Affine-invariant contracting-point methods for Convex Optimization

Nikita Doikov

Joint work with Yurii Nesterov

UCLouvain, Belgium

Symposium on Numerical Analysis and Optimization, UFPR
March 4, 2021

Plan of the Talk

1. Introduction

2. Contracting-Point Methods

3. Inexact and Stochastic Algorithms

4. Conclusions

Plan of the Talk

1. Introduction
2. Contracting-Point Methods
3. Inexact and Stochastic Algorithms
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

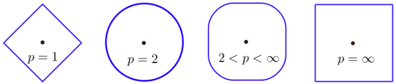
1. Let $Q \subset \mathbb{R}^n$ be a simple bounded convex set. We can use

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

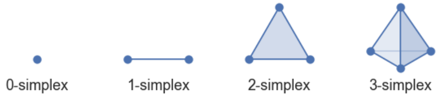
$\Rightarrow$ Then our problem is $\boxed{\min\limits_{x \in Q} f(x)}$

Examples of sets:

- The Ball: $Q = \left\{ x \in \mathbb{R}^n \ : \ \|x\|_p := \left(\sum_{i=1}^n |x^{(i)}|^p\right)^{1/p} \leq \frac{D}{2} \right\}$.



- The standard Simplex: $Q = \mathbb{S}_n := \left\{ x \in \mathbb{R}^n_+ \ : \ \sum_{i=1}^n x^{(i)} = 1 \right\}$.



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.

- Machine Learning, Statistics. Empirical Risk Minimization:

$$f(x) \;=\; \tfrac{1}{M}\sum_{i=1}^{M} f_i(x).$$

Generalized Linear Models: $f_i(x) = \phi(\langle a_i, x\rangle - b_i)$, where $\{a_i, b_i\}$ is given data, ϕ is a convex loss.

Logistic Regression: $\phi(x) = \log(1 + e^x)$.

- Smooth Approximation of a Nondifferentiable Function.

$$f(x) \;\approx\; \bar{f}(x) \;=\; \max_{i=1}^{m}\{\langle a_i, x\rangle - b_i\} \quad \text{(pointwise maximum).}$$

Log-sum-exp (SoftMax): $f(x) = \mu \log\left(\sum_{i=1}^{m} e^{(\langle a_i, x\rangle - b_i)/\mu}\right)$.

The Gradient Method:

$$\boxed{x_{k+1} \;=\; \underset{y}{\operatorname{argmin}}\Big\{\langle \nabla f(x_k), y - x_k\rangle + \tfrac{L}{2}\|y - x_k\|^2 + \psi(y)\Big\}.}$$

Note: when $\psi(x) = \mathrm{Ind}_Q(x)$ and the norm is Euclidean, the iterations can be rewritten in a canonical form:

$$x_{k+1} \;=\; \mathrm{proj}_Q(x_k - \tfrac{1}{L}\nabla f(x_k)).$$

- Global convergence: $F(x_K) - F^* \leq \varepsilon$ for $K = \mathcal{O}(\frac{1}{\varepsilon})$, if the gradient is Lipschitz continuous.
- The method depends on the norm $\|\cdot\|$.
- Cheap iterations, but the rate of convergence is slow.

The Newton's Method:

$$\boxed{\begin{aligned} x_{k+1} \;\; = \;\; \operatorname*{argmin}_{y} \Big\{ & \langle \nabla f(x_k), y - x_k \rangle + \tfrac{1}{2}\langle \nabla^2 f(x_k)(y - x_k), y - x_k \rangle \\ & + \; \psi(y) \Big\}. \end{aligned}}$$

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).

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 Algorithms
4. Conclusions

$g_k(x) = f(\gamma_k x + (1-\gamma_k)x_k), \;\; \gamma_k \in (0,1]$. Then

$$
\begin{array}{rcl}
Dg_k(x) & = & \gamma_k Df(\gamma_k x + (1-\gamma_k)x_k), \\
D^2 g_k(x) & = & \gamma_k^2 D^2 f(\gamma_k x + (1-\gamma_k)x_k), \;\; \ldots
\end{array}
$$

- Smoothness properties of $g_k(\cdot)$ can be better than $f(\cdot)$.

Conceptual Contracting-Point Method:

$$
\boxed{
\begin{array}{rcl}
v_{k+1} & \in & \mathop{\mathrm{Argmin}}\limits_x \Big\{ \, f(\gamma_k x + (1-\gamma_k)x_k) + \gamma_k \psi(x) \, \Big\}, \\
x_{k+1} & = & \gamma_k v_{k+1} + (1-\gamma_k)x_k .
\end{array}
}
$$

Lemma: Inexact Contracting Step

The subproblem: $\min\limits_x \Big\{ F_k(x) \overset{\mathrm{def}}{=} f(\gamma_k x + (1-\gamma_k)x_k) + \gamma_k \psi(x) \Big\}$.

Lemma.

Let v_{k+1} be an approximate solution: $F_k(v_{k+1}) - F_k^* \leq \delta_{k+1}$. Then

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

Proof: For any $v \in \mathrm{dom}\,\psi$, it holds

$$
\begin{aligned}
& (1-\gamma_k)F(x_k) + \gamma_k F(v) \\
& \quad = \quad (1-\gamma_k)f(x_k) + \gamma_k f(v) \; + \; (1-\gamma_k)\psi(x_k) + \gamma_k \psi(v) \\
& \overset{\text{convexity}}{\geq} \quad F_k(v) + (1-\gamma_k)\psi(x_k) \\
& \overset{\text{def. of } v_{k+1}}{\geq} \quad F_k(v_{k+1}) + (1-\gamma_k)\psi(x_k) - \delta_{k+1} \\
& \overset{\text{convexity}}{\geq} \quad F(\gamma_k v_{k+1} + (1-\gamma_k)x_k) - \delta_{k+1} \quad = \quad F(x_{k+1}) - \delta_{k+1}. \qquad \square
\end{aligned}
$$

Corollary: Global Convergence

The Lemma provides us with the following guarantee of one inexact Contracting-Point step:

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

Let us take an arbitrary increasing sequence $\{A_k\}_{k \geq 1}, A_0 := 0$, and denote:

$$a_{k+1} \quad := \quad A_{k+1} - A_k \; > \; 0, \qquad \gamma_k \; := \; \tfrac{a_{k+1}}{A_{k+1}}.$$

Substituting this into $(*)$, we obtain

$$A_{k+1}F(x_{k+1}) \quad \leq \quad A_k F(x_k) + a_{k+1}F^* + A_{k+1}\delta_{k+1}.$$

Corollary: $F(x_k) - F^* \; \leq \frac{1}{A_k} \sum\limits_{i=1}^{k} A_i \delta_i.$

- If A_k grow $\nearrow$ sufficiently fast, and δ_i 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 $\forall \gamma_k \in (0,1]$:

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

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

Contracting-Point Tensor Method:

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

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

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

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

Proof: Let us consider one Contracting-Point Tensor step. We have, for any $x \in \mathrm{dom}\,\psi$

$$
\begin{aligned}
F_k(x) \;&\equiv\; f(\gamma_k x + (1-\gamma_k)x_k) + \gamma_k \psi(x) \\
&\overset{\text{Taylor's}}{\geq} f(x_k) + \sum_{i=1}^{p} \frac{\gamma_k^i}{i!} D^i f(x_k)[x - x_k]^i + \gamma_k \psi(x) - \frac{\gamma_k^{p+1}\mathcal{V}^{(p+1)}_{\mathrm{dom}\,\psi}(f)}{(p+1)!} \\
&\overset{\text{Step}}{\geq} f(x_k) + \sum_{i=1}^{p} \frac{\gamma_k^i}{i!} D^i f(x_k)[v_{k+1} - x_k]^i + \gamma_k \psi(v_{k+1}) - \frac{\gamma_k^{p+1}\mathcal{V}^{(p+1)}_{\mathrm{dom}\,\psi}(f)}{(p+1)!} \\
&\overset{\text{Taylor's}}{\geq} f(\gamma_k v_{k+1} + (1-\gamma_k)x_k) + \gamma_k \psi(v_{k+1}) - \frac{2\gamma_k^{p+1}\mathcal{V}^{(p+1)}_{\mathrm{dom}\,\psi}(f)}{(p+1)!} \\
&\equiv\; F_k(v_{k+1}) - \frac{2\gamma_k^{p+1}\mathcal{V}^{(p+1)}_{\mathrm{dom}\,\psi}(f)}{(p+1)!}.
\end{aligned}
$$

Global Convergence Rate – Continue with the Proof

Hence, one step of the Tensor method approximates a step of the conceptual Contracting-Point scheme:

$$F_k(v_{k+1}) - F_k^* \;\; \leq \;\; \delta_{k+1} \;\; := \;\; \tfrac{2\gamma_k^{p+1}\mathcal{V}^{(p+1)}_{\mathrm{dom}\,\psi}(f)}{(p+1)!}.$$

The Corollary implies the global convergence:

$$F(x_k) - F^* \;\; \leq \;\; \tfrac{1}{A_k}\sum_{i=1}^k A_i\delta_i \;\; = \;\; \tfrac{2\mathcal{V}^{(p+1)}_{\mathrm{dom}\,\psi}(f)}{(p+1)!} \cdot \tfrac{1}{A_k}\sum_{i=1}^k \tfrac{a_i^{p+1}}{A_i^p}.$$

Note: $A_k \approx k^{p+1}$, then $a_k = A_k - A_{k-1} \approx k^p$, $\frac{a_i^{p+1}}{A_i^p} \approx$ const, and

$$\tfrac{1}{A_k}\textstyle\sum_{i=1}^k \tfrac{a_i^{p+1}}{A_i^p} \;\; \approx \;\; \tfrac{1}{k^p}.$$

An appropriate choice

$$A_k \;\; := \;\; k\cdot(k+1)\cdot\ldots\cdot(k+p) \;\; \Rightarrow \;\; \boxed{\gamma_k \;=\; \tfrac{a_{k+1}}{A_{k+1}} \;=\; \tfrac{p+1}{k+p+1}}$$

$\square$

Plan of the Talk

$p = 1$: **Conditional Gradient Method:**

$$\boxed{\begin{array}{rcl} v_{k+1} & \in & \operatorname*{Argmin}_x \Big\{ \langle \nabla f(x_k), x - x_k \rangle + \psi(x) \Big\}, \\ x_{k+1} & = & \gamma_k v_{k+1} + (1 - \gamma_k) x_k. \end{array}}$$

- $F(x_k) - F^* \le \mathcal{O}(1/k)$. The same rate as of the GM.

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

$$\boxed{\begin{array}{rcl} v_{k+1} & \in & \operatorname*{Argmin}_x \Big\{ \langle \nabla f(x_k), x - x_k \rangle + \frac{\gamma_k}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle \\ & & \qquad\qquad + \psi(x) \Big\}, \\ x_{k+1} & = & \gamma_k v_{k+1} + (1 - \gamma_k) x_k. \end{array}}$$

- $F(x_k) - F^* \le \mathcal{O}(1/k^2)$.

Contracting Newton Method (reformulation):

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

Regularization of quadratic model by the assymmetric trust region.

» How to solve the subproblem?

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.

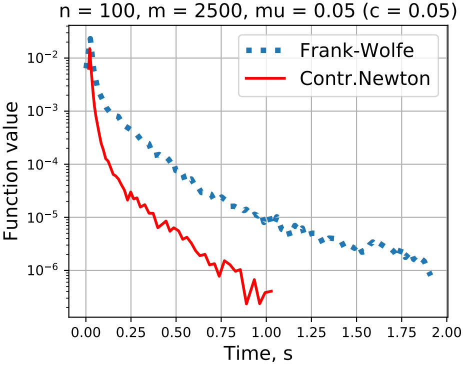
See books: [Conn-Gould-Toint, 2000], [Nocedal-Wright, 2006].

- At each step we can solve the subproblem inexactly by the first-order Conditional Gradient Method.
- We have the full control over the required accuracy.

Theorem. To reach $F(x_K) - F^* \leq \varepsilon$ we need $K = \mathcal{O}\big(\frac{1}{\sqrt{\varepsilon}}\big)$ outer iterations (oracle calls for f).
The total number of linear minimization oracle calls for ψ is $\mathcal{O}(\frac{1}{\varepsilon})$.

Experiment: Log-sum-exp over the Simplex

$$\min_{x \in \mathbb{S}_n} f(x), \qquad \mathbb{S}_n \stackrel{\mathrm{def}}{=} \left\{ x \in \mathbb{R}^n_+ \; : \; \sum_{i=1}^n x^{(i)} = 1 \right\}.$$



- 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:

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

$S_k^g, S_k^H \subseteq \{1,\ldots,M\}$ 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:

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

Theorem. At iteration k, set $m_k^g = \mathcal{O}(k^4)$, $m_k^H = \mathcal{O}(k^2)$. 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].

$$
\begin{aligned}
\hat{g}_k &:= \frac{1}{m_k^g} \sum_{i \in S_k^g} (\nabla f_i(x_k) - \nabla f_i(z_k) + \nabla f(z_k)), \\
H_k &:= \frac{1}{m_k^H} \sum_{i \in S_k^H} \nabla^2 f_i(x_k),
\end{aligned}
$$

where z_k is being updated not often.

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

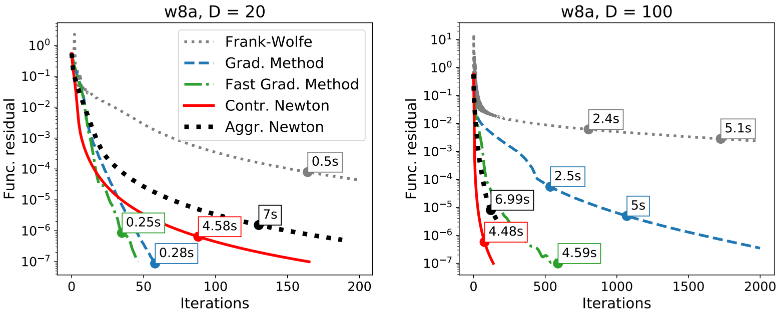
- During N iterations, we recompute the full gradient only $\log_2 N$ times.

Theorem. It is enough to set $m_k^g = m_k^H = \mathcal{O}(k^2)$. Then we have

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

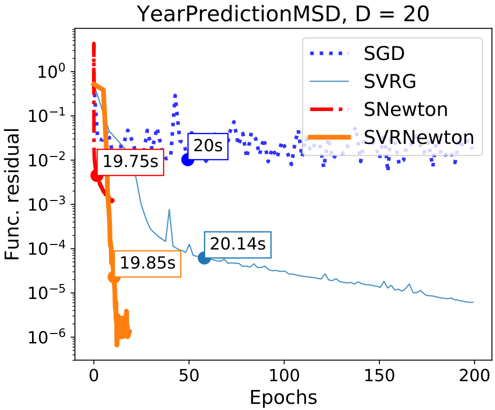
$$\min_{\|x\|_2 \leq \frac{D}{2}} \sum_{i=1}^{M} f_i(x), \qquad f_i(x) \;=\; \log(1 + \exp(\langle a_i, x \rangle)).$$

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

Takeaway Points

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 $\Rightarrow$ Frank-Wolfe algorithm.
2. Second-order approximation $\Rightarrow$ Contracting Newton Method.
3. Third-order $\Rightarrow \ldots$?

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

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

Open Problems

- Lower complexity bounds?

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

- Implementation for $p = 3$ (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. "Affine-invariant contracting-point methods for Convex Optimization". In: *arXiv:2009.08894* (2020)

Nikita Doikov and Yurii Nesterov. "Convex optimization based on global lower second-order models". In: *arXiv:2006.08518* (2020)

Thank you for your attention!