Modern Theory of 2nd-Order Methods (Dec 2019)

Lecture 4: Implementable Tensor Methods

Yurii Nesterov (CORE/INMA, UCLouvain)

Minicourse: January 20-23, 2020 (Munich)

Contents

Taylor Approximation

Let $x \in \mathrm{int}\,(\mathrm{dom}\, f)$. Then

$$f(x+h) \;=\; \Phi_{x,p}(h) + o(\|h\|^p), \quad x+h \in \mathrm{dom}\, f,$$

where $\boxed{\Phi_{x,p}(y) \stackrel{\mathrm{def}}{=} f(x) + \sum\limits_{i=1}^{p} \frac{1}{i!} D^i f(x)[y-x]^i,\ y \in \mathbb{E}}$ and

$$D^p f(x)[h_1, \dots, h_p]$$

is the directional derivative of f at x along directions $h_i \in \mathbb{E}$, $i = 1, \dots, p$.

Note:

1. $D^p f(x)[\cdot]$ is a *symmetric p-linear form*.
2. If $h_1 = \cdots = h_p$, we use notation $\boxed{D^p f(x)[h]^p}$

Measuring the quality of approximations

Let us fix a norm $\| \cdot \|$ in $\mathbb{E}$ and define the norm

$$\| D^p f(x) \| = \max_h \left\{ \left| D^p f(x)[h]^p \right| : \; \|h\| \le 1 \right\}.$$

Then we can introduce Lipschitz constants for derivatives:

$$\boxed{\| D^p f(x) - D^p f(y) \| \le L_p \| x - y \|, \quad x, y \in \mathrm{dom}\, f}$$

These constants ensure the high-quality of local approximations:

A. *Function*: $\boxed{| f(y) - \Phi_{x,p}(y) | \le \frac{L_p}{(p+1)!} \| y - x \|^{p+1}}$

B. *Gradient:* $\boxed{\| f'(y) - \Phi'_{x,p}(y) \|_* \le \frac{L_p}{p!} \| y - x \|^{p}}$

C. *Hessian:* $\boxed{\| f''(y) - \Phi''_{x,p}(y) \| \le \frac{L_p}{(p-1)!} \| y - x \|^{p-1}}$

and so on ...

And what?

Note that:

1. For $p \geq 3$, $\Phi_{x,p}(y)$ is a *non-convex multivariate polynomial*.

2. Up to now, Algebraic Geometry cannot provide us with efficient tools for computing even its stationary points

(not speaking about the global minimum)

Consequence

Practical Optimization goes up to the 2nd-order methods.

Let us look ...

Let us fix $B = B^* \succ 0 : \mathbb{E} \to \mathbb{E}^*$ and define the norms

$$\|x\| = \langle Bx, x\rangle^{1/2}, \quad x \in \mathbb{E}, \quad \|g\|_* = \langle g, B^{-1}g\rangle^{1/2}, \quad g \in E^*.$$

Let us introduce the *power function* $\boxed{d_p(x) = \frac{1}{p}\|x\|^p, \; p \geq 2}$ with

$$\begin{aligned} d_p'(x) &= \|x\|^{p-2}Bx, \\ d_p''(x) &= \|x\|^{p-2}B + (p-2)\|x\|^{p-4}Bxx^*B \\ &\succeq \|x\|^{p-2}B. \end{aligned}$$

Define $\boxed{\Omega_{x,p,M}(y) = \Phi_{x,p}(y) + \frac{M}{p!}d_{p+1}(y-x)}$

NB: 1. If $M \geq L_p$, then $f(y) \overset{(A)}{\leq} \Omega_{x,p,M}(y)$ for all $y \in \mathbb{E}$.

2. The epigraph $\{(x,t) : \; t \geq f(x)\}$ is a <u>*convex set*</u>.

Question: Is it easy to put a <u>*nonconvex object*</u> into the convex one?

The answer is: NO!

Main Theorem

Let $M \geq pL_p$. Then function $\Omega_{x,p,M}(\cdot)$ is convex.

Proof. $\Phi''_{x,p}(\cdot)$ is a Taylor approximation of $f''(\cdot)$.

Therefore, for any $y \in \mathbb{E}$ we have

$$\begin{aligned} 0 \ \preceq \ f''(y) \ &\preceq \ \Phi''_{x,p}(y) + \tfrac{L_p}{(p-1)!}\|y-x\|^{p-1}B \\ &\preceq \ \Phi''_{x,p}(y) + \tfrac{M}{p!}\|y-x\|^{p-1}B \\ &\preceq \ \Omega''_{x,p,M}(y). \qquad \square \end{aligned}$$

Consequences

1. For $M > pL_p$ the point $T_{p,M}(x) = \arg\min\limits_{y\in\mathbb{E}} \Omega_{x,p,M}(y)$ is well defined.
2. It can be computed by the techniques of *Convex Optimization*.
3. It can be used for solving the problem $\quad f_* = \min\limits_{x\in\mathbb{E}} f(x)$

in the case $L_p(f) < +\infty$.

Properties of the Tensor Step

Let $T = T_{p,M}(x)$ be the solution of the equation

$$\boxed{\Phi'_{x,p}(T) + \frac{M}{p!} r^{p-1} B(T-x) = 0}$$

where $r = \| T - x \|$.

$$\boxed{\| f'(T) \| \leq \frac{L_p + M}{p!} r^p}$$

Proof.

$$\begin{aligned} \| f'(T) \| &= \| f'(T) - \Phi'_{x,p}(T) - \frac{M}{p!} r^{p-1} B(T-x) \| \\ &\leq \| f'(T) - \Phi'_{x,p}(T) \| + \frac{M}{p!} r^p \leq \frac{M+L_p}{p!} r^p. \end{aligned}$$

□

$$\boxed{\langle f'(T), x - T \rangle \geq \frac{M - L_p}{p!} r^{p+1}}$$

Proof.

$$\begin{aligned} \langle f'(T), x - T \rangle &= \langle f'(T) - \Phi'_{x,p}(T) - \frac{M}{p!} r^{p-1} B(T-x), x - T \rangle \\ &\geq \frac{M - L_p}{p!} r^{p+1}. \end{aligned}$$

□

Local Method

For $M \geq pL_p$, consider the process

$$\boxed{x_{t+1} \;=\; T_{p,M}(x_t), \;\; t \geq 0.}$$

Theorem 2. For all $t \geq 0$ we have $f(x_{t+1}) \leq f(x_t)$.

At the same time, $\boxed{f(x_t) - f_* \leq \frac{(M+L_p)D^{p+1}}{p!} \left(\frac{p+1}{t}\right)^p, \quad t \geq 1}$

where $D = \max\limits_{x \in E}\{\|x - x^*\| : \; f(x) \leq f(x_0)\}$.

Proof. We have

$$\begin{aligned} f(x_k) - f(x_{k+1}) \;&\geq\; O(r_k^{p+1}) \;\geq O(\|f'(x_{k+1})\|^{\frac{p+1}{p}}) \\ &\geq\; O((f(x_{k+1}) - f^*)^{\frac{p+1}{p}}). \end{aligned}$$

□

Accelerated Tensor Method

NB: We apply the standard technique of *estimating sequences*

We choose $M \geq pL_p$ and recursively update the following sequences.

1. Sequence of estimating functions

$$\psi_k(x) = \ell_k(x) + \tfrac{C}{p!} d_{p+1}(x - x_0), \quad k \geq 1,$$

where $\ell_k(\cdot)$ are linear functions in $x \in \mathbb{E}$, and $C > 0$.

2. Minimizing sequence $\{x_k\}_{k=1}^{\infty}$.

3. Sequence of scaling parameters $\{A_k\}_{k=1}^{\infty}$: $A_{k+1} \stackrel{\mathrm{def}}{=} A_k + a_k$, $k \geq 1$.

For these objects, we are going to maintain the following relations:

$$\mathcal{R}_k^1 : \; A_k f(x_k) \leq \psi_k^* \; \equiv \; \min_{x \in \mathbb{E}} \psi_k(x),$$

$$\mathcal{R}_k^2 : \; \psi_k(x) \leq A_k f(x) + \tfrac{M + L_p + C}{p!} d_{p+1}(x - x_0), \; \forall x \in \mathbb{E}, \; k \geq 1.$$

Define $A_k = \left[\frac{(p-1)(M^2-p^2L_p^2)}{4(p+1)M^2}\right]^{\frac{p}{2}} \left(\frac{k}{p+1}\right)^{p+1}$, $a_{k+1} = A_{k+1} - A_k$, $k \geq 0$.

Initialization: Choose $x_0 \in \mathbb{E}$ and $M > pL_p$.

Define $C = \frac{p}{2}\sqrt{\frac{(p+1)}{(p-1)}(M^2 - p^2 L_p^2)}$ and $\psi_0(x) = \frac{C}{p!} d_{p+1}(x - x_0)$.

Iteration k, ($k \geq 1$):

1. Compute $v_k = \arg\min\limits_{x\in\mathbb{E}} \psi_k(x)$ and choose $y_k = \frac{A_k}{A_{k+1}} x_k + \frac{a_{k+1}}{A_{k+1}} v_k$.

2. Compute $x_{k+1} = T_{p,M}(y_k)$ and update

$$\psi_{k+1}(x) = \psi_k(x) + a_{k+1}[f(x_{k+1}) + \langle f'(x_{k+1}), x - x_{k+1}\rangle].$$

Convergence:

$$f(x_k) - f(x^*) \leq \frac{M+L_p+C}{(p+1)!}\left[\frac{4(p+1)M^2}{(p-1)(M^2-p^2L_p^2)}\right]^{\frac{p}{2}} \left(\frac{p+1}{k}\right)^{p+1} \|x_0 - x^*\|^{p+1}.$$

Lower Complexity Bounds

Assumption: Method can move only to the point generated by pth-order information.

Difficult function. Define $\eta_{p+1}(x) = \frac{1}{p+1} \sum\limits_{i=1}^{n} |x^{(i)}|^{p+1}, \quad x \in \mathbb{R}^n$.

Let $U_k = \begin{pmatrix} 1 & -1 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ & & \dots & \\ 0 & 0 & \dots & -1 \\ 0 & 0 & \dots & 1 \end{pmatrix} \in \mathbb{R}^{k \times k}$, and $A_k = \begin{pmatrix} U_k & 0 \\ 0 & I_{n-k} \end{pmatrix}$.

Consider the function $\boxed{f_k(x) = \eta_{p+1}(A_k x) - x^{(1)}, \quad 2 \le k \le p}$

Theorem 3. Let for any function f with $L_p(f) < +\infty$ method $\mathcal{M}$ ensures the rate of convergence

$$\min_{0 \le k \le t} f(x_k) - f_* \le \frac{L_p \|x_0 - x^*\|^{p+1}}{(p+1)!\ \kappa(t)}, \ t \ge 1.$$

Then for all $t:\ 2t+1 \le n$ we have $\kappa(t) \le \frac{1}{3p} 2^{p+1} (2t+2)^{\frac{3p+1}{2}}$.

NB: for $p = 2$ the lower bound is $\boxed{O\left(\frac{1}{k^{3.5}}\right)}$

Degree of Non-Optimality

Accelerated method:

- Rate of convergence: $O\left(\left(\frac{1}{t}\right)^{p+1}\right)$.
- Complexity bound: $O\left(\left(\frac{1}{\epsilon}\right)^{\frac{1}{p+1}}\right)$.

Lower bound:

- Rate of convergence: $O\left(\left(\frac{1}{t}\right)^{\frac{3p+1}{2}}\right)$.
- Complexity bound: $O\left(\left(\frac{1}{\epsilon}\right)^{\frac{2}{3p+1}}\right)$.

Extra factor: $O\left(\left(\frac{1}{\epsilon}\right)^{\frac{p-1}{(p+1)(3p+1)}}\right)$.

NB: $p = 1 \Rightarrow \left(\frac{1}{\epsilon}\right)^0$, $p = 2 \Rightarrow \left(\frac{1}{\epsilon}\right)^{\frac{1}{21}}$, $p = 3 \Rightarrow \left(\frac{1}{\epsilon}\right)^{\frac{1}{20}}$.

At the same time, $2^{20} \approx 10^6$.

Conclusion: "Optimal methods" with expensive line search should not work in practice.

Third-order methods: implementation details

Taylor polynomial:

$$\Phi_x(h) = \langle f'(x), h\rangle + \tfrac{1}{2}\langle f''(x)h, h\rangle + \tfrac{1}{6} D^3 f(x)[h]^3.$$

Auxiliary Problem: $\Omega_{x,M}(h) \stackrel{\mathrm{def}}{=} \Phi_x(h) + \frac{M}{24}\|h\|^4 \;\to\; \min\limits_{h\in\mathbb{E}}.$

Main Theorem: for all $h \in \mathbb{E}$ we have

$$0 \preceq f''(x) + D^3 f(x)[\pm h] + \tfrac{1}{2} L_3 \|h\|^2 B.$$

Conclusion: For any $h \in \mathbb{E}$, the Hessian $\Phi_x''(h)$ is *similar* to the Hessian of the function

$$\rho_x(h) = \tfrac{1}{2}(1 - \tfrac{1}{\tau})\langle f''(x)h, h\rangle + \tfrac{M-\tau L_3}{10}\|h\|^4$$

with some $\tau > 1$.

Relative Smoothness Condition

Definition: Function $f(\cdot)$ satisfies the strong relative smoothness condition with respect to $\rho(\cdot)$ if

$$\mu\rho''(x) \preceq f''(x) \preceq L\rho''(x).$$

Define the Bregman distance $\beta_\rho(x, y) = \rho(y) - \rho(x) - \langle \rho'(x), y - x \rangle$.

Consider the method:

$$x_{k+1} = \arg\min_{y \in \mathbb{E}} \{ \langle f'(x_k), y - x_k \rangle + L\beta_\rho(x_k, y) \}. \quad (*).$$

Theorem 4. $\quad f(x_k) - f^* \leq \frac{\mu\beta_\rho(x_0, x^*)}{\left(\frac{L}{L-\mu}\right)^k - 1}$.

NB: 1. For 3rd-order method with $\rho = \rho_x$, we have $\mu = 1$, $L = \frac{\tau+1}{\tau-1}$.

2. Solution of problem (*) is simple:

$$\min_{h \in \mathbb{E}} \{ \langle g, h \rangle + \tfrac{1}{2}\langle Gh, h \rangle + \gamma \|h\|^4 \},$$

especially after an appropriate factorization of matrix $G \succeq 0$.

Remarks

1. There exists an accelerated 3rd order schemes for minimizing smooth convex functions with the global rate of convergence $O(\frac{1}{k^4})$.

This is the fastest sublinear rate known so far.

2. These schemes are *implementable*. Complexity of each iteration is comparable with that of the 2nd-order methods:

- Linear convergence rate of auxiliary process depends only on absolute constant.
- Algorithmic complexity of one iteration is $O(n^2)$.
- The oracle is simple: we need to compute the vector $D^3 f(x)[h]^2$.

  (e.g. Separable Optimization: $\sum\limits_{i=1}^{N} f_i(\langle a_i, x\rangle)$, functions with explicit structure (by fast backward differentiation), etc.)
- The vector $D^3 f(x)[h]^2$ can be approximate by the <u>*2nd-order oracle*</u>.

  Then we get 2nd-order method with the rate of convergence $O(\frac{1}{k^4})$.

  No contradiction with the lower bounds since this is for another problem class.