Efficiency of Quasi-Newton methods on strictly positive functions

Yurii Nesterov, CORE/INMA (UCL)

October 12, 2010, UCLA

IPAM Workshop II: Numerical methods for Continuous Optimization

Outline

Standard approach: Absolute accuracy

Problem: $f^* \stackrel{\mathrm{def}}{=} \min\limits_{x \in Q} f(x)$, where $Q \subseteq R^n$ is a closed convex set.

Definition:

For $\epsilon > 0$, find $\quad \bar{x} \in Q \quad$ satisfying $\quad f(\bar{x}) \leq f^* + \epsilon$.

Problem classes

1. *Bounds on the growth.* (Strong) convexity with $\mu \geq 0$:
$$f(y) \geq f(x) + \langle f'(x), y - x \rangle + \tfrac{1}{2}\mu \|y - x\|^2, \quad x, y \in Q.$$

2. *Bounds on derivatives.* For example,
$$\|f'(x)\|_* \leq M, \quad \text{or}, \quad \|f''(x)\| \leq L, \quad \text{etc.}$$

Important: operation $\quad f \;\Rightarrow\; f + \text{const} \quad$ does not change complexity.

Relative accuracy (RA)

Problem: $f^* \stackrel{\mathrm{def}}{=} \min\limits_{x \in Q} f(x) \; > 0$, where Q is a closed convex set.

Definition:

For $\delta \in (0,1)$, find $\bar{x} \in Q$ satisfying $(1-\delta) f(\bar{x}) \leq f^* \leq f(\bar{x})$.

Condition $f^* > 0$ must be guaranteed. *How?*

1. Homogeneous model [N.08]: Let $0 \notin Q$. For $f(x) = \max\limits_{s \in B} \langle s, x \rangle$ with $0 \in \mathrm{int}\, B$ define γ_0, γ_1: $\gamma_0 \|x\| \leq f(x) \leq \gamma_1 \|x\|$, $\alpha = \frac{\gamma_0}{\gamma_1}$.
Then, by smoothing technique, we get complexity $O^*(\frac{1}{\alpha\delta})$.

2. Polyhedral model [N.09]: if $B = \mathrm{Conv}\,(\pm a_i, i = 1 \dots m) \subset R^n$, then we need $O^*(\frac{n^{1/2}}{\delta})$ iterations.
(An appropriate norm is constructed by preprocessing.)

Question: Can we address RA in Black-Box Framework? $\Rightarrow$ Need new problem classes. (Invariant with respect to multiplication.)

Barrier subgradient method [N.10]

Problem: $\max_{x \in Q} \phi(x)$,

- Q is a closed convex set endowed with a ν-self-concordant barrier $F(x)$.
- ϕ is a concave function, which is *non-negative* on Q.

Method: (potential $f = \ln \phi$)

$$x_{k+1} = \arg\max_{x \in Q} \left[\sum_{i=0}^{k} \langle \frac{\phi'(x_i)}{\phi(x_i)}, x - x_i \rangle - \left(1 + \sqrt{\frac{k+1}{\nu}}\right) F(x) \right].$$

Convergence: $\max_{0 \leq i \leq k} \phi(x_i) \geq \phi^* \left(1 - O^*\left(\sqrt{\frac{\nu}{k+1}} + \frac{\nu}{k+1}\right)\right)$.

Complexity: $O^*\left(\frac{\nu}{\delta^2}\right)$ iterations.

Strictly positive functions

Definition

Convex function f is called strictly positive on Q if

$$f(y) + f(x) + \langle f'(x), y - x \rangle \geq 0, \quad x, y \in Q.$$

Corollary: $f(y) \geq |f(x) + \langle f'(x), y - x \rangle|, \quad x, y \in Q.$

Simple properties

- $f(x) \equiv \text{const} > 0$ is strictly positive.
- Strict positivity is an *affine-invariant* property.
- Class of strictly positive functions is a convex cone.

Simple examples

Lemma 1. Let B be bounded, closed, and centrally symmetric.

Then $f(x) = \max\limits_{x \in B} \langle s, x \rangle$ is strictly positive on R^n.

Proof: Since $f(x) = \langle f'(x), x \rangle$ and $-f'(x) \in B$, we have

$$f(y) \;\geq\; \langle -f'(x), y \rangle \;=\; -f(x) - \langle f'(x), y - x \rangle. \qquad \square$$

The simplest examples of strictly positive functions are *norms*.

Lemma 2. Let $f_1(x)$ and $f_2(x)$ be strictly positive on Q.

Then $f(x) = \max\{f_1(x), f_2(x)\}$ is also strictly positive.

Proof: For arbitrary $x \in Q$, assume $f_1(x) \geq f_2(x)$. Then,

$$\begin{aligned} f(y) \;&\geq\; f_1(y) \;\geq\; -f_1(x) - \langle f_1'(x), y - x \rangle \\ &=\; -f(x) - \langle f'(x), y - x \rangle. \qquad \square \end{aligned}$$

Particular examples

All functions below are strictly positive:

$$
\begin{aligned}
f(x) &= \max_{1 \le i \le m} \|A_i x - b_i\|, \\
f(x) &= \sum_{i=1}^{m} \|A_i x - b_i\|, \\
f(x) &= \sigma_{\max}\left(\sum_{i=1}^{n} A_i x^{(i)} \right), \\
f(x) &= \sum_{j=1}^{m} \sigma_j \left(\sum_{i=1}^{n} A_i x^{(i)} \right),
\end{aligned}
$$

where $A_i \in R^{m \times n}$, and $b_i \in R^m$, $i = 1 \dots n$.

General convex functions

Theorem 1. Let ϕ be convex function on Q with uniformly bounded subgradients: $\|\phi'(x)\|^* \leq L, \quad x \in Q$.

Then $f(x) = \max\{\phi(x), L\|x\|\}$ is strictly positive on Q.

Proof: Clearly, $\|f'(x)\|^* \leq L$. Therefore,

$$\begin{aligned} f(y) + f(x) + \langle f'(x), y - x\rangle &\geq L\|y\| + L\|x\| + \langle f'(x), y - x\rangle \\ &\geq L\|y\| + L\|x\| - L\|y - x\| \geq 0. \end{aligned}$$

□

Shifted general optimization problem

Consider the problem: $\min\limits_{x \in Q} \phi(x)$, where ϕ has bounded subgradients. Let $x^* \in Q$ be its optimal solution.

Lemma 3. For $x_0 \in Q$ define

$$f(x) = \max\{\phi(x) - \phi(x_0) + 2LR, L\|x - x_0\|\}.$$

It is strictly positive. If $\|x - x_0\| \leq R$ then $f(x) \equiv \phi(x) + \text{const}$.

If $\|x_0 - x^*\| \leq R$, then the optimal value f^* of the equivalent problem $\min\limits_{x \in Q} f(x)$ satisfies $LR \leq f^* \leq 2LR$.

Proof: If $\|x - x_0\| \leq R$, then

$$\phi(x) - \phi(x_0) + 2LR \geq 2LR - L\|x - x_0\| \geq L\|x - x_0\|.$$

Further, $f^* \leq f(x_0) = 2LR$, and

$$f(x) \geq \max\{2LR - L\|x - x_0\|, L\|x - x_0\|\} \geq LR. \qquad \square$$

Optimization problem with squared objective

Problem: $\min_{x \in Q} f(x)$, where f is strictly positive on Q.

New objective: $\hat{f}(x) = \frac{1}{2} f^2(x), \quad \hat{f}'(x) = f(x) \cdot f'(x).$

Equivalent problem: $\min_{x \in Q} \hat{f}(x).$

Lemma 4. Let f be strictly positive on Q. Then for $x, y \in Q$

$$\hat{f}(y) \geq \hat{f}(x) + \langle \hat{f}'(x), y - x \rangle + \tfrac{1}{2} \langle f'(x), y - x \rangle^2.$$

Proof: Indeed,

$$\begin{aligned} \hat{f}(y) &= \tfrac{1}{2} f^2(y) \geq \tfrac{1}{2} [f(x) + \langle f'(x), y - x \rangle]^2 \\ &= \hat{f}(x) + \langle \hat{f}'(x), y - x \rangle + \tfrac{1}{2} \langle f'(x), y - x \rangle^2. \end{aligned}$$

□

Important: We have *nonlinear support function*!

Quasi-Newton Method

Let us fix $G_0 \succ 0$, starting point $x_0 \in Q$, and accuracy $\delta \in (0,1)$. Define $\psi_0(x) = \frac{1}{2}\|x - x_0\|^2_{G_0}$. For $k \geq 0$, consider the process:

$$\begin{aligned} x_k &= \arg\min_{x \in Q} \psi_k(x), \\ \psi_{k+1}(x) &= \psi_k(x) + a_k \left[\hat{f}(x_k) + \langle \hat{f}'(x_k), x - x_k \rangle + \tfrac{1}{2}\langle f'(x_k), x - x_k \rangle^2 \right], \end{aligned}$$

where

$$a_k = \frac{\delta}{1-\delta} \cdot \frac{1}{(\|f'(x_k)\|^*_{G_k})^2}, \quad G_k = \psi''_k(x), \quad k \geq 0,$$

and $\|h\|_G = \langle Gh, h \rangle^{1/2}$, $\|g\|^*_G = \langle g, G^{-1} g \rangle^{1/2}$.

Denote $A_k = \sum\limits_{i=0}^{k-1} a_i$. Clearly, $\boxed{\psi_k(x) \leq A_k \hat{f}(x) + \psi_0(x),\ x \in Q.}$

We can use the technique of estimate sequences!

Evolution of the Hessians

Since $\psi_k(x)$ are quadratic, their Hessians $G_k \succ 0$ are updated as

$$G_{k+1} = G_k + a_k \cdot f'(x_k) f'(x_k)^T = G_k + \frac{\delta}{1-\delta} \cdot \frac{f'(x_k) f'(x_k)^T}{(\| f'(x_k) \|^*_{G_k})^2}, \quad k \geq 0.$$

Therefore, $G_{k+1}^{-1} = G_k^{-1} - \delta \cdot \frac{G_k^{-1} f'(x_k) f'(x_k)^T G_k^{-1}}{(\| f'(x_k) \|^*_{G_k})^2}$.

Important: $\det G_{k+1} = \frac{1}{1-\delta} \det G_k = \frac{1}{(1-\delta)^{k+1}} \det G_0$.

Moreover,

$$\begin{aligned} \tfrac{1}{2} a_k^2 (\| \hat{f}'(x_k) \|^*_{G_{k+1}})^2 &= a_k^2 \cdot \hat{f}(x_k) \cdot (\| f'(x_k) \|^*_{G_{k+1}})^2 \\ &= a_k^2 \cdot \hat{f}(x_k) \cdot (1-\delta) \cdot (\| f'(x_k) \|^*_{G_k})^2 \\ &= \delta \cdot a_k \cdot \hat{f}(x_k). \end{aligned}$$

Main Lemma

For any $k \geq 0$, $\boxed{\psi_k^* \stackrel{\mathrm{def}}{=} \min\limits_{x \in Q} \psi_k(x) \;\geq\; (1-\delta)\sum\limits_{i=0}^{k-1} a_i \hat{f}(x_i).}$

Proof: Assume this is true for some $k \geq 0$. Since $\psi_k(x)$ quadratic,

$$\psi_k(x) = \psi_k^* + \langle \psi_k'(x_k), x - x_k\rangle + \tfrac{1}{2}\|x - x_k\|^2_{G_k} \geq \psi_k^* + \tfrac{1}{2}\|x - x_k\|^2_{G_k}.$$

Therefore,

$$\psi_{k+1}^* \geq \psi_k^* +$$

$$\min_{x \in Q}\left\{ \tfrac{1}{2}\|x - x_k\|^2_{G_k} + a_k[\hat{f}(x_k) + \langle \hat{f}'(x_k), x - x_k\rangle + \tfrac{1}{2}\langle f'(x_k), x - x_k\rangle^2] \right\}$$

$$= \psi_k^* + a_k \hat{f}(x_k) + \min_{x \in Q}\left\{ \tfrac{1}{2}\|x - x_k\|^2_{G_{k+1}} + a_k\langle \hat{f}'(x_k), x - x_k\rangle \right\}$$

$$\geq \psi_k^* + a_k \hat{f}(x_k) - \tfrac{1}{2}a_k^2\|\hat{f}'(x_k)\|^2_{G_{k+1}} \;=\; \psi_k^* + (1-\delta)\cdot a_k \hat{f}(x_k). \qquad \square$$

Rate of convergence

Denote $\tilde{x}_k = \frac{1}{A_k} \sum\limits_{i=0}^{k-1} a_i x_i$. Recall: $G_{k+1} = G_k + a_k \cdot f'(x_k) f'(x_k)^T$.

Theorem: Assume that for SP-function f, $\| f'(\cdot) \|^*_{G_0} \leq L$.

Then, $(1-\delta)\hat{f}(\tilde{x}_k) \leq \hat{f}(x^*) + \frac{L^2 \|x_0 - x^*\|^2_{G_0}}{2n\left[e^{\delta(k+1)/n} - 1\right]}$.

Proof: We have $(1-\delta)\hat{f}(x^*_k) \leq \hat{f}(x^*) + \frac{1}{2A_{k+1}} \|x_0 - x^*\|^2_{G_0}$.

Let us estimate the growth of A_k. Denote $\bar{G}_k = G_0^{-1/2} G_k G_0^{-1/2}$.

$$A_k = \sum_{i=0}^{k-1} a_i \geq \frac{1}{L^2} \sum_{i=0}^{k-1} a_i \|f'(x_i)\|^2_{G_0} = \frac{1}{L^2} \left[\mathrm{Trace}\, \bar{G}_k - n \right]$$

$$\geq \frac{n}{L^2} \left[\frac{1}{(1-\delta)^{k/n}} - 1 \right] \geq \frac{n}{L^2} \left[e^{\delta k/n} - 1 \right]. \qquad \square$$

Mixed accuracy

Definition: point $\bar{x} \in Q$ is a solution with *mixed* (ϵ, δ)-accuracy if

$$(1-\delta)\hat{f}(\bar{x}) \quad \leq \quad \hat{f}(x^*) + \epsilon.$$

- $\epsilon > 0$ serves as an absolute accuracy.
- $\delta \in (0,1)$ represents the relative accuracy.

Complexity: $N_n(\epsilon, \delta) \stackrel{\text{def}}{=} \frac{n}{\delta} \ln \left(1 + \frac{L^2 R^2}{2n \cdot \epsilon}\right)$ iterations of Q-N scheme.

Note:

- High absolute accuracy is *easy* to achieve.
- High relative accuracy is *difficult*. (No need?)
- $\#$ of iterations is proportional to $\frac{n}{\delta}$. (Compare with BSM.)
- We have a uniform bound: $N_n(\epsilon, \delta) < N_\infty(\epsilon, \delta) \stackrel{\text{def}}{=} \frac{L^2 R^2}{2\epsilon\delta}$.

Relative accuracy

Our goal: generate $\bar{x} \in Q$ satisfying $\quad (1-\delta) f(\bar{x}) \leq f^*$.

After k iterations of Q-N method, we have

$$\begin{aligned} (1-\delta)(f(x_k^*) - f^*) f^* & \leq (1-\delta)(\hat{f}(x_k^*) - \hat{f}(x^*)) \\ & \leq \delta \hat{f}(x^*) + \frac{L^2 R^2}{2n\left[e^{\delta(k+1)/n} - 1\right]} \overset{(?)}{\leq} \delta (f^*)^2. \end{aligned}$$

Thus, we need $k = R_n(\delta) \overset{\text{def}}{=} \frac{n}{\delta} \ln \left(1 + \frac{L^2 R^2}{n \delta (f^*)^2}\right)$ iterations.

Note:

- The main factor $\frac{n}{\delta}$ does not depend on the data. (*Fully polynomial approximation scheme.*)
- Dependence in n is the same as for optimal methods.
- Each iteration of Q-N method is very simple, same as in Ellipsoid Method (complexity $O(n^2 \ln \frac{LR}{\delta f^*})$ iterations).
- We have a uniform bound $R_n(\delta) < R_\infty(\delta) \overset{\text{def}}{=} \frac{L^2 R^2}{\delta^2 (f^*)^2}$.

Absolute accuracy

Our goal: for problem $\min_{x \in Q} \phi(x)$ find $\bar{x} \in Q: \ \phi(\bar{x}) \leq \phi^* + \epsilon$.
Assume ϕ has bounded subgradients and $\|x - x_0\| \leq R$, $\forall x \in Q$.
Define now a new SP-objective

$$\begin{aligned} f(x) &= \max\{\phi(x) - \phi(x_0) + 2LR, L\|x - x_0\|\} \\ &= \phi(x) - \phi(x_0) + 2LR \quad \forall x \in Q. \end{aligned}$$

Applying now Q-N method to $\hat{f}$, we get

$$\begin{aligned} \phi(x_k^*) - \phi^* &= f(x_k^*) - f^* \leq \tfrac{\delta f^*}{2(1-\delta)} + \tfrac{L^2 R^2}{2n\left[e^{\delta(k+1)/n} - 1\right] \cdot (1-\delta) f^*} \\ &\leq LR \left[\tfrac{\delta}{1-\delta} + \tfrac{1}{2n\left[e^{\delta(k+1)/n} - 1\right] \cdot (1-\delta)} \right]. \end{aligned}$$

Choice of parameters

Let us find $\delta = \delta(\epsilon)$ from equation

$$\frac{\delta}{1-\delta} \;=\; \frac{\epsilon}{2LR} \quad \Rightarrow \quad \delta(\epsilon) \;=\; \frac{\epsilon}{\epsilon+2LR}.$$

Then, we need at most

$$\begin{aligned} k \;&=\; R_n(\epsilon) \stackrel{\mathrm{def}}{=} \frac{n}{\delta(\epsilon)} \ln\left(1 + \frac{LR}{n\epsilon(1-\delta(\epsilon))}\right) \\ &=\; n\left(1 + 2\frac{LR}{\epsilon}\right) \cdot \ln\left(1 + \frac{\epsilon+2LR}{2n\epsilon}\right) \end{aligned}$$

iterations. At the same time,

$$R_n(\epsilon) \;<\; R_\infty(\epsilon) \;=\; \frac{1}{2}\left(1 + 2\frac{LR}{\epsilon}\right)^2.$$

Note: worst-case complexity bound of Q-N method is <u>always better</u> than that of the standard subgradient scheme.

Discussion

Schemes of Q-N methods look very natural.

1. Minimization in relative scale: (No parameters!)

$$x_{k+1} = \arg\min_{x \in Q} \left[\|x - x_0\|_{G_0}^2 + \tfrac{\delta}{1-\delta} \sum_{i=0}^{k} \tfrac{(f(x_i) + \langle f'(x_i), x - x_i \rangle)^2}{(\|f'(x_i)\|_{G_i}^*)^2} \right].$$

2. Minimization in absolute scale:

$$x_{k+1} = \arg\min_{x \in Q} \left[\|x - x_0\|_{G_0}^2 + \tfrac{\epsilon}{LR} \sum_{i=0}^{k} \tfrac{(\phi(x_i) - \phi(x_0) + \langle \phi'(x_i), x - x_i \rangle + 2LR)^2}{(\|\phi'(x_i)\|_{G_i}^*)^2} \right].$$

Compare: Dual gradient method

$$x_{k+1} = \arg\min_{x \in Q} \left[\|x - x_0\|_{G_0}^2 + \tfrac{2}{L_k} \sum_{i=0}^{k} (\phi(x_i) + \langle \phi'(x_i), x - x_i \rangle) \right].$$

for $C_L^{1,1}$: $L_k \equiv L$, and for $C_M^{1,0}$: $L_k \approx \sqrt{k} \cdot \frac{M}{R}$.

Revival of old questions

1. Roles of *dimension* and *accuracy* in complexity estimates.
Denote by $\mathcal{T}$ complexity of the oracle.

Available methods	Complexity
Subgradient*	$\frac{L^2R^2}{\epsilon^2} \cdot (n + \mathcal{T})$
Quasi-Newton	$\frac{nLR}{\epsilon} \ln \left(1 + \frac{LR}{n\epsilon}\right) \cdot (n^2 + \mathcal{T})$
Ellipsoids	$n^2 \ln \frac{LR}{\epsilon} \cdot (n^2 + \mathcal{T})$
Inscribed ellipsoids*	$n \ln \frac{LR}{\epsilon} \cdot (n^3 + \mathcal{T})$

If $\frac{1}{\epsilon} < n \ln \frac{1}{\epsilon}$ and $\mathcal{T} \approx n^2$, then QN is the best.

Questions:

1. What are the best methods for all spectrum of n, ϵ, and $\mathcal{T}$?
2. Are n, ϵ, and $\mathcal{T}$ really independent? $\Rightarrow$ Accuracy of the model?
   (Finite elements, Truss topology, Optimal Control, PDE, etc.)

Revival of old questions

2. Do we have a future for Quasi-Newton methods?

- Two decades of intensive research (60's, 70's).
  Good computational performance.
- r-algorithm by Shor for nonsmooth minimization.
- Excellence and failure of Ellipsoid Method.
  (Wrong application?)
- 25 years of complete silence.

Can we finally do a proper global complexity analysis for these schemes?