant function,

$\mathcal{X}$ – convex compact with atomic or another Linear Minimization Oracle (LMO) friendly structure: ℓ_1-ball, Spectrahedron, etc. $\Rightarrow$ Frank-Wolfe (FW)/Conditional Gradient (CG) methods [Frank & Wolfe, 1956], [Levitin & Polyak, 1966], [Jaggi, 2013].

$$\min_{x \in \mathcal{X} \subset \mathsf{E}} f(x), \tag{P}$$

where f is M_f-self-concordant function,

$\mathcal{X}$ – convex compact with atomic or another Linear Minimization Oracle (LMO) friendly structure: ℓ_1-ball, Spectrahedron, etc. $\Rightarrow$ Frank-Wolfe (FW)/Conditional Gradient (CG) methods [Frank & Wolfe, 1956], [Levitin & Polyak, 1966], [Jaggi, 2013].

Standard analysis relies on Lipschitz gradient/bounded curvature.

$$\min_{x \in \mathcal{X} \subset \mathsf{E}} f(x), \tag{P}$$

where f is M_f-self-concordant function,

$\mathcal{X}$ – convex compact with atomic or another Linear Minimization Oracle (LMO) friendly structure: ℓ_1-ball, Spectrahedron, etc. $\Rightarrow$ Frank-Wolfe (FW)/Conditional Gradient (CG) methods [Frank & Wolfe, 1956], [Levitin & Polyak, 1966], [Jaggi, 2013].

Standard analysis relies on Lipschitz gradient/bounded curvature.

Related works

- [Bach, 2010], [Ostrovskii & Bach, 2018] Non-Lipschitz smooth losses in ML.

$$\min_{x\in\mathcal{X}\subset\mathsf{E}} f(x), \tag{P}$$

where f is M_f-self-concordant function,

$\mathcal{X}$ – convex compact with atomic or another Linear Minimization Oracle (LMO) friendly structure: ℓ_1-ball, Spectrahedron, etc. $\Rightarrow$ Frank-Wolfe (FW)/Conditional Gradient (CG) methods [Frank & Wolfe, 1956], [Levitin & Polyak, 1966], [Jaggi, 2013].

Standard analysis relies on Lipschitz gradient/bounded curvature.

Related works

- [Bach, 2010], [Ostrovskii & Bach, 2018] Non-Lipschitz smooth losses in ML.
- [Odor et al., 2016] FW algorithm for Poisson inverse problem in phase retrieval.

$$\min_{x \in \mathcal{X} \subset \mathsf{E}} f(x), \tag{P}$$

where f is M_f-self-concordant function,

$\mathcal{X}$ – convex compact with atomic or another Linear Minimization Oracle (LMO) friendly structure: ℓ_1-ball, Spectrahedron, etc. $\Rightarrow$ Frank-Wolfe (FW)/Conditional Gradient (CG) methods [Frank & Wolfe, 1956], [Levitin & Polyak, 1966], [Jaggi, 2013].

Standard analysis relies on Lipschitz gradient/bounded curvature.

Related works

- [Bach, 2010], [Ostrovskii & Bach, 2018] Non-Lipschitz smooth losses in ML.
- [Odor et al., 2016] FW algorithm for Poisson inverse problem in phase retrieval.
- [Liu et al., 2020] Newton-FW algorithm for minimizing self-concordant functions.

$$\min_{x \in \mathcal{X} \subset \mathsf{E}} f(x), \qquad \text{(P)}$$

where f is M_f-self-concordant function,

$\mathcal{X}$ – convex compact with atomic or another Linear Minimization Oracle (LMO) friendly structure: ℓ_1-ball, Spectrahedron, etc. $\Rightarrow$ Frank-Wolfe (FW)/Conditional Gradient (CG) methods [Frank & Wolfe, 1956], [Levitin & Polyak, 1966], [Jaggi, 2013].

Standard analysis relies on Lipschitz gradient/bounded curvature.

Related works

- [Bach, 2010], [Ostrovskii & Bach, 2018] Non-Lipschitz smooth losses in ML.
- [Odor et al., 2016] FW algorithm for Poisson inverse problem in phase retrieval.
- [Liu et al., 2020] Newton-FW algorithm for minimizing self-concordant functions.
- [Carderera & Pokutta, 2020] Newton-FW approach for objectives with Lipschitz Hessians.

$$\min_{x \in \mathcal{X} \subset \mathsf{E}} f(x), \tag{P}$$

where f is M_f-self-concordant function,

$\mathcal{X}$ – convex compact with atomic or another Linear Minimization Oracle (LMO) friendly structure: ℓ_1-ball, Spectrahedron, etc. $\Rightarrow$ Frank-Wolfe (FW)/Conditional Gradient (CG) methods [Frank & Wolfe, 1956], [Levitin & Polyak, 1966], [Jaggi, 2013].

Standard analysis relies on Lipschitz gradient/bounded curvature.

Related works

- [Bach, 2010], [Ostrovskii & Bach, 2018] Non-Lipschitz smooth losses in ML.
- [Odor et al., 2016] FW algorithm for Poisson inverse problem in phase retrieval.
- [Liu et al., 2020] Newton-FW algorithm for minimizing self-concordant functions.
- [Carderera & Pokutta, 2020] Newton-FW approach for objectives with Lipschitz Hessians.
- [Zhao & Freund, 2020] FW for composite minimization involving LHSCB.

Preliminaries:

Linear minimization oracle: $s(x) = \operatorname{argmin}_{s \in \mathcal{X}} \langle \nabla f(x), s \rangle$.

Preliminaries:

Linear minimization oracle: $s(x) = \operatorname{argmin}_{s\in\mathcal{X}} \langle \nabla f(x), s\rangle$.

FW gap $\mathsf{Gap}(x) = \langle \nabla f(x), x - s(x)\rangle$ (NB: $\mathsf{Gap}(x) \geq f(x) - f^*$).

Preliminaries:

Linear minimization oracle: $s(x) = \mathrm{argmin}_{s \in \mathcal{X}} \langle \nabla f(x), s \rangle$.

FW gap $\mathsf{Gap}(x) = \langle \nabla f(x), x - s(x) \rangle$ (NB: $\mathsf{Gap}(x) \geq f(x) - f^*$).

Frank-Wolfe method for SCF:

While $\mathsf{Gap}(x^k) > \varepsilon$ do

1. Obtain $s^k = s(x^k)$;

Preliminaries:

Linear minimization oracle: $s(x) = \mathrm{argmin}_{s \in \mathcal{X}} \langle \nabla f(x), s \rangle$.

FW gap $\mathsf{Gap}(x) = \langle \nabla f(x), x - s(x) \rangle$ (NB: $\mathsf{Gap}(x) \geq f(x) - f^*$).

Frank-Wolfe method for SCF:

While $\mathsf{Gap}(x^k) > \varepsilon$ do

1. Obtain $s^k = s(x^k)$;
2. Set $\alpha_k = \min\left\{1, \frac{\mathsf{Gap}(x^k)}{M_f \|s^k - x^k\|_{x^k} (\mathsf{Gap}(x^k) + M_f \|s^k - x^k\|_{x^k})}\right\}$;

Preliminaries:

Linear minimization oracle: $s(x) = \operatorname{argmin}_{s \in \mathcal{X}} \langle \nabla f(x), s \rangle$.

FW gap $\mathsf{Gap}(x) = \langle \nabla f(x), x - s(x) \rangle$ (NB: $\mathsf{Gap}(x) \geq f(x) - f^*$).

Frank-Wolfe method for SCF:

While $\mathsf{Gap}(x^k) > \varepsilon$ do

1. Obtain $s^k = s(x^k)$;
2. Set $\alpha_k = \min \left\{ 1, \frac{\mathsf{Gap}(x^k)}{M_f \|s^k - x^k\|_{x^k} (\mathsf{Gap}(x^k) + M_f \|s^k - x^k\|_{x^k})} \right\}$;
3. Update $x^{k+1} = x^k + \alpha_k (s^k - x^k)$.

Let

$$S(x^0) = \{x \in \mathcal{X} | f(x) \leq f(x^0)\}, \text{ and } L_{\nabla f} = \max_{x \in S(x^0)} \lambda_{\max}(\nabla^2 f(x)).$$

Let

$$S(x^0) = \{x \in \mathcal{X} | f(x) \leq f(x^0)\}, \text{ and } L_{\nabla f} \quad = \max_{x \in S(x^0)} \lambda_{\max}(\nabla^2 f(x)).$$

Complexity theorem for FW-SCF [D., Ostroukhov, Safin, Shtern, Staudigl, 2020]

For given $\varepsilon > 0$, define $N_\varepsilon(x^0) = \min\{k \geq 0 | f(x^k) - f^* \leq \varepsilon\}$. Then,

$$N_\varepsilon(x^0) \leq \frac{1}{c_1} \ln\left(\frac{c_1}{(f(x^0) - f^*)c_2}\right) + \frac{4L_{\nabla f}\mathrm{diam}(\mathcal{X})}{\varepsilon},$$

where c_1, c_2 are explicit constants depending on $M_f, L_{\nabla f}, \mathrm{diam}(\mathcal{X})$.

We also propose extensions:

- Minimization of generalized self-concordant functions [Sun & Tran-Dinh, 2018];

We also propose extensions:

- Minimization of generalized self-concordant functions [Sun & Tran-Dinh, 2018];
- Line-search variants;

We also propose extensions:

- Minimization of generalized self-concordant functions [Sun & Tran-Dinh, 2018];
- Line-search variants;
- Linearly Convergent Variants on polytopes;

We also propose extensions:

- Minimization of generalized self-concordant functions [Sun & Tran-Dinh, 2018];
- Line-search variants;
- Linearly Convergent Variants on polytopes;
- A conditional gradient homotopy method for conic-constrained problems:

$$\min_{x} g(x) \quad \text{s.t. } x \in X, Ax \in K \subseteq H, \tag{P}$$

  where g is a closed convex lsc function, $X \subset \mathsf{E}$ is a LMO-friendly convex compact, $A : \mathsf{E} \to H$ is an affine mapping, and K is a closed convex pointed cone.

We also propose extensions:

- Minimization of generalized self-concordant functions [Sun & Tran-Dinh, 2018];
- Line-search variants;
- Linearly Convergent Variants on polytopes;
- A conditional gradient homotopy method for conic-constrained problems:

$$\min_x g(x) \quad \text{s.t. } x \in X, Ax \in K \subseteq H, \tag{P}$$

where g is a closed convex lsc function, $X \subset \mathsf{E}$ is a LMO-friendly convex compact, $A : \mathsf{E} \to H$ is an affine mapping, and K is a closed convex pointed cone.

P. Dvurechensky, P. Ostroukhov, K. Safin, S. Shtern, M. Staudigl, Self-Concordant Analysis of Frank-Wolfe Algorithms, ICML 2020

P. Dvurechensky, K. Safin, S. Shtern, M. Staudigl, Generalized Self-Concordant Analysis of Frank-Wolfe algorithms, Math. Progr., 2022

P. Dvurechensky, S. Shtern, M. Staudigl, A conditional gradient homotopy method with applications to Semidefinite Programming, arXiv:2207.03101, 2022

Thank you for your attention!

P. Dvurechensky, M. Staudigl, Hessian barrier algorithms for non-convex conic optimization, Mathematical Programming, 2024 (arXiv:2111.00100, 2021).

P. Dvurechensky, M. Staudigl, Barrier Algorithms for Constrained Non-Convex Optimization, ICML 2024.

P. Dvurechensky, Y. Nesterov. Global performance guarantees of second-order methods for unconstrained convex minimization. CORE Discussion Paper 2018/32.

P. Dvurechensky, Y. Nesterov. Improved global performance guarantees of second-order methods in convex minimization. arXiv:2408.11022.

P. Dvurechensky. Newton method with gradient regularization for minimizing self-concordant functions. To appear on arXiv.

P. Dvurechensky, P. Ostroukhov, K. Safin, S. Shtern, M. Staudigl, Self-Concordant Analysis of Frank-Wolfe Algorithms, ICML, 2020

P. Dvurechensky, K. Safin, S. Shtern, M. Staudigl, Generalized Self-Concordant Analysis of Frank-Wolfe algorithms, Math. Progr., 2022

P. Dvurechensky, S. Shtern, M. Staudigl, A conditional gradient homotopy method with applications to Semidefinite Programming, arXiv:2207.03101, 2022

Regularized non-linear regression problem: training input convex neural networks (ICNN) with sparsity penalty

ICNN: $\Phi(z, x)$, where z is the input data and x are parameters. If $x \geq 0$ and ReLU nonlinearity is used, then $\Phi(\cdot, x)$ is convex. But, the training problem is non-convex.

$$\min_{x \geq 0} \left\{ f(x) = \|\Phi(\hat{z}, x) - \hat{y}\|_2^2 + \lambda \|x\|_p^p \right\}, \tag{33}$$

where $\ell(x)$ is a non-convex loss function, $\lambda > 0$, $p \in (0, 1)$.

Recent interest in non-Lipschitz smooth losses

- [Bach, 2010] Logistic regression as a generalized self-concordant function.
- [Owen, 2013] Self-concordance for empirical likelihood.
- [Odor et al., 2016] Poisson inverse problem in phase retrieval.
- [Ostrovskii & Bach, 2018] Finite-sample analysis of M-estimators using self-concordance.
- [Marteau-Ferey et al., 2019] Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance.

[Nesterov & Nemirovski, 1994]

- **Portfolio Optimization**

$$f(x) = -\sum_{t=1}^{T} \ln(\langle r_t, x\rangle), x \in \mathcal{X} = \Delta_n$$

- **Covariance Estimation:**

$$f(x) = -\ln(\det(x)) + \mathrm{tr}(\Sigma x),$$
$$x \in \mathcal{X} = \{x \in \mathcal{S}_+^n : \|\mathrm{vec}(x)\|_1 \leq R\}.$$

- **Poisson Inverse Problem**

$$f(x) = \sum_{i=1}^{m} \langle w_i, x\rangle - \sum_{i=1}^{m} y_i \ln(\langle w_i, x\rangle),$$
$$x \in \mathcal{X} = \{x \in \mathbb{R}^n | \|x\|_1 \leq R\}.$$

- **Logistic Loss** ($\nu = 2$ or $\nu = 3$).

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} \ln\left(1 + \exp(b_i\langle a_i, x\rangle)\right) + \frac{\mu}{2}\|x\|_2^2.$$

where $b_i \in \{-1, 1\}, \mu > 0, a_i \in \mathbb{R}^n$.

- **Robust regression** ($\nu = 2$)

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} \varphi(b_i - \langle a_i, x\rangle),\ \varphi(u) = \ln(e^u + e^{-u}).$$

- **Distance-Weighted Discrimination** ($\nu = 2(q+3)/(q+2)$)

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} (a_i^\top w + \beta y_i + \xi_i)^{-q} + \langle c, \xi\rangle,\ x = (w, \beta, \xi).$$