Affine-invariant contracting-point methods for Convex Optimization

Nikita Doikov

Joint work with Yurii Nesterov

UCLouvain, Belgium

Workshop on Advances in Continuous Optimization, EUROPT
July 29, 2022

Plan of the Talk

1. Introduction

2. Contracting-Point Methods

3. Inexact and Stochastic Contracting Newton

4. Conclusions

Composite Optimization Problem

$$\min_x \Big\{ F(x) \overset{\text{def}}{=} f(x) + \psi(x) \Big\}$$

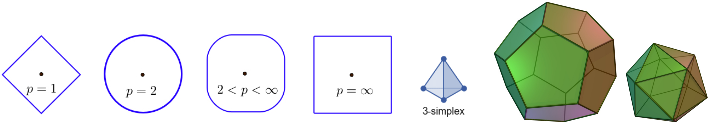
- f is convex and several times differentiable (the *difficult part*).
- $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a *simple* convex function.
- We assume that the domain of ψ,

$$\mathrm{dom}\,\psi \overset{\text{def}}{=} \Big\{ x \in \mathbb{R}^n \,:\, \psi(x) < +\infty \Big\},$$

  is **bounded**.

Example: Indicator of a Set

1. Let $Q \subset \mathbb{R}^n$ be a simple bounded convex set.



We can use

$$\psi(x) \quad = \quad \mathrm{Ind}_Q(x) \quad := \quad \begin{cases} 0, & x \in Q \\ +\infty, & \text{otherwise.} \end{cases}$$

$\Rightarrow$ Then our problem is $\boxed{\min_{x \in Q} f(x)}$

2. Let

$$\psi(x) \;=\; \begin{cases} \lambda\|x\|_1, & x \in Q \\ +\infty, & \text{otherwise.} \end{cases}$$

$\Rightarrow$ Adding ℓ_1-Regularizer to the problem:

$$\min_{x\in Q} f(x) + \lambda\|x\|_1.$$

Enforce solutions to be sparse.

Let $\nabla f(x)$ be Lipschitz continuous: $\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|$

The Gradient Method [Cauchy, 1847]:

$$x_{k+1} = \operatorname*{argmin}_y \left\{ f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \tfrac{L}{2}\|y - x_k\|^2 + \psi(y) \right\}$$

- The method depends on the norm $\|\cdot\|$
- Global convergence: $F(x_k) - F^* \leq O(\frac{1}{k})$

The Conditional Gradient Method [Frank-Wolfe, 1956]:

$$\begin{aligned} v_{k+1} &= \operatorname*{argmin}_y \left\{ f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \psi(y) \right\}, \\ x_{k+1} &= \gamma_k v_{k+1} + (1 - \gamma_k) x_k \end{aligned}$$

- Set $\gamma_k = \frac{2}{k+2}$. Then $F(x_k) - F^* \leq O(\frac{1}{k})$

Note: Near-optimal for $\|\cdot\|_\infty$-balls [Guzmán-Nemirovski, 2015]

The Newton's Method:

[Newton, 1669; Raphson, 1690; Fine-Bennett, 1916; Kantorovich, 1948]

$$x_{k+1} = \operatorname*{argmin}_y \Big\{ \langle \nabla f(x_k), y - x_k \rangle + \tfrac{1}{2} \langle \nabla^2 f(x_k)(y - x_k), y - x_k \rangle + \psi(y) \Big\}$$

If $\psi(x) \equiv 0$, then

$$x_{k+1} = x_k - \big(\nabla^2 f(x_k)\big)^{-1} \nabla f(x_k)$$

- Quadratic convergence $\mathcal{O}(\log \log \frac{1}{\varepsilon})$, if $\nabla^2 f(x^*) \succ 0$ and x_0 close to x^*
- No global convergence. A heuristic: use line-search in practice
- The method is affine-invariant (it does not use any norms)

This Work

The goal: to develop second- and high-order algorithms with global convergence guarantees

- The rate of second-order methods should be better than that of first-order methods

We propose a general framework of Contracting-Point Methods

- New affine-invariant algorithms of different order $p \geq 1$
- We prove: $F(x_k) - F^* \leq \mathcal{O}(1/k^p)$

Plan of the Talk

1. Introduction

2. Contracting-Point Methods

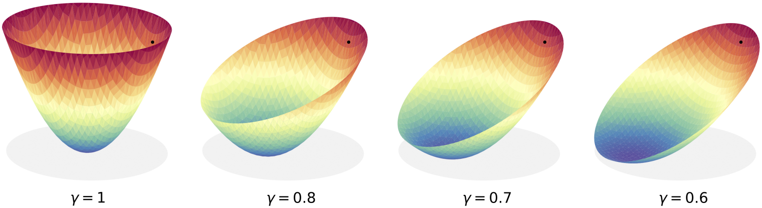
3. Inexact and Stochastic Contracting Newton

4. Conclusions

Contraction Technique

Let us consider contraction of the objective:

$$g(x) \quad := \quad f(\gamma x + (1-\gamma)\bar{x}), \qquad \gamma \in [0,1].$$



Note:

$$\nabla g(x) \quad = \quad \gamma \nabla f(\gamma x + (1-\gamma)\bar{x}),$$

$$\nabla^2 g(x) \quad = \quad \gamma^2 \nabla^2 f(\gamma x + (1-\gamma)\bar{x}),$$

$$\ldots$$

Smoothness properties of $g(\cdot)$ are better than that of $f(\cdot)$

Idea: use γ to balance the error of $g(x) \approx f(x)$ and smoothness

Conceptual Contracting-Point Method. Iterate, $k \geq 0$:

$$\boxed{\begin{array}{rcl} v_{k+1} & \approx & \underset{x}{\operatorname{argmin}} \Big\{ f(\gamma_k x + (1-\gamma_k) x_k) + \gamma_k \psi(x) \Big\}, \\ x_{k+1} & = & \gamma_k v_{k+1} + (1-\gamma_k) x_k \end{array}}$$

- Denote $F_k(x) \stackrel{\text{def}}{=} f(\gamma_k x + (1-\gamma_k)x_k) + \gamma_k \psi(x)$.

Lemma. Let v_{k+1} be an approximate minimizer of $F_k(\cdot)$:

$$F_k(v_{k+1}) - F_k^* \quad \leq \quad \delta_{k+1}.$$

Then

$$F(x_{k+1}) \quad \leq \quad (1-\gamma_k) F(x_k) + \gamma_k F^* + \delta_{k+1}.$$

- If $\gamma_k \to 0$ with an appropriate rate, and δ_{k+1} are small, we have global convergence

Affine-Invariant Smoothness Condition

Fix $p \geq 1$. For a bounded convex set Q, denote

$$\mathcal{V}_Q^{(p+1)}(f) \stackrel{\mathrm{def}}{=} \sup_{x,y,v \in Q} \left| D^{p+1} f(y)[v-x]^{p+1} \right|.$$

Note: for a fixed norm, we have $\mathcal{V}_Q^{(p+1)}(f) \leq L_p (\mathrm{diam}\, Q)^{p+1}$, where L_p is the Lipschitz constant for pth derivative.

It holds, $\forall x, x_k \in Q$ and $\forall \gamma_k \in (0,1]$:

$$\left| f(\gamma_k x + (1-\gamma_k) x_k) - f(x_k) - \sum_{i=1}^{p} \frac{\gamma_k^i}{i!} D^i f(x_k)[x - x_k]^i \right|$$

$$\leq \frac{\gamma_k^{p+1}}{(p+1)!} \mathcal{V}_Q^{(p+1)}(f) \equiv \delta_{k+1} \qquad \text{(Taylor's Theorem).}$$

Contracting-Point Tensor Method:

$$
\boxed{\begin{array}{rcl} v_{k+1} & = & \operatorname*{argmin}_x \Big\{ \sum\limits_{i=1}^{p} \frac{\gamma_k^i}{i!} D^i f(x_k)[x - x_k]^i + \gamma_k \psi(x) \Big\}, \\ x_{k+1} & = & \gamma_k v_{k+1} + (1 - \gamma_k) x_k \end{array}}
$$

Since $\mathrm{dom}\,\psi$ is bounded, the subproblem is well-defined.

Theorem. Set $\gamma_k := \frac{p+1}{k+p+1}$. Then $F(x_k) - F^* \leq O\Big(\frac{\mathcal{V}^{(p+1)}_{\mathrm{dom}\,\psi}(f)}{k^p}\Big)$

- $p = 1$: The Conditional Gradient Method [Frank-Wolfe, 1956]
- $p = 2$: Contracting Newton **(new)**

Plan of the Talk

1. Introduction

2. Contracting-Point Methods

3. Inexact and Stochastic Contracting Newton

4. Conclusions

Contracting Newton Method

- $p = 2$: **Contracting Newton**

$$
\begin{aligned}
v_{k+1} &= \operatorname*{argmin}_x \Big\{ \langle \nabla f(x_k), x - x_k \rangle + \tfrac{\gamma_k}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle \\
&\qquad + \psi(x) \Big\}, \\
x_{k+1} &= \gamma_k v_{k+1} + (1 - \gamma_k) x_k
\end{aligned}
$$

- $F(x_k) - F^* \leq \mathcal{O}(1/k^2)$.
- Acceleration of the Conditional Gradient Method by employing second-order information

Contracting Newton Method (reformulation):

$$x_{k+1} = \operatorname*{argmin}_y \Big\{ \langle \nabla f(x_k), y - x_k \rangle + \tfrac{1}{2}\langle \nabla^2 f(x_k)(y - x_k), y - x_k \rangle + \gamma_k \psi(x_k + \tfrac{1}{\gamma_k}(y - x_k)) \Big\}$$

- $\gamma_k = 1$: The classical Newton's Method
- Interpretation: regularization of quadratic model by the assymmetric trust region

If $\psi(x) = \mathrm{Ind}_Q(x)$, where $Q = \{x \in \mathbb{R}^n \ : \ \|x\| \leq \frac{D}{2}\}$ is the ball, we can use techniques developed for Trust-Region methods [Conn-Gould-Toint, 2000].

Inexact Contracting Newton

Let $\psi(x) = \mathrm{Ind}_Q(x)$ for an arbitrary bounded convex set Q.

$$x_{k+1} = \operatorname*{argmin}_y \Big\{ \langle \nabla f(x_k), y - x_k \rangle + \frac{1}{2}\langle \nabla^2 f(x_k)(y - x_k), y - x_k \rangle \; : \; y \in x_k + \gamma_k (Q - x_k) \Big\}$$

How to compute the iteration?

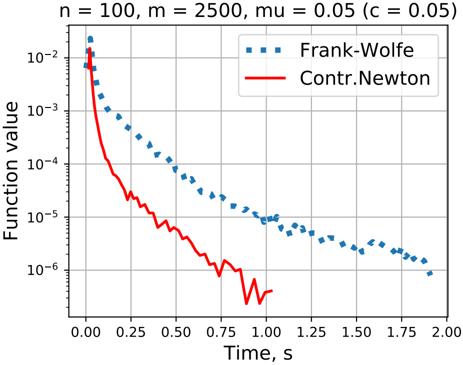
- We can solve the subproblem inexactly by the first-order Frank-Wolfe algorithm
- We have full control over the required accuracy

Theorem. To reach $F(x_K) - F^* \leq \varepsilon$ it needs

- $K = \mathcal{O}\big(\frac{1}{\sqrt{\varepsilon}}\big)$ oracle calls for f
- $\mathcal{O}(\frac{1}{\varepsilon})$ linear minimization oracle calls for ψ totally

Experiment: Log-sum-exp over the Simplex

$$\min_{x \in \mathbb{R}^n_+} \left\{ f(x) \;=\; \mu \log\Big(\sum_{i=1}^m e^{(\langle a_i, x\rangle - b_i)/\mu} \Big) \;:\; \sum_{i=1}^n x^{(i)} = 1 \right\}$$



two times faster

Finite-sum minimization: $f(x) = \frac{1}{M} \sum_{i=1}^{M} f_i(x)$.

- M can be very big in modern applications (several millions).
- Machine Learning: M is the size of the dataset.

It is expensive to compute the full gradient and Hessian:

$$\nabla f(x) = \frac{1}{M} \sum_{i=1}^{M} \nabla f_i(x), \qquad \nabla^2 f(x) = \frac{1}{M} \sum_{i=1}^{M} \nabla^2 f_i(x).$$

Random estimators:

$$\nabla f(x_k) \approx g_k := \frac{1}{m_k^g} \sum_{i \in S_k^g} \nabla f_i(x_k),$$
$$\nabla^2 f(x_k) \approx H_k := \frac{1}{m_k^H} \sum_{i \in S_k^H} \nabla^2 f_i(x_k).$$

$S_k^g, S_k^H \subseteq \{1, \ldots, M\}$ are random subsets (sampled uniformly) for a fixed batchsize $m_k^g = |S_k^g|$, and $m_k^H = |S_k^H|$.

$$
\begin{aligned}
g_k &:= \frac{1}{m_k^g} \sum_{i \in S_k^g} \nabla f_i(x_k), \\
H_k &:= \frac{1}{m_k^H} \sum_{i \in S_k^H} \nabla^2 f_i(x_k).
\end{aligned}
$$

Stochastic Contracting Newton:

$$
\begin{aligned}
x_{k+1} = \operatorname*{argmin}_y \Big\{ & \langle g_k, y - x_k \rangle + \tfrac{1}{2}\langle H_k(y - x_k), y - x_k \rangle \\
& + \gamma_k \psi(x_k + \tfrac{1}{\gamma_k}(y - x_k)) \Big\}
\end{aligned}
$$

Theorem. At iteration k, set $m_k^g = (1+k)^4$, $m_k^H = (1+k)^2$. Then,

$$
\mathbb{E}\big[F(x_k) - F^*\big] \leq \mathcal{O}(1/k^2).
$$

- Idea: at some iterations, recompute the full gradient [Schmidt-Roux-Bach, 2017]

$$\hat{g}_k := \frac{1}{m_k^g} \sum_{i \in S_k^g} (\nabla f_i(x_k) - \nabla f_i(z_k) + \nabla f(z_k)),$$

$$H_k := \frac{1}{m_k^H} \sum_{i \in S_k^H} \nabla^2 f_i(x_k),$$

where z_k is being updated not often.

$$z_k := x_{\pi(k)}, \qquad \pi(k) \stackrel{\text{def}}{=} \begin{cases} 2^{\lfloor \log_2 k \rfloor}, & k > 0 \\ 0, & k = 0. \end{cases}$$

- During N iterations, we recompute the full gradient only $\log_2 N$ times.

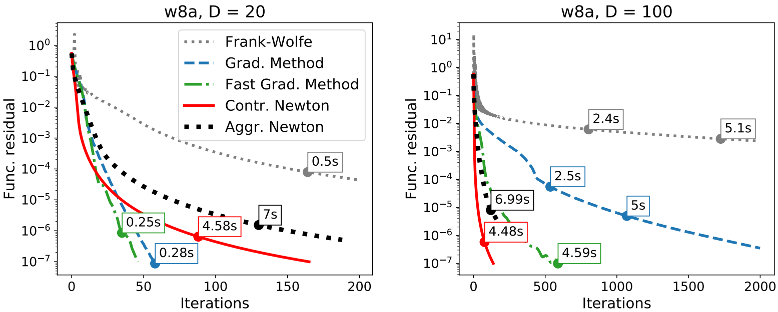
Theorem. It is enough to set $m_k^g = m_k^H = (1+k)^2$. Then we have

$$\mathbb{E}\big[F(x_k) - F^*\big] \leq \mathcal{O}(1/k^2).$$

Experiments: Logistic Regression

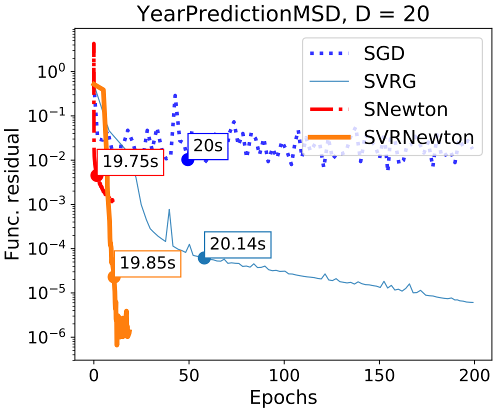
$$\min_{\|x\|_2 \leq \frac{D}{2}} \sum_{i=1}^{M} f_i(x), \qquad f_i(x) = \log\big(1 + \exp(\langle a_i, x\rangle)\big)$$

D plays the role of regularization parameter



For bigger D the problem becomes more *ill-conditioned*

Approximate $\nabla f(x)$, $\nabla^2 f(x)$ by stochastic estimates



The problem with big dataset size ($M = 463715$) and small dimension ($n = 90$)

Plan of the Talk

1. Introduction

2. Contracting-Point Methods

3. Inexact and Stochastic Contracting Newton

4. Conclusions

Conclusions

Using the contraction of the objective

$$g_k(x) \quad := \quad f(\gamma_k x + (1 - \gamma_k) x_k),$$

we are able to construct new algorithms for Convex Optimization, endowed with the global complexity bounds.

1. First-order Taylor's approximation ⇒ Frank-Wolfe algorithm
2. Second-order approximation ⇒ Contracting Newton Method

- The methods are affine-invariant (do not depend on a norm).
- There is a complementary *Proximal-Point approach*:

$$g_k(x) \quad := \quad f(x) + \frac{\alpha_k}{2}\|x - x_k\|^2.$$

Open Questions

- Lower complexity bounds?

Note: Frank-Wolfe algorithm is near-optimal for $\| \cdot \|_\infty$-balls [Guzmán-Nemirovski, 2015]

- Implementation for $p \geq 3$ (the subproblem is not convex)?

Third-order Proximal-type Tensor Methods admits effective implementation [Grapiglia-Nesterov, 2019]

- Variance reduction for the Hessian

References

Nikita Doikov and Yurii Nesterov. "Convex optimization based on global lower second-order models". In: *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020)

Nikita Doikov and Yurii Nesterov. "Affine-invariant contracting-point methods for convex optimization". In: *Mathematical Programming* (2022), pp. 1–23

Thank you for your attention!