Optimization Methods for Fully Composite Problems

Nikita Doikov

Joint work with Yurii Nesterov

UCLouvain, Belgium

French German Portuguese Conference on Optimization, Porto
May 5, 2022

Optimization problems have a number of different **formulations**

- Unconstrained minimization $\min_{x \in \mathbb{R}^n} f(x)$
- Constrained minimization $\min_{x \in \mathcal{X}} f(x)$
  - A given (simple) feasible set $\mathcal{X}$ (a ball, an ellipsoid)
  - Functional constraints $\min_x \left\{ f_1(x) : f_2(x) \leq 0 \right\}$
- Additive composite minimization $\min_x f(x) + \psi(x)$
- Max-type objectives $\min_x \left\{ \max_{i=1}^m f_i(x) \right\}$
- ...

The Goal: unified *Fully Composite* formulation for convex problems

$\Rightarrow$ **new** efficient general optimization methods of different order

Plan of the Talk

I. Fully Composite Problems

- Formulation
- Examples

II. Efficient Methods

- Basic Tensor Method
- Fully Composite Methods
- Subgomogeneous Objective

III. Conclusions

Optimization Problem

Let $x \in \mathbb{E}$, where $\mathbb{E}$ is a fixed vector space ($\mathbb{R}^n$, $\mathbb{R}^{n\times n}$, S^n, etc.)

Fully Composite Problem:

$$\min_{x \in \mathrm{dom}\,\varphi} \Big\{ \varphi(x) \stackrel{\mathrm{def}}{=} F(x, f(x)) \Big\}$$

- where $f(x) = (f_1(x), \ldots, f_m(x))^T$ is a vector-function

$$f : \mathrm{dom}\, f \to \mathbb{R}^m$$

each component $f_i(x) : \mathrm{dom}\, f_i \to \mathbb{R}$ is differentiable and convex (the difficult parts of the problem)

- $F(x, u)$ is a *simple* convex component given by the problem structure

$$F : \mathbb{E} \times \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$$

Assumption: F is *monotone* in a second argument:

$$F(x, u) \;\; \leq \;\; F(x, v), \quad \forall u, v \in \mathbb{R}^m \text{ s.t. } u \leq v$$

Example: Unconstrained Minimization

Let's take

$$F(x, u) \;\equiv\; u^{(1)}$$

Then

$$\varphi(x) \;=\; F(x, f(x)) \;=\; f_1(x)$$

$\Rightarrow$ **unconstrained minimization of $f_1(\cdot)$:**

$$\boxed{\min_x f_1(x)}$$

Let's take

$$F(x, u) \;\equiv\; \psi(x) + u^{(1)},$$

where $\psi : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ is a proper closed convex function

$\Rightarrow$ **(additive) composite optimization:**

$$\boxed{\min_x f_1(x) + \psi(x)}$$

[Beck-Teboulle, 2009; Nesterov, 2013]

Let's take

$$F(x,u) \quad \equiv \quad u^{(1)} + \sum_{i=2}^{m} \mathrm{Ind}_{\leq 0}(u^{(i)}),$$

where

$$\mathrm{Ind}_{\leq 0}(t) \quad \stackrel{\mathrm{def}}{=} \quad \begin{cases} 0, & t \leq 0 \\ +\infty, & \text{otherwise} \end{cases}$$

Then $\min_x \bigl\{ \varphi(x) = F(x, f(x)) \bigr\}$ is minimization with **functional constraints**:

$$\min_{x \in \mathbb{E}} \Bigl\{ \; f_1(x) \; : \; f_2(x) \leq 0, \ldots, f_m(x) \leq 0 \Bigr\}$$

[Powell, 1978; Burke, 1985]

Let's take

$$F(x,u) \;\equiv\; \max_{i=1}^{m} u^{(i)}$$

Then, we have a problem with **Max-type objective**:

$$\min_{x \in \mathbb{E}} \Big\{ \varphi(x) \;=\; \max_{i=1}^{m} f_i(x) \Big\}$$

Note: $\varphi(x)$ is nonsmooth while all components $f_i(\cdot)$ are smooth

Plan of the Talk

I. Fully Composite Problems

- Formulation
- Examples

II. Efficient Methods

- Basic Tensor Method
- Fully Composite Methods
- Subgomogeneous Objective

III. Conclusions

Unconstrained minimization: $\boxed{\min\limits_{x \in \mathbb{E}} f(x)}$

Fix $p \geq 1$. Denote by $\Omega_p(f, x; y)$ Taylor's polynomial of f at x:

$$f(y) \;\approx\; \Omega_p(f, x; y) \;\stackrel{\mathrm{def}}{=}\; f(x) + \sum_{i=1}^{p} \tfrac{1}{i!} D^i f(x)[y - x]^i$$

Assume that p-th derivative is Lipschitz continuous:

$$\|D^p f(x) - D^p f(y)\| \;\leq\; L_p \|x - y\|, \quad \forall x, y$$

- Then we have a global bound for the approximation error:
$$|f(y) - \Omega_p(f, x; y)| \;\leq\; \tfrac{L_p}{(p+1)!} \|y - x\|^{p+1}, \quad \forall x, y$$

$$\boxed{\min_{x \in \mathbb{E}} f(x)}$$

Tensor Method of order $p \geq 1$:

$$x_{k+1} \;=\; \operatorname*{argmin}_y \Big\{ \Omega_p(f, x_k; y) + \tfrac{H}{(p+1)!} \|y - x_k\|^{p+1} \Big\}$$

$H > 0$ is a regularization parameter

$p = 1$: **The Gradient Method** [Cauchy, 1847]

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

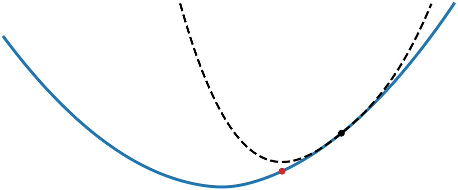
$p = 2$: **Newton Method with Cubic Regularization**
[Nesterov-Polyak, 2006]

$$x_{k+1} \;=\; x_k - \Big(\nabla^2 f(x_k) + \tfrac{H r_k}{2} I \Big)^{-1} \nabla f(x_k)$$

Convergence

$\boxed{\min_{x \in \mathbb{E}} f(x)}$ **Tensor Method of order** $p \geq 1$:

$$x_{k+1} \;=\; \operatorname*{argmin}_{y} \Big\{ \Omega_p(f, x_k; y) + \tfrac{H}{(p+1)!} \|y - x_k\|^{p+1} \Big\}$$

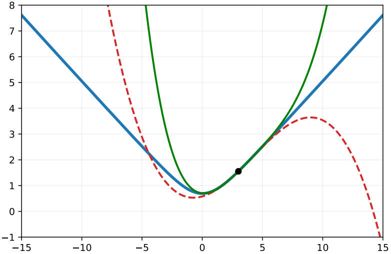


Theorem [Baes, 2009]:

Let $H := L_p \quad \Rightarrow \quad$ global rate $\quad f(x_k) - f^* \leq O\big(1/k^p\big)$.

- How to solve subproblem?

Note: $\Omega_p(f, x; y)$ is **nonconvex** for $p \geq 3$.



- **Theorem** [Nesterov, 2018]:

Let $f(\cdot)$ be a convex function and $H \geq pL_p$. Then $\forall x$ the model

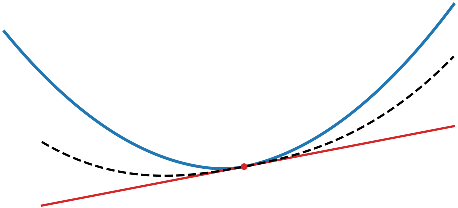
$$M(y) \;\; := \;\; \Omega_p(f, x; y) + \frac{H}{(p+1)!}\|y - x\|^{p+1}$$

is convex in y

- **For $p = 3$:** efficient implementation using only second-order oracle is available [Nesterov, 2019]. The cost is $\mathcal{O}(n^3) + \tilde{O}(n)$.

f is called uniformly convex of degree $q \geq 2$ with $\sigma_q > 0$, iff

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\sigma_q}{q} \|y - x\|^q, \quad \forall x, y$$



- Strongly convex functions: $q = 2$
- Example: $f(x) = \frac{1}{q}\|x - x_0\|^q$ is uniformly convex of degree q with constant $\sigma_q = 2^{2-q}$
- Sum of convex and uniformly convex functions gives uniformly convex

Global Lower Model

Fix $p \geq 1$.

- Let $D^p f$ be Lipschitz continuous
- Let f be uniformly convex of degree $q = p + 1$

Theorem. We have a *global lower bound*, $\forall x, y$:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \sum_{i=2}^{p} \frac{\gamma^{i-1}}{i!} D^i f(x)[y-x]^i + \frac{\gamma^p p L_p}{(p+1)!} \|y - x\|^{p+1}$$

where $\gamma \geq \frac{1}{2} \min\{1, \frac{p!}{p+1} \frac{\sigma_{p+1}}{L_p}\}$ is a **condition number**.

Recall the global *upper model*:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \sum_{i=2}^{p} \frac{1}{i!} D^i f(x)[y-x]^i + \frac{L_p \|y-x\|^{p+1}}{(p+1)!}$$

Plan of the Talk

I. Fully Composite Problems

- Formulation
- Examples

II. Efficient Methods

- Basic Tensor Method
- Fully Composite Methods
- Subgomogeneous Objective

III. Conclusions

FC Problem: $\min_x \Big\{ \varphi(x) \;=\; F(x, f(x)) \Big\}$

where

$$f(x) \;=\; [\, f_1(x), \ldots, f_m(x)\,]^T$$

denote the directional derivatives of f by

$$D^i f(x)[h]^i \;\stackrel{\text{def}}{=}\; \big[\, D^i f_1(x)[h]^i, \ldots, D^i f_m(x)[h]^i \,\big]^T, \quad \forall h \in \mathbb{E}$$

Assume:

- $D^p f_i$ are Lipschitz continuous
- f_i are uniformly convex of degree $p+1$

Basic Method of order $p \geq 1$:

$$x_{k+1} \;=\; \operatorname*{argmin}_y F\big(y, \Omega_p(f, x_k; y) + \tfrac{H}{(p+1)!}\|y - x_k\|^{p+1}\big), \quad k \geq 0$$

Set $H := p[L_p(f_1), \ldots, L_p(f_m)]^T$

Theorem. $\varphi(x_{k+1}) - \varphi^* \;\leq\; (1-\gamma)\big(\varphi(x_k) - \varphi^*\big)$
Proof:

$$\gamma\varphi^* + (1-\gamma)\varphi(x_k)$$

$$\overset{\text{convex}}{\geq} \; F\big(\gamma x^* + (1-\gamma)x_k,\; \gamma f(x^*) + (1-\gamma) f(x_k)\big)$$

$$\overset{\text{lower}}{\geq} \; F\Big(\gamma x^* + (1-\gamma)x_k,\; f(x_k) + \sum_{i=1}^{k} \tfrac{\gamma^i}{i!} D^i f(x_k)[x^* - x_k]^i + \tfrac{\gamma^p p L_p \|x^* - x_k\|^{p+1}}{(p+1)!}\Big)$$

Denote $y := x_k + \gamma(x^* - x_k)$. Then $y - x_k = \gamma(x^* - x_k)$

Proof: continue

We have

$$
\begin{aligned}
&\gamma\varphi^* + (1-\gamma)\varphi(x_k) \\
&\geq \; F\Big(y, f(x_k) + \sum_{i=1}^{k} \tfrac{1}{i!} D^i f(x_k)[y-x_k]^i + \tfrac{pL_p\|y-x_k\|^{p+1}}{(p+1)!}\Big) \\
&\overset{\text{method}}{\geq} F\Big(x_{k+1}, f(x_k) + \sum_{i=1}^{k} \tfrac{1}{i!} D^i f(x_k)[x_{k+1}-x_k]^i + \tfrac{pL_p\|x_{k+1}-x_k\|^{p+1}}{(p+1)!}\Big) \\
&\overset{\text{upper}}{\geq} F\big(x_{k+1}, f(x_{k+1})\big) \;=\; \varphi(x_{k+1})
\end{aligned}
$$

Hence,

$$
\varphi(x_{k+1}) - \varphi^* \;\leq\; (1-\gamma)\big(\varphi(x_k) - \varphi^*\big).
$$

□

Example: Functional Constraints

Problem: $\boxed{\min_x \left\{ f_1(x) \;:\; f_2(x) \le 0 \right\}}$

- $p = 1$

Method step:

$$\begin{aligned} x^+ \;&=\; \operatorname*{argmin}_y \Big\{ f_1(\bar{x}) + \langle \nabla f_1(\bar{x}), y - \bar{x} \rangle + \tfrac{H_1}{2} \|y - \bar{x}\|^2 \\ &\qquad : \quad f_2(\bar{x}) + \langle \nabla f_2(\bar{x}), y - \bar{x} \rangle + \tfrac{H_2}{2} \|y - \bar{x}\|^2 \;\le\; 0 \Big\} \end{aligned}$$

[Auslender-Shefi-Teboulle, 2010]

- f_1 and f_2 are strongly convex and have Lipschitz gradient

$\Rightarrow$ the method has global linear convergence

The step is $x^+ \;=\; x - \frac{1}{\lambda_1 H_1 + \lambda_2 H_2} \big(\lambda_1 \nabla f_1(x) + \lambda_2 \nabla f_2(x) \big)$

Problem: $$\boxed{\min_x \left\{ f_1(x) \ : \ f_2(x) \le 0 \right\}}$$

- $p = 2$

Method step:

$$x^+ = \operatorname*{argmin}_y \Big\{ f_1(\bar{x}) + \langle \nabla f_1(\bar{x}), y - \bar{x} \rangle + \tfrac{1}{2} \nabla^2 f_1(\bar{x})[y - \bar{x}]^2 + \tfrac{H_1}{6} \| y - \bar{x} \|^3$$
$$: \ \ f_2(\bar{x}) + \langle \nabla f_2(\bar{x}), y - \bar{x} \rangle + \tfrac{1}{2} \nabla^2 f(\bar{x})[y - \bar{x}]^2 + \tfrac{H_2}{6} \| y - \bar{x} \|^3 \ \le \ 0 \Big\}$$

- Global linear rate for uniformly convex functions with Lipschitz Hessian

The step is $x^+ \ = \ x - M^{-1}\big(\lambda_1 \nabla f_1(\bar{x}) + \lambda_2 \nabla f_2(\bar{x})\big),$

with $M = \lambda_1 \nabla^2 f_1(\bar{x}) + \lambda_2 f_2(\bar{x}) + \tau I$

What if objective is not uniformly convex?

- We can do regularization

Initial problem: $\min_x \varphi(x) = F(x, f(x))$

Regularized problem: $\min_x \varphi_\mu(x) = F(x, f(x) + \frac{\mu}{p+1}\|x\|^{p+1})$

Set

$$\boxed{\mu \quad \approx \quad \delta^2}$$

where δ is the target relative accuracy:

$$\varphi(\bar{x}) - \varphi^* \quad \leq \quad \delta(\varphi(x_0) - \varphi^*)$$

- The method of order $p \geq 1$ needs $\tilde{O}\big((\frac{1}{\delta})^{\frac{2}{p}}\big)$ iterations

Plan of the Talk

I. Fully Composite Problems

- Formulation
- Examples

II. Efficient Methods

- Basic Tensor Method
- Fully Composite Methods
- Subgomogeneous Objective

III. Conclusions

$g : \mathrm{dom}\, g \to \mathbb{R}$ is called subhomogeneous iff

$$g(tx) \quad \leq \quad tg(x), \qquad \forall t \geq 1$$

- $g(x) = \max\limits_{i=1}^{m} x^{(i)}$ is subhomogeneous
- $g(x) = \ln\left(\sum_{i=1}^{m} \exp x^{(i)}\right)$ is subhomogeneous

Assume that $F(x, u)$ is subhomogeneous in u

Note: this with monotonicity and $0 \in \mathrm{int}\,\mathrm{dom}\, F(x, \cdot)$ implies that $\mathrm{dom}\, F(x, \cdot) = \mathbb{R}^m$

Subgomogeneous Problem

Problem: $\boxed{\min_x \varphi(x) \;=\; F(x, f(x))}$

- $F(x,\cdot)$ is subhomogeneous and monotone

(e.g. $\varphi(x) = \max_{i=1}^{m} f_i(x)$)

Basic Gradient Method:

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

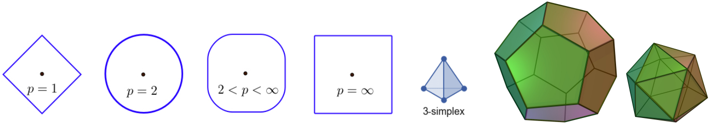
[Burke, 1985; Cartis-Gould-Toint, 2011; Lewis-Wright, 2016]

Theorem. Global convergence: $\varphi(x_k) - \varphi^* \leq O(1/k)$

Contracting Conditional Gradient Method

Problem: $$\min_{x \in \mathrm{dom}\,\varphi} \varphi(x) \;=\; F(x, f(x))$$

- $F(x, \cdot)$ is subhomogeneous and monotone
- $Q := \mathrm{dom}\,\varphi$ is bounded



Contracting Conditional Gradient Method:

$$x_{k+1} \;=\; \operatorname*{argmin}_{y} \Big\{ F\big(y, f(x_k) + \langle \nabla f(x_k), y - x_k \rangle\big) \;:\; x_k + \tfrac{1}{\gamma_k}(y - x_k) \in Q \Big\}$$

[Frank-Wolfe, 1956]

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

Problem: $\boxed{\min_x \varphi(x) \; = \; F(x, f(x))}$

- $F(x, \cdot)$ is subhomogeneous and monotone

Fast Gradient Method:

$$\begin{aligned} y_k &= \tfrac{a_{k+1} v_k + A_k x_k}{A_{k+1}} \\ x_{k+1} &= \operatorname*{argmin}_y \Big\{ F\big(y, f(y_k) + \langle \nabla f(y_k), y - y_k \rangle\big) + \tfrac{H}{2} \|y - y_k\|^2 \Big\} \\ v_{k+1} &= x_{k+1} + \tfrac{A_k}{a_{k+1}} (x_{k+1} - x_k) \end{aligned}$$

[Nesterov, 1983]

Theorem. Set $a_k = k^2, A_k = \sum_{i=1}^{k} a_i$. Then $\varphi(x_k) - \varphi^* \le O(1/k^2)$

Plan of the Talk

I. Fully Composite Problems

- Formulation
- Examples

II. Efficient Methods

- Basic Tensor Method
- Fully Composite Methods
- Subgomogeneous Objective

III. Conclusions

Conclusions

- Fully Composite formulation $\varphi(x) = F(x, f(x)) \to \min\limits_{x}$
  - includes functional constraints
  - max-type minimization
- High-order $(p \geq 1)$ methods
- Global linear rate when the components are uniformly convex
- Subgomogeneous assumption $\Rightarrow$ more efficient methods

Thank you for your attention!