First-order methods for nonlinear optimization with a few constraints

Yangyang Xu

Mathematical Sciences, Rensselaer Polytechnic Institute

7th International Conference on Continuous Optimization

July 26, 2022

Partly supported by NSF awards #1719549 and #2053493.

Outline

1. Lower complexity bounds

2. Near-optimal FOMs for problems with $O(1)$ functional constraints

   Examples with $O(1)$ nonlinear constraints

   Cutting-plane first-order method for primal subproblems

   Numerical experiments on QCQP

3. Conclusions

Part I: lower complexity bound for affine-constrained problems

(joint work with Yuyuan Ouyang from Clemson University)

Why lower complexity bounds

- provide understanding of the fundamental limit of a class of methods and the difficulty of a class of problems
- tell if existing methods could be improved
- guide to design "optimal" methods

First-order methods for smooth convex problems [Nesterov'04]

Consider problem

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}}\, f(\mathbf{x})$$

- f is convex and L-smooth, i.e.,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x}-\mathbf{y}\|,\ \forall\, \mathbf{x},\mathbf{y}$$

- lower bound: $f(\mathbf{x}^k) - f(\mathbf{x}^*) \geq \frac{3L\|\mathbf{x}^0-\mathbf{x}^*\|^2}{32(k+1)^2}$ if $k \leq \frac{n-1}{2}$ and

$$\mathbf{x}^k \in \mathbf{x}^0 + \text{span}\{\nabla f(\mathbf{x}^0), \nabla f(\mathbf{x}^1), \ldots, \nabla f(\mathbf{x}^{k-1})\}$$

- upper bound: $f(\mathbf{x}^k) - f(\mathbf{x}^*) \leq \frac{4L\|\mathbf{x}^0-\mathbf{x}^*\|^2}{(k+1)^2}$

First-order methods for nonsmooth convex problems [Nesterov'04]

Consider problem

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}}\, f(\mathbf{x})$$

- f is convex and M-Lipschitz continuous on $X = \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}^0\| \leq R\}$, i.e.,

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq M\|\mathbf{x} - \mathbf{y}\|,\ \forall\, \mathbf{x}, \mathbf{y} \in X$$

- lower bound: $f(\mathbf{x}^k) - f(\mathbf{x}^*) \geq \frac{MR}{2(1+\sqrt{k+1})}$ if $k \leq n-1$, and

$$\mathbf{x}^k \in \mathbf{x}^0 + \text{span}\{\mathbf{g}^0, \mathbf{g}^1, \ldots, \mathbf{g}^{k-1}\}$$

  where $\mathbf{g}^t \in \partial f(\mathbf{x}^t)$

- upper bound: $f(\mathbf{x}^k) - f(\mathbf{x}^*) \leq \frac{2MR}{\sqrt{k+1}}$

Remark: non-smooth problems are harder than smooth ones.

More examples

- first-order methods for stochastic convex problems [Agarwal et. al'12]
- first-order methods for finite-sum convex problems [Woodworth-Srebro'16]
- first-order and higher-order methods for nonconvex problems [Carmon et. al'19a, Carmon et. al'19b]
- first-order decentralized methods for convex or nonconvex problems [Scaman et. al'19, Sun-Hong'19]
- first-order methods for convex-concave saddle-point problems [Zhang-Hong-Zhang'21]
- .....

linearly constrained problems by linear span [Ouyang-X.'21][1]

Consider

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}),\ \text{s.t. } \mathbf{A}\mathbf{x}=\mathbf{b}$$

- problem class I: f is smooth (i.e., ∇f is Lipschitz) and convex
- problem class II: f is smooth and strongly convex
- algorithm class:

$$\mathbf{x}^t \in \mathbf{x}^0 + \mathrm{Span}\{\nabla f(\mathbf{x}^0), \mathbf{A}^\top \mathbf{r}^0, \ldots, \nabla f(\mathbf{x}^{t-1}), \mathbf{A}^\top \mathbf{r}^{t-1}\} \qquad \text{(Span)}$$

  where $\mathbf{r}^t = \mathbf{A}\mathbf{x}^t - \mathbf{b}$

  - Without loss of generality, assume that $\mathbf{x}^0 = \mathbf{0}$
- error measure: $|f(\mathbf{x}^t) - f^*|$ and $\|\mathbf{A}\mathbf{x}^t - \mathbf{b}\|$, or $\|\mathbf{x}^t - \mathbf{x}^*\|^2$

[1] Results for bilinear convex-concave problems by general first-order methods are shown in "Ouyang and Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. Math Prog."

lower bound for convex case

Setting of problem class:

- given positive integers $m \leq n$, and $t < \frac{m}{2}$
- given positive numbers L_A and L_f

Conclusion: there exists an instance of smooth linearly constrained convex problem such that

- ∇f is L_f-Lipschitz continuous, $\|\mathbf{A}\|_2 = L_A$
- it has a unique primal-dual solution $(\mathbf{x}^*, \mathbf{y}^*)$, i.e., satisfying KKT system
- in addition, for (Span), it holds

$$|f(\mathbf{x}^t) - f(\mathbf{x}^*)| \geq \frac{3L_f\|\mathbf{x}^*\|^2}{64(t+1)^2} + \frac{\sqrt{3}L_A\|\mathbf{x}^*\| \cdot \|\mathbf{y}^*\|}{16(t+1)},$$

$$\|\mathbf{A}\mathbf{x}^t - \mathbf{b}\| \geq \frac{\sqrt{3}L_A\|\mathbf{x}^*\|}{4\sqrt{2}(t+1)}.$$

Tightness: the lower bounds match with upper bounds [Nesterov'05; X.'17]

Worst-case instance

$$\underset{\mathbf{x}}{\text{minimize}}\, \frac{1}{2}\mathbf{x}^\top \mathbf{H}\mathbf{x} - \mathbf{h}^\top \mathbf{x},\ \text{s.t. } \mathbf{A}\mathbf{x} = \mathbf{b}. \qquad \text{(QP-Inst)}$$

Here,

$$\mathbf{H} = \frac{L_f}{4}\begin{bmatrix}\mathbf{B}^\top\mathbf{B} & \\ & \mathbf{I}_{n-2k}\end{bmatrix} \in \mathbb{R}^{n\times n}, \mathbf{h} = \frac{L_f}{2}\mathbf{e}_{2k,n}, \mathbf{A} = \frac{L_A}{2}\mathbf{\Lambda}, \mathbf{b} = \frac{L_A}{2}\mathbf{c},$$

and

$$\mathbf{\Lambda} = \begin{bmatrix}\mathbf{B} & \mathbf{O}\\ \mathbf{O} & \mathbf{G}\end{bmatrix} \in \mathbb{R}^{m\times n},\ \mathbf{c} = \begin{bmatrix}\mathbf{1}_{2k}\\ \mathbf{0}\end{bmatrix},\ \mathbf{B} := \begin{bmatrix} & & & -1 & 1\\ & & \iddots & \iddots & \\ & -1 & 1 & & \\ & 1 & & & \end{bmatrix} \in \mathbb{R}^{2k\times 2k}$$

with $\mathbf{G} \in \mathbb{R}^{(m-2k)\times(n-2k)}$ is any matrix of full row rank such that $\|\mathbf{G}\| = 2$.

Remark: condition number of $\mathbf{B}$ proportional to k

lower bound for strongly convex case

Setting of problem class:

- given positive integers $m \leq n$, and $t < \frac{m}{2}$
- given positive numbers L_A and μ

Conclusion: there exists an instance of smooth linearly constrained problem such that

- f is smooth and μ-strongly convex, $\|\mathbf{A}\|_2 = L_A$
- it has a unique primal-dual solution $(\mathbf{x}^*, \mathbf{y}^*)$
- in addition, for (Span), it holds

$$\|\mathbf{x}^t - \mathbf{x}^*\|^2 \geq \frac{5L_A^2\|\mathbf{y}^*\|^2}{256\mu^2(t+1)^2}.$$

- Again, higher than the result for unconstrained problems.

Part II: Near-optimal FOMs for problems with $O(1)$ functional constraints

Complexity comparison for unconstrained and constrained problems

Oracle complexity of an optimal FOM to produce an ε-solution $\bar{\mathbf{x}}$ of the problem $F^* = \min_{\mathbf{x}} \Big\{F(\mathbf{x}) \equiv f(\mathbf{x}) + r(\mathbf{x})\Big\}$, i.e., $F(\bar{\mathbf{x}}) - F^* \leq \varepsilon$:

- convex composite case: $O(\sqrt{\frac{L}{\varepsilon}})$ from convergence rate $O(\frac{L}{k^2})$
- strongly-convex composite case: $O(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon})$ from convergence rate $O\Big((1-\sqrt{\mu/L})^k\Big)$

Oracle complexity of the best FOM to produce an ε-solution $\bar{\mathbf{x}} \in X$ of problem $f^* = \min_{\mathbf{x}\in X} \Big\{f(\mathbf{x}),\ \text{s.t. } \mathbf{A}\mathbf{x} = \mathbf{b}\Big\}$, i.e., $|f(\bar{\mathbf{x}}) - f^*| \leq \varepsilon, \|\mathbf{A}\bar{\mathbf{x}} - \mathbf{b}\| \leq \varepsilon$

- convex smooth case: $O(1/\varepsilon)$ [Lan-Monteiro'16, Ouyang et. al'15, Hamedani-Aybat'21, Sabach-Teboulle'20, ...]
- strongly-convex smooth case: $O(1/\sqrt{\varepsilon})$ [Hamedani-Aybat'21, Sabach-Teboulle'20, Xu'21, Lan-Ouyang-Zhou'21, ...]

It is impossible to close the gap **in general**, but possible for special problems.

Problem formulation

$$\begin{aligned} &\min_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}) := f(\mathbf{x}) + h(\mathbf{x}), \\ &\text{s.t. } \mathbf{g}(\mathbf{x}) := [g_1(\mathbf{x}), \ldots, g_m(\mathbf{x})] \leq \mathbf{0}, \end{aligned} \tag{1}$$

- f is L-smooth and μ-strongly convex with $\mu > 0$ (convexity can be relaxed)
- h is closed convex and admits a simple proximal mapping,
- each g_i is convex and smooth.
- **specialty:** m is small

Example I: Neyman-Pearson classification [Scott-Nowak'05]

$$\min_{\mathbf{w}} \frac{1}{N_+} \sum_{i \in \mathcal{N}_+} \ell(\mathbf{w}; \mathbf{x}_i, 1), \text{ s.t. } \frac{1}{N_-} \sum_{i \in \mathcal{N}_-} \ell(\mathbf{w}; \mathbf{x}_i, -1) \leq \gamma \qquad \text{(NPC)}$$

- $\{\mathbf{x}_i\}_{i \in \mathcal{N}_+}$ positive data and $\{\mathbf{x}_i\}_{i \in \mathcal{N}_-}$ negative data
- ℓ is a loss function, e.g., logistic regression loss
- $\gamma > 0$ is a false-positive error level
- Application examples: spam detection and medical diagnosis [Rigollet-Tong'11]

Observation: a single nonlinear constraint but difficult to project onto the feasible set

Example II: fairness-constrained regression [Komiyama et. al'18]

$$\min_{\mathbf{w}_s, \mathbf{w}_u} \mathbf{w}_s^\top \mathbf{V}_s \mathbf{w}_s + \mathbf{w}_u^\top \mathbf{V}_u \mathbf{w}_u - 2\mathbb{E}(y\mathbf{s}^\top \mathbf{w}_s + y\mathbf{u}^\top \mathbf{w}_u)$$

$$\text{s.t. } \frac{\mathbf{w}_s^\top \mathbf{V}_s \mathbf{w}_s}{\mathbf{w}_s^\top \mathbf{V}_s \mathbf{w}_s + \mathbf{w}_u^\top \mathbf{V}_u \mathbf{w}_u} \leq \gamma$$

- $(\mathbf{s}, \mathbf{u}, y)$ denotes one data point
  - $\mathbf{s}$ the sensitive attributes (e.g., gender, race), $\mathbf{u}$ non-sensitive attributes, and y the label
- model derived based on linear prediction $\hat{y} = \mathbf{s}^\top \mathbf{w}_s + \mathbf{u}^\top \mathbf{w}_u$
- $\mathbf{V}_s$ covariance of $\mathbf{s}$ and $\mathbf{V}_u$ covariance of $\mathbf{u}$
- $\gamma \in [0, 1]$ user-specified fairness parameter
  - $\gamma = 0$: completely fair; $\gamma = 1$ fully fairness-ignorant
- equivalent to a convex QCQP with two quadratic inequality constraints

First-order augmented Lagrangian method

Classic augmented Lagrangian function [Rockafellar'73]:

$$\mathcal{L}_\beta(\mathbf{x},\mathbf{z}) = f(\mathbf{x}) + h(\mathbf{x}) + \frac{\beta}{2}\left\|[\mathbf{g}(\mathbf{x}) + \frac{\mathbf{z}}{\beta}]_+\right\|^2 - \frac{\|\mathbf{z}\|^2}{2\beta}.$$

Algorithm 1: First-order inexact augmented Lagrangian method for (1)

1. **Initialization:** choose $\mathbf{x}^0, \mathbf{z}^0$, and $\beta_0 > 0$
2. **for** $k = 0, 1, \ldots$ **do**
3. Apply a first-order method to find $\mathbf{x}^{k+1}$ as an approximate solution of $\min_{\mathbf{x}} \mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{z}^k)$.
4. Update $\mathbf{z}$ by $\mathbf{z}^{k+1} = [\mathbf{z}^k + \beta_k \mathbf{g}(\mathbf{x}^{k+1})]_+$.
5. Choose $\beta_{k+1} \geq \beta_k$.
6. **if** *a stopping condition is satisfied* **then**
7. Output $(\mathbf{x}^{k+1}, \mathbf{z}^{k+1})$ and stop

- [Yan-He'20, Xu'21, ...]: Suppose (1) has a KKT point. Given $\varepsilon > 0$, if $\beta_{k+1} = \sigma\beta_k$ with $\sigma > 1$, then an ε-solution can be found by solving $O(\log \frac{1}{\beta_0 \varepsilon})$ primal subproblems, each to an $O(\varepsilon)$ accuracy.

Using accelerated proximal gradient (APG) as subroutine

- If $\mathrm{dom}(h)$ is bounded, then $\mathcal{L}_\beta(\cdot, \mathbf{z})$ is $O(\beta)$-smooth.
- To find an ε-solution of $\min_{\mathbf{x}} \mathcal{L}_\beta(\mathbf{x}, \mathbf{z})$, it suffices to run $O(\sqrt{\frac{\beta}{\mu}} \log \frac{1}{\varepsilon})$ APG iterations.
- In order to have an ε-solution of (1), β_k needs to increase to $\Theta(\frac{1}{\varepsilon})$.
- Total APG iterations are $O(\varepsilon^{-\frac{1}{2}} \log \frac{1}{\varepsilon} \log \frac{1}{\beta_0 \varepsilon})$, close to the lower bound $\Theta(\varepsilon^{-\frac{1}{2}})$ but worse than $O(\log \frac{1}{\varepsilon})$.

Can we make a first-order subroutine better than the optimal APG?

Key idea to have a better first-order subroutine

Let $\boldsymbol{\theta}(\mathbf{x}) = \mathbf{g}(\mathbf{x}) + \frac{\mathbf{z}}{\beta}$. Then $\min_{\mathbf{x}} \mathcal{L}_\beta(\mathbf{x}, \mathbf{z})$ is equivalent to

$$\min_{\mathbf{x}\in\mathbb{R}^n} \max_{\mathbf{y}\geq\mathbf{0}} \ \Phi(\mathbf{x}, \mathbf{y}) := F(\mathbf{x}) + \beta \left(\mathbf{y}^\top \boldsymbol{\theta}(\mathbf{x}) - \tfrac{1}{2}\|\mathbf{y}\|^2\right).$$

Let $d(\mathbf{y}) = \min_{\mathbf{x}} \Phi(\mathbf{x}, \mathbf{y})$ and solve $\max_{\mathbf{y}\geq\mathbf{0}} d(\mathbf{y})$ to a certain accuracy.

- d is smooth and strongly concave, and $\nabla d(\mathbf{y}) = \beta(\boldsymbol{\theta}(\mathbf{x}(\mathbf{y})) - \mathbf{y})$
- Given $\mathbf{y} \geq \mathbf{0}$, finding δ-solution of $\min_{\mathbf{x}} \Phi(\mathbf{x}, \mathbf{y})$ requires $O(\sqrt{\frac{\beta\|\mathbf{y}\|+L}{\mu}} \log \frac{1}{\delta})$ first-order oracles.

  first key: solution of $\max_{\mathbf{y}} d(\mathbf{y})$ satisfies $\|\bar{\mathbf{y}}\| \leq \frac{2\|\mathbf{z}^*\|+\|\mathbf{z}\|}{\beta}$
- However, accelerated gradient ascent to $\max_{\mathbf{y}\geq\mathbf{0}} d(\mathbf{y})$ has high complexity because the smoothness constant of d is proportional to $\frac{\beta^2}{\mu}$

  second key: cutting planes can be generated by strong-concavity of d

The case with a single nonlinear constraint

Main idea (to approximately solve $\max_{\mathbf{y}\geq\mathbf{0}} d(\mathbf{y})$):

1. Given $\hat{\mathbf{y}} \geq \mathbf{0}$, solve $\min_{\mathbf{x}} \Phi(\mathbf{x}, \hat{\mathbf{y}})$ to have a sufficiently accurate solution $\hat{\mathbf{x}}$.
2. Use $\beta(\boldsymbol{\theta}(\hat{\mathbf{x}}) - \hat{\mathbf{y}})$ as an approximation of $d'(\hat{\mathbf{y}})$.
3. If $|[\boldsymbol{\theta}(\hat{\mathbf{x}})]_+ - \hat{\mathbf{y}}|$ is sufficiently small, accept $\hat{\mathbf{y}}$ as an approximate solution of $\max_{\mathbf{y}\geq\mathbf{0}} d(\mathbf{y})$
4. Otherwise, $d'(\hat{\mathbf{y}})$ has the same sign of $\boldsymbol{\theta}(\hat{\mathbf{x}}) - \hat{\mathbf{y}}$, and thus the bisection method can be applied.

Formal result:

Lemma 8 *Given $\delta > 0$ and $\widehat{\mathbf{y}} \geq \mathbf{0}$, let $\widehat{\mathbf{x}} \in \mathrm{dom}(h)$ be a point satisfying $\mathrm{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\Phi(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})\big) \leq \frac{\mu\delta}{4B_g}$. If $\big|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - \widehat{\mathbf{y}}\big| \leq \frac{3\delta}{4}$, then $|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}| \leq \delta$. Otherwise, $|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}| > \frac{\delta}{2}$, and $\nabla d(\widehat{\mathbf{y}})(\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}) > 0$.*

- B_g is the bound of the Jacobi matrix of $\mathbf{g}$ on $\mathrm{dom}(h)$.

Overall complexity for the case of a single nonlinear constraint

Given a target accuracy $\varepsilon > 0$,

1. $O(\sqrt{\kappa}\log\frac{1}{\varepsilon})$ first-order oracles for approximately solving $\min_{\mathbf{x}} \Phi(\mathbf{x}, \hat{\mathbf{y}})$
2. $O(\log\frac{1}{\varepsilon})$ halves to reduce an interval sufficiently short
3. For every ALM subproblem, $O(\sqrt{\kappa}(\log\frac{1}{\varepsilon})^2)$ first-order oracles suffice
4. Hence, in total $O(\sqrt{\kappa}(\log\frac{1}{\varepsilon})^2\log\frac{1}{\beta_0\varepsilon})$ first-order oracles to produce an ε-solution of (1).

Formal result:

Theorem 8 (Iteration complexity when $m = 1$) *Suppose that Assumptions 1 through 4 hold, and $m = 1$ in (1). Let $(\beta_0, \sigma, \varepsilon, \gamma_1, \gamma_2)$ be the input of Algorithm 7 and $\{(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)\}_{k\geq 0}$ be the generated sequence. Suppose $\bar{\varepsilon} = \min\left\{\varepsilon, \sqrt{\frac{\varepsilon\mu(\sigma-1)}{8\sigma+1}}\right\} \leq \left\{\varepsilon, \frac{24B_g(\mu+\beta_k B_g^2)}{\mu}\right\}, \forall\, k \geq 0$. Let $\varepsilon_k = \bar{\varepsilon}$ for all $k \geq 0$. Then Algorithm 7 needs at most $T_{\text{total}} = O\big(\sqrt{\frac{L_f + L_g(1+\|\mathbf{z}^*\|)}{\mu}}|\log\varepsilon|^3\big)$ evaluations on f, ∇f, $\mathbf{g}$, and $J_{\mathbf{g}}$ to produce an ε-KKT point of (1).*

- One factor $\log\frac{1}{\varepsilon}$ is to search $\hat{\mathbf{y}}$; another factor $\log\frac{1}{\varepsilon}$ is to increase β in ALM.

The case with multiple nonlinear constraints

Main idea (to approximately solve $\max_{\mathbf{y}\geq\mathbf{0}} d(\mathbf{y})$):

1. Given $\hat{\mathbf{y}} \geq \mathbf{0}$, solve $\min_{\mathbf{x}} \Phi(\mathbf{x}, \hat{\mathbf{y}})$ to have a sufficiently accurate solution $\hat{\mathbf{x}}$.
2. Use $\beta(\boldsymbol{\theta}(\hat{\mathbf{x}}) - \hat{\mathbf{y}})$ as an approximation of $\nabla d(\hat{\mathbf{y}})$.
3. If $|[\boldsymbol{\theta}(\hat{\mathbf{x}})]_+ - \hat{\mathbf{y}}|$ is sufficiently small, accept $\hat{\mathbf{y}}$ as an approximate solution of $\max_{\mathbf{y}\geq\mathbf{0}} d(\mathbf{y})$
4. Otherwise, $\langle \boldsymbol{\theta}(\hat{\mathbf{x}}) - \hat{\mathbf{y}}, \mathbf{y} - \hat{\mathbf{y}} \rangle \geq 0$ for all $\mathbf{y}$ near the solution of $\max_{\mathbf{y}\geq\mathbf{0}} d(\mathbf{y})$, and thus a cutting-plane method can be applied.

Formal result:

LEMMA 3.11. *Let $b > 0$, and suppose $\|\bar{\mathbf{y}}\| \leq b$. Given $\delta > 0$ and $\widehat{\mathbf{y}} \geq \mathbf{0}$, let $\widehat{\mathbf{x}} \in \mathrm{dom}(h)$ be a point satisfying $\mathrm{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\Phi(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})\big) \leq \min\{\frac{\mu\delta}{4B_g}, \frac{\mu^2\delta}{8B_g(\mu+\beta B_g^2)}\}$. If $\|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - \widehat{\mathbf{y}}\| \leq \frac{3\delta}{4}$, then $\|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}\| \leq \delta$. Otherwise, $\big\|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}\big\| > \frac{\delta}{2}$, and also $\langle \boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}, \mathbf{y} - \widehat{\mathbf{y}} \rangle > 0$ for any $\mathbf{y} \in \mathcal{B}_\eta(\bar{\mathbf{y}}) \cap \mathcal{B}_b^+$, where $\eta = \min\{b, \eta_+\}$, and η_+ is the positive root of the equation*

$$\frac{\mu+\beta B_g^2}{\mu}\left(\eta + \sqrt{\frac{2\eta B_d}{\beta}}\right) = \frac{\delta}{4}, \quad \text{with} \quad B_d = \max_{\mathbf{y}\in\mathcal{B}_b^+} \nabla d(\mathbf{y}). \tag{3.13}$$

Overall complexity for the case of multiple nonlinear constraints

Given a target accuracy $\varepsilon > 0$,

1. $O(\sqrt{\kappa}\log\frac{1}{\varepsilon})$ first-order oracles for approximately solving $\min_{\mathbf{x}} \Phi(\mathbf{x}, \hat{\mathbf{y}})$
2. $O(m\log\frac{1}{\varepsilon})$ cutting planes by the volumetric-center [Vaidya'96] cutting-plane method to reduce a polytope sufficiently small
3. For every subproblem, $O(m\sqrt{\kappa}(\log\frac{1}{\varepsilon})^2)$ first-order oracles suffice
4. Hence, in total $O(m\sqrt{\kappa}(\log\frac{1}{\varepsilon})^2\log\frac{1}{\beta_0\varepsilon})$ first-order oracles to produce an ε-solution of (1).

Formal result:

THEOREM 4.2 (oracle complexity). *Suppose that Assumptions 1–4 hold. Let* $(\beta_0, \sigma, \varepsilon, \gamma_1, \gamma_2)$ *be the input of Algorithm 8 and* $\{(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)\}_{k\geq 0}$ *be the generated sequence. Suppose* $\bar{\varepsilon} = \min\left\{\varepsilon, \sqrt{\frac{\varepsilon\mu(\sigma-1)}{8\sigma+1}}\right\} \leq \left\{\varepsilon, \frac{24B_g(\mu+\beta_k B_g^2)}{\mu}\right\}$ *for all* $k \geq 0$. *Let* $\varepsilon_k = \bar{\varepsilon}$ *for all* $k \geq 0$. *Then, to produce an* ε*-KKT point of* (1.1), *Algorithm 8 needs at most* $T_{\text{total}} = O\big(m\sqrt{\frac{L_f+L_g(1+\|\mathbf{z}^*\|)}{\mu}}|\log\varepsilon|^2(\log m + |\log\varepsilon|)\big)$ *evaluations on* f, ∇f, $\mathbf{g}$, *and* $J_{\mathbf{g}}$.

- Cutting-plane based FOM can be better than APG-based FOM in the regime of $m = o(\frac{1}{\sqrt{\varepsilon}})$.

Extension to convex cases

Suppose that f is convex in (1). Apply the previous described method to the perturbed problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) + \tfrac{\varepsilon}{4D_h}\|\mathbf{x}-\mathbf{x}^0\|^2 + h(\mathbf{x}),$$
$$\text{s.t. } \mathbf{g}(\mathbf{x}) := [g_1(\mathbf{x}),\ldots,g_m(\mathbf{x})] \le \mathbf{0},$$

or apply a proximal augmented Lagrangian method with a cutting-plane based first-order method for solving each subproblem.

- total number of first-order oracles: $\tilde{O}(m/\sqrt{\varepsilon})$ as compared to $O(1/\varepsilon)$.

Extension to problems with nonconvex objective but convex constraints

Suppose that f is nonconvex L-smooth in (1). Apply the previous described method in the framework of proximal point method to each subproblem

$$\bar{\mathbf{x}}^{k+1} \approx \arg\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) + L_f\|\mathbf{x} - \bar{\mathbf{x}}^k\|^2 + h(\mathbf{x}),$$
$$\text{s.t. } \mathbf{g}(\mathbf{x}) := [g_1(\mathbf{x}), \ldots, g_m(\mathbf{x})] \leq \mathbf{0},$$

- total number of first-order oracles: $\tilde{O}(m/\varepsilon^2)$ as compared to $O(1/\varepsilon^{2.5})$.

Experiments on QCQP

$$\begin{aligned} &\min_{\mathbf{x}\in\mathbb{R}^n} \tfrac{1}{2}\mathbf{x}^\top \mathbf{Q}_0 \mathbf{x} + \mathbf{x}^\top \mathbf{c}_0, \\ &\text{s.t. } \tfrac{1}{2}\mathbf{x}^\top \mathbf{Q}_j \mathbf{x} + \mathbf{x}^\top \mathbf{c}_j + d_j \leq 0,\ j = 1, \ldots, m; \\ &\quad x_i \in [l_i, u_i],\ i = 1, \ldots, n. \end{aligned} \tag{QCQP}$$

- $\mathbf{Q}_0$ is generated to be positive definite
- $\mathbf{Q}_j$ is positive semidefinite but rank-deficient for each $j = 1, \ldots, m$
- $l_i = -10$ and $u_i = 10$ for each i
- All d_j are negative so the Slater's condition holds.

Numerical results for $m = 1, n = 1000$

		APG-based iALM					proposed cutting-plane iALM			
out.Iter	β	#grad	#func	pres	dres	compl	#grad	pres	dres	compl
trial 1		total running time = 774.2 sec.					total running time = 12.4 sec.			
1	1	5056	9420	5.13e-02	9.65e-05	2.63e-03	2136	5.13e-02	6.40e-11	2.64e-03
2	10	16802	31298	1.65e-06	9.46e-05	8.46e-08	1434	4.23e-07	9.20e-11	2.17e-08
3	10^2	55359	103112	5.40e-08	9.77e-05	2.77e-09	1068	4.22e-10	2.63e-10	2.17e-11
4	10^3	179877	335030	6.51e-09	9.96e-05	3.34e-10	1080	0.00e+00	1.74e-08	4.84e-11
5	10^4	584145	1087988	0.00e+00	9.95e-05	4.57e-11	1104	2.29e-11	9.23e-09	1.17e-12
trial 2		total running time = 760.0 sec.					total running time = 12.1 sec.			
1	1	4969	9258	5.78e-02	9.78e-05	3.34e-03	1926	5.78e-02	4.94e-09	3.34e-03
2	10	16466	30672	2.10e-06	9.99e-05	1.21e-07	1440	5.85e-07	3.41e-10	3.38e-08
3	10^2	54617	101730	4.57e-08	9.85e-05	2.64e-09	1050	0.00e+00	7.90e-09	4.03e-10
4	10^3	177171	329990	6.44e-09	9.93e-05	3.72e-10	1074	0.00e+00	2.18e-07	1.42e-10
5	10^4	580377	1080970	0.00e+00	1.00e-04	4.06e-11	1104	2.75e-10	1.84e-09	1.59e-11
trial 3		total running time = 780.9 sec.					total running time = 12.4 sec.			
1	1	5100	9502	4.37e-02	9.66e-05	1.91e-03	2088	4.37e-02	2.53e-09	1.91e-03
2	10	17035	31732	0.00e+00	9.33e-05	8.08e-08	1428	4.34e-07	7.52e-09	1.90e-08
3	10^2	56348	104954	1.43e-07	9.79e-05	6.25e-09	1092	0.00e+00	2.75e-13	2.36e-10
4	10^3	182583	340070	0.00e+00	9.63e-05	5.12e-10	1122	4.33e-09	4.76e-07	1.89e-10
5	10^4	595012	1108228	1.81e-10	9.99e-05	7.92e-12	1164	0.00e+00	1.88e-09	2.01e-11

- For each subproblem, both methods use a random generated starting point.

Observation: cutting-plane based iALM requires far fewer gradient evaluations

Numerical results for $m = 2, n = 1000$

		APG based iALM					proposed cutting-plane iALM			
out.Iter	β	#grad	#func	pres	dres	compl	#grad	pres	dres	compl
trial 1		total running time = 1348.0 sec.					total running time = 51.0 sec.			
1	1	5551	10342	4.45e-02	8.71e-05	1.40e-03	3342	4.45e-02	1.06e-09	1.40e-03
2	10	18330	34144	0.00e+00	9.62e-05	6.47e-08	3384	3.19e-07	9.17e-09	9.98e-09
3	10^2	59680	111160	8.81e-08	9.77e-05	2.71e-09	3522	6.01e-09	9.15e-10	2.44e-10
4	10^3	194236	361774	0.00e+00	9.94e-05	9.15e-11	3582	1.36e-10	3.84e-09	6.17e-12
5	10^4	629359	1172200	0.00e+00	9.99e-05	7.65e-12	3678	2.66e-11	1.60e-09	8.13e-13
trial 2		total running time = 1299.4 sec.					total running time = 49.5 sec.			
1	1	5362	9990	6.60e-02	9.05e-05	3.10e-03	3180	6.60e-02	8.27e-09	3.10e-03
2	10	17646	32870	2.74e-06	9.26e-05	1.34e-07	3282	6.17e-07	2.67e-10	2.91e-08
3	10^2	57832	107718	1.41e-08	9.82e-05	2.79e-09	3372	5.91e-10	9.05e-11	2.61e-11
4	10^3	187544	349310	0.00e+00	9.88e-05	2.70e-10	3450	4.97e-10	6.76e-09	2.34e-11
5	10^4	606432	1129498	9.88e-11	9.97e-05	7.38e-12	3528	1.82e-11	5.23e-09	1.95e-12
trial 3		total running time = 1337.1 sec.					total running time = 49.2 sec.			
1	1	5464	10180	5.50e-02	9.51e-05	2.25e-03	3156	5.50e-02	6.27e-09	2.25e-03
2	10	18039	33602	1.78e-06	9.90e-05	8.15e-08	3324	5.16e-07	1.76e-10	2.07e-08
3	10^2	59505	110834	2.88e-08	9.95e-05	1.86e-09	3384	5.93e-09	8.30e-09	2.49e-10
4	10^3	192301	358170	3.78e-09	9.99e-05	1.45e-10	3504	0.00e+00	1.02e-09	3.00e-11
5	10^4	627235	1168244	6.81e-11	1.00e-04	9.17e-12	3528	5.23e-11	2.78e-09	1.70e-12

- For each subproblem, both methods use a random generated starting point.

Observation: cutting-plane based iALM requires far fewer gradient evaluations but scales worse to m.

Numerical results for $m = 5, n = 1000$

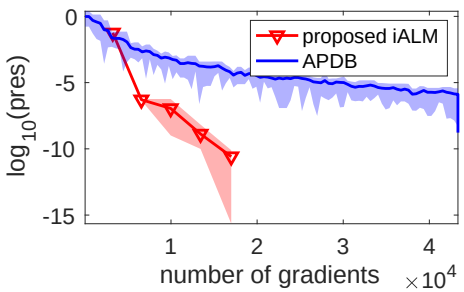
		APG based iALM					proposed cutting-plane iALM			
out.Iter	β	#grad	#func	pres	dres	compl	#grad	pres	dres	compl
trial 1		total running time = 2833.1 sec.					total running time = 156.8 sec.			
1	1	5537	10316	7.93e-02	9.91e-05	2.90e-03	6714	7.93e-02	2.91e-09	2.90e-03
2	10	18417	34306	1.12e-06	9.83e-05	4.28e-08	6984	8.93e-07	4.32e-09	3.27e-08
3	10^2	60058	111864	5.83e-08	9.62e-05	2.25e-09	7158	4.64e-09	1.50e-09	2.02e-10
4	10^3	195894	364862	3.14e-09	9.88e-05	1.64e-10	7314	4.37e-10	4.28e-09	1.64e-11
5	10^4	640357	1192684	9.40e-10	9.97e-05	3.51e-11	7614	2.79e-11	8.77e-09	1.74e-12
trial 2		total running time = 2786.0 sec.					total running time = 160.7 sec.			
1	1	5537	10316	6.77e-02	8.21e-05	2.42e-03	6900	6.77e-02	6.16e-09	2.42e-03
2	10	18170	33846	6.24e-07	9.21e-05	2.43e-08	7110	7.39e-07	2.64e-09	2.75e-08
3	10^2	59607	111024	2.66e-08	9.73e-05	1.71e-09	7224	2.81e-09	9.46e-09	1.90e-10
4	10^3	194483	362234	1.21e-08	9.99e-05	3.19e-10	7512	6.61e-10	4.34e-09	2.53e-11
5	10^4	636109	1184772	7.58e-11	9.94e-05	1.76e-11	7698	3.94e-11	7.84e-09	1.73e-12
trial 3		total running time = 2820.0 sec.					total running time = 155.3 sec.			
1	1	5595	10424	8.47e-02	8.51e-05	3.26e-03	6594	8.47e-02	9.82e-09	3.26e-03
2	10	18461	34388	7.78e-07	9.55e-05	3.07e-08	6882	8.64e-07	5.52e-09	3.33e-08
3	10^2	60422	112542	3.78e-09	9.93e-05	4.10e-09	7116	3.42e-09	1.52e-10	1.83e-10
4	10^3	196869	366678	7.70e-09	9.87e-05	3.05e-10	7260	7.35e-11	5.28e-09	1.91e-11
5	10^4	640997	1193876	3.63e-10	9.95e-05	1.37e-11	7488	6.86e-11	6.05e-09	2.72e-12

- For each subproblem, both methods use a random generated starting point.

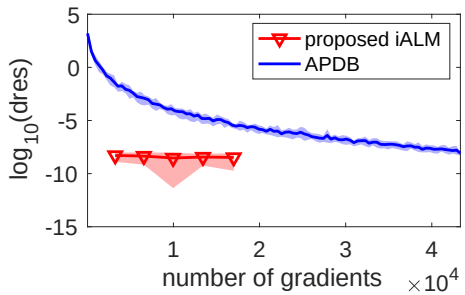
Observation: cutting-plane based iALM requires far fewer gradient evaluations but scales almost linearly to m.

Comparison to an accelerated primal-dual method [Hamedani-Aybat'21]

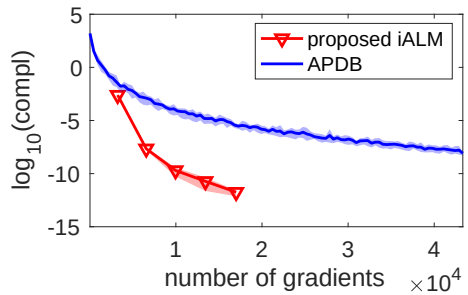
primal residual



dual residual



complementarity



- QCQP with $m = 2$ constraints
- A 10^{-8}-KKT point is targeted
- 10 independent trials performed

Observation: the proposed method uses fewer gradients.

Comparison to an interior-point method

	Proposed cutting-plane iALM					SDPT3			
Trial	Time(h:m:s)	#grad	pres	dres	compl	Time(h:m:s)	pres	dres	compl
Problem size: $m = 2, n = 1000$									
1	0:0:35	16776	0.00e+00	1.13e-10	3.42e-12	0:0:11	3.30e-10	1.03e-09	4.12e-11
2	0:0:36	16812	0.00e+00	1.89e-09	8.75e-13	0:0:16	2.14e-10	4.40e-10	9.25e-12
3	0:0:35	17004	4.09e-11	1.19e-09	1.91e-12	0:0:11	0.00e+00	2.04e-09	8.31e-11
4	0:0:36	16698	3.53e-11	2.69e-09	2.27e-12	0:0:11	0.00e+00	8.00e-09	1.61e-08
5	0:0:35	16578	2.32e-11	3.19e-09	2.77e-12	0:0:17	1.58e-09	8.16e-10	9.10e-11
Problem size: $m = 2, n = 5000$									
1	0:11:9	21630	2.58e-11	5.85e-10	1.24e-12	0:40:44	0.00e+00	8.26e-09	5.71e-10
2	0:11:11	21642	3.58e-11	9.17e-10	1.63e-12	0:52:23	6.55e-08	1.18e-09	2.84e-09
3	0:11:6	21504	1.95e-11	6.10e-10	7.19e-13	0:50:39	5.45e-08	NaN	NaN
4	0:11:12	21678	3.13e-11	4.67e-09	1.04e-12	0:40:38	0.00e+00	1.12e-08	1.59e-09
5	0:11:7	21516	1.99e-11	8.59e-09	9.04e-13	0:36:17	2.71e-08	1.10e-08	1.28e-09
Problem size: $m = 2, n = 10000$									
1	1:16:1	22332	0.00e+00	6.32e-10	2.37e-13	5:55:22	2.41e-07	3.10e-08	1.33e-08
2	1:8:33	22296	4.99e-12	6.36e-09	1.30e-12	6:20:3	0.00e+00	4.60e-10	3.19e-09
3	0:58:16	22296	1.73e-11	2.54e-09	7.43e-13	6:13:5	0.00e+00	3.44e-08	8.17e-09
4	0:58:9	22368	2.05e-11	1.14e-08	9.17e-13	7:3:39	0.00e+00	2.16e-08	7.70e-09
5	1:15:19	22182	7.95e-12	1.04e-08	1.30e-12	8:31:16	0.00e+00	3.70e-08	1.48e-09

Observation: the proposed method performs worse for small-size instances but better for large-size ones.

Conclusions

- Gave a lower complexity bound result of first-order methods for solving affine-constrained convex smooth problems
  - The bound is higher than that for unconstrained problems
- Presented a cutting-plane based first-order method for solving problems with a few nonlinear convex constraints
  - When there are $O(1)$ constraints, the oracle complexity is in almost the same order as an optimal method for solving unconstrained *convex or nonconvex* problems
  - Demonstrated the performance of the proposed method on QCQP

References

1. Y. Ouyang and **Y. Xu**. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. Mathematical Programming, Series A, 185, 1-35, 2021.
2. **Y. Xu**. First-order methods for problems with O(1) functional constraints can have almost the same convergence rate as for unconstrained problems. SIAM J. Optimization, forthcoming. arXiv:2010.02282

Thank you!!!