CORE DISCUSSION PAPER

2006/39

Cubic regularization of Newton's method for convex problems with constraints

Yu. Nesterov *

March 31, 2006

Abstract

In this paper we derive efficiency estimates of the regularized Newton's method as applied to constrained convex minimization problems and to variational inequalities. We study a one-step Newton's method and its multistep accelerated version, which converges on smooth convex problems as $O(\frac{1}{k^3})$, where k is the iteration counter. We derive also the efficiency estimate of a second-order scheme for smooth variational inequalities. Its global rate of convergence is established on the level $O(\frac{1}{k})$.

Keywords: convex optimization, variational inequalities, Newton's method, cubic regularization, worst-case complexity, global complexity bounds.

*Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 34 voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium; e-mail: nesterov@core.ucl.ac.be.

The research results presented in this paper have been supported by a grant "Action de recherche concertè ARC 04/09-315" from the "Direction de la recherche scientifique - Communautè française de Belgique". The scientific responsibility rests with its author.

1 Introduction

Motivation. Starting from the very beginning [1], the behavior of the Newton's method was studied mainly in a small neighborhood of non-degenerate solution (see [4]; a comprehensive exposition of the state of art in this field can be found in [2, 3]). However, the recent development of cubic regularization of the Newton's method [8] opened a possibility for global efficiency analysis of the second order schemes on different problem classes. After [8], the next step in this direction was done in [7]. Namely, it was shown that, on the class of smooth convex unconstrained minimization problems, the rate of convergence of one-step regularized Newton's method [8] can be improved by a multistep strategy from $O(\frac{1}{k^2})$ up to $O(\frac{1}{k^3})$, where k is the iteration counter. It is interesting that a similar idea is used for accelerating the usual gradient method for minimizing convex functions with Lipschitz-continuous gradient (see Section 2.2. in [6]).

In this paper we analyze the global efficiency of the second-order schemes as applied to convex constrained minimization problems and to variational inequalities. For generalizing the regularized Newton's step onto the constrained situation, we compute the next test point as an exact minimum of the *upper* second-order model of the objective function, taking into account the feasible set. This approach is based on the same idea as *gradient mapping*, which is employed in many first-order schemes (see, for example, Section 2.2.3 [6]).

Contents. In Section 2 we present different properties of *cubic regularization* of the support functions of convex sets. Our results can be seen as an extension of the standard approach based on regularization by *strongly convex* functions, which is intensively used in the first-order methods (see, for example, Section 2 [5]). In Section 3 we introduce a cubic regularization of the Newton step for a constrained variational inequality problem with sufficiently smooth monotone operator. Section 4 is devoted to the second-order schemes for constrained minimization problems. We derive the efficiency estimates for one-step regularized Newton's method and for its accelerated multistep variant. In Section 5 we present a second-order scheme for a variational inequality with smooth monotone operator. We show that the schemes converges with the rate $O(\frac{1}{k})$. We analyze also its efficiency as applied to strongly monotone operators. In the end of the section we prove the local quadratic convergence of the regularized Newton method. In the last Section 6 we discuss some implementation issues.

Notation. In what follows E denotes a finite-dimensional real vector space, and E^* the dual space, which is formed by all linear functions on E. The value of function $s \in E^*$ at $x \in E$ is denoted by $\langle s, x \rangle$.

Let us fix a positive definite self-adjoint operator $B : E \to E^*$. Define the following norms:

$$\begin{aligned} \|h\| &= \langle Bh, h \rangle^{1/2}, \quad h \in E, \\ \|s\|_* &= \langle s, B^{-1}s \rangle^{1/2}, \quad s \in E^*, \\ \|A\| &= \max_{\|h\| \le 1} \|Ah\|_*, \quad A : E \to E^*. \end{aligned}$$

For a self-adjoint operator $A = A^*$, the same norm can be defined as

$$\|A\| \;=\; \max_{\|h\| \le 1} |\langle Ah, h\rangle|. \tag{1.1}$$

Further, for function $f(x)$, $x \in E$, we denote by $\nabla f(x)$ its *gradient* at x:

$$f(x+h) \;=\; f(x) + \langle \nabla f(x), h\rangle + o(\|h\|), \quad h \in E.$$

Clearly $\nabla f(x) \in E^*$. Similarly, we denote by $\nabla^2 f(x)$ the *Hessian* of f at x:

$$\nabla f(x+h) \;=\; \nabla f(x) + \nabla^2 f(x)h + \mathbf{o}(\|h\|), \quad h \in E.$$

Of course, $\nabla^2 f(x)$ is a self-adjoint linear operator from E to E^*. We keep the same notation for gradients of functions defined on E^*. But then, of course, such a gradient is an element of E.

Finally, for a nonlinear operator $g : E \to E^*$, we denote by $g'(x)$ is *Jacobian*:

$$g(x+h) \;=\; g(x) + g'(x)h + \mathbf{o}(\|h\|), \quad h \in E.$$

Thus, $g'(x)$ can be seen as a linear operator from E to E^*. Clearly, $(\nabla f(x))' = \nabla^2 f(x)$.

2 Cubic regularization of support functions

Consider the following *cubic* prox function

$$d(x,y) \;=\; \tfrac{1}{3}\|x-y\|^3, \quad x, y \in E.$$

In accordance to Lemma 4 in [7], for any fixed $\bar{x} \in E$, we have

$$d(\bar{x},y) \;\ge\; d(\bar{x},x) + \langle \nabla_2 d(\bar{x},x), y-x\rangle + \tfrac{1}{6}\|y-x\|^3, \quad x, y \in E, \tag{2.1}$$

where ∇_2 denotes the gradient of corresponding function with respect to its second variable. Thus, $d(\bar{x},\cdot)$ is a uniformly convex on E function of degree $p = 3$ with convexity parameter $\sigma = \frac{1}{2}$.

Let Q be a closed convex set in E. We allow Q to be unbounded (for example, $Q \equiv E$). Let us fix some $x_0 \in Q$, which we treat as a *center* of this set. For our analysis we need to define two support-type functions of the set Q:

$$\begin{aligned} \xi_D(s) \;&=\; \max_{x \in Q} \left\{ \langle s, x - x_0\rangle :\; d(x) \le \tfrac{1}{3}D^3 \right\}, \quad s \in E, \\ W_\beta(x,s) \;&=\; \max_{y \in Q} \{\langle s, y-x\rangle - \beta d(x,y)\}, \qquad\quad s \in E, \end{aligned} \tag{2.2}$$

where x is an arbitrary point from Q, and parameters D and β are positive. The first function is a usual support function for the set

$$\mathcal{F}_D = \{x \in Q :\; \|x - x_0\| \le D\}.$$

The second one is a proximal-type approximation of the support function of set Q with respect to x. Since $d(x, \cdot)$ is uniformly convex (see (2.1)), for any positive D and β we have $\operatorname{dom} \xi_D = \operatorname{dom} W_\beta = E^*$. Note that both of the functions are nonnegative.

Let us mention some properties of function $W(\cdot, \cdot)$. If $\beta_2 \geq \beta_1 > 0$, then for any $x \in E$ and $s \in E^*$ we have

$$W_{\beta_2}(x, s) \leq W_{\beta_1}(x, s). \tag{2.3}$$

The support functions (2.2) are related as follows.

Lemma 1 *For any positive β and D, and any $s \in E^*$ we have*

$$\xi_D(s) \;\leq\; \tfrac{1}{3}\beta D^3 + W_\beta(x_0, s). \tag{2.4}$$

Proof:
Indeed,

$$\begin{aligned}
\xi_D(s) &= \max_y \{\langle s, y - x_0\rangle : \; y \in Q, \; d(x_0, y) \leq \tfrac{1}{3}D^3\} \\
&= \max_{y \in Q} \min_{\beta \geq 0} \{\langle s, y - x_0)\rangle + \tfrac{1}{3}\beta \cdot (D^3 - \|y - x_0\|^3))\} \\
&= \min_{\beta \geq 0} \max_{y \in Q} \{\langle s, y - x_0)\rangle + \tfrac{1}{3}\beta \cdot (D^3 - \|y - x_0\|^3))\} \\
&\leq \tfrac{1}{3}\beta D^3 + W_\beta(x_0, s).
\end{aligned}$$

□

We need some bounds on the rate of variation of function $W_\beta(x, s)$ in both arguments. Denote

$$\pi_\beta(x, s) = \arg\max_y \{\langle s, y - x\rangle - \beta d(x, y) : \; y \in Q\}.$$

Note that function $W_\beta(x, s)$ is differentiable in s and

$$\nabla_2 W_\beta(x, s) \;=\; \pi_\beta(x, s) - x.$$

Lemma 2 *Let us choose arbitrary $x \in Q$ and $s, \delta \in E^*$. Then*

$$W_\beta(x, s + \delta) \leq W_\beta(x, s) + \langle \delta, \nabla_2 W_\beta(x, s)\rangle + W_{\beta/2}(\pi_\beta(x, s), \delta). \tag{2.5}$$

Moreover, for any $y \in Q$ and $\delta \in E^*$ we have

$$W_\beta(y, \delta) \;\leq\; \tfrac{2}{3} \cdot \tfrac{1}{\sqrt{\beta}} \cdot \|\delta\|_*^{3/2}. \tag{2.6}$$

Proof:
Denote $\pi = \pi_\beta(x, s)$. From the first-order optimality condition for the second maximization problem in (2.2) we have

$$\langle s - \beta \cdot \nabla_2 d(x, \pi), \; y - \pi\rangle \leq 0 \quad \forall y \in Q.$$

Hence,

$$\langle s, y - x\rangle \leq \langle s, \pi - x\rangle + \beta \cdot \langle \nabla_2 d(x, \pi), y - \pi\rangle \quad \forall y \in Q.$$

Therefore,

$$\begin{aligned} W_\beta(x, s+\delta) &= \max_{y \in Q} \{\langle s+\delta, y-x\rangle - \beta \cdot d(x,y)\} \\ &\leq \max_{y \in Q} \{\langle \delta, y-x\rangle + \langle s, \pi - x\rangle + \beta \langle \nabla_2 d(x,\pi), y-\pi\rangle - \beta \cdot d(x,y)\} \\ &\overset{(2.1)}{\leq} \max_{y \in Q} \{\langle \delta, y-x\rangle + \langle s, \pi - x\rangle - \beta \cdot d(x,\pi) - \tfrac{1}{6}\beta \|y-\pi\|^3\} \\ &= W_{\beta/2}(\pi, \delta) + \langle \delta, \pi - x\rangle + W_\beta(x,s). \end{aligned}$$

It remains to note that

$$\begin{aligned} W_\beta(y,\delta) &= \max_{x \in Q}\{\langle \delta, x-y\rangle - \tfrac{1}{3}\beta \cdot \|x-y\|^3\} \\ &\leq \max_{x \in E}\{\langle \delta, x-y\rangle - \tfrac{1}{3}\beta \cdot \|x-y\|^3\} \;=\; \tfrac{2}{3} \cdot \tfrac{1}{\sqrt{\beta}} \cdot \|\delta\|_*^{3/2}. \end{aligned}$$

□

Thus, the level of smoothness of function $W_\beta(x, \cdot)$ can be controlled by the parameter β. This function can be used for measuring the size of the second argument. Of course, the result depends on the choice of the center x. Let us estimate from above the change of the measurement when the center moves.

Lemma 3 . *Let $x, \pi \in Q$, and $\alpha \in (0,1]$. Define $y = x + \alpha(\pi - x)$. Then, for any $\delta \in E^*$, and $\beta > 0$ we have*

$$W_\beta(\pi, \alpha\delta) \;\leq\; W_{\beta/\alpha^3}(y, \delta). \tag{2.7}$$

Proof:
Indeed,

$$\begin{aligned} W_\beta(\pi, \alpha\delta) &= \max_{v \in Q}\{\alpha\langle \delta, v-\pi\rangle - \beta d(\pi, v)\} \\ (y = x + \alpha(\pi - x)) &= \max_{v \in Q}\left\{\langle \delta, w-y\rangle - \tfrac{\beta}{3\alpha^3}\|w-y\|^3 :\; w = x + \alpha(v-x)\right\} \\ (x + \alpha(Q-x) \subseteq Q) &\leq \max_{w \in Q}\left\{\langle \delta, w-y\rangle - \tfrac{\beta}{3\alpha^3}\|w-y\|^3\right\} \;=\; W_{\beta/\alpha^3}(y,\delta). \end{aligned}$$

□

3 Cubic regularization of the Newton step

Consider a nonlinear differentiable monotone operator $g(x) : Q \to E^*$:

$$\langle g(x) - g(y), x - y\rangle \;\geq\; 0, \quad \forall x, y \in Q. \tag{3.1}$$

This condition is equivalent to positive semidefiniteness of its Jacobian:

$$\langle g'(x)h, h\rangle \ \geq \ 0, \quad \forall x \in Q, \ h \in E. \tag{3.2}$$

We assume also that the Jacobian of g is Lipschitz-continuous:

$$\|g'(x) - g'(y)\|_* \ \leq \ L\|x - y\|, \quad \forall x, y \in Q. \tag{3.3}$$

A well known consequence of this assumption is as follows:

$$\|g(y) - g(x) - g'(x)(y - x)\| \ \leq \ \tfrac{1}{2}L\|y - x\|^2, \quad x, y \in Q, \tag{3.4}$$

(see, for example, [9]).

An important example of such an operator is given by the gradient of the distance function:

$$\nabla_2 d(x, y) \ = \ \|y - x\| \cdot B(y - x), \quad x, y \in E,$$

with $L = 2$ (see Lemma 5 [7]).

For any $x \in Q$ and $M > 0$ we can define a regularized operator

$$U_{M,x}(y) \ = \ g(x) + g'(x)(y - x) + \tfrac{1}{2}M\nabla_2 d(x, y), \quad y \in E.$$

This is a uniformly monotone operator. Therefore, the following variational inequality:

$$\text{Find } T \in Q: \quad \langle U_{M,x}(T), y - T\rangle \ \geq \ 0 \quad \forall y \in Q, \tag{3.5}$$

has a unique solution $T \equiv T_M(x)$. Denote $r_M(x) = \|T_M(x) - x\|$, and

$$\Delta_M(x) \ = \ \langle g(x), x - T_M(x)\rangle - \tfrac{1}{2}\langle g'(x)(x - T_M(x)), x - T_M(x)\rangle - \tfrac{M}{6}r_M^3(x).$$

Using inequality (3.5) with $y = x$, we obtain

$$\langle g(x), x - T_M(x)\rangle - \langle g'(x)(x - T_M(x)), x - T_M(x)\rangle - \tfrac{M}{2}r_M^3(x) \ \geq \ 0.$$

Hence,

$$\begin{aligned} \Delta_M(x) \ &\geq \ \tfrac{1}{2}\langle g'(x)(x - T_M(x)), x - T_M(x)\rangle + \tfrac{M}{3}r_M^3(x) \\ &\overset{(3.2)}{\geq} \ \tfrac{M}{3}r_M^3(x). \end{aligned} \tag{3.6}$$

In what follows, we often use several simple consequences of (3.5).

Theorem 1 *Let operator g satisfy (3.2), (3.3). Then, for any $y \in Q$, and any positive λ and M, we have*

$$\langle g(T_M(y)), y - T_M(y)\rangle \ \geq \ \tfrac{M-L}{2} \cdot r_M^3(y), \tag{3.7}$$

$$\begin{aligned} W_\beta(y, -\lambda g(T_M(y))) \ &\leq \ \lambda\langle g(T_M(y)), y - T_M(y)\rangle \\ &\quad + \left(\sqrt{\tfrac{\lambda^3(L+M)^3}{18\beta}} - \lambda\tfrac{M-L}{2}\right) \cdot r_M^3(y). \end{aligned} \tag{3.8}$$

Proof:
Denote $T = T_M(y)$, and $r = r_M(y)$. Then, by (3.5),

$$\begin{aligned} 0 &\leq \langle g(y) + g'(y)(T-y) - g(T) + g(T) + \tfrac{1}{2}Mr(T-y), y-T\rangle \\ &= \langle g(y) + g'(y)(T-y) - g(T), y-T\rangle + \langle g(T), y-T\rangle - \tfrac{1}{2}Mr^3 \\ &\overset{(3.4)}{\leq} \tfrac{1}{2}Lr^3 + \langle g(T), y-T\rangle - \tfrac{1}{2}Mr^3, \end{aligned}$$

and that is (3.7).

Denote now $\bar{g} = g(y)$, and $\bar{g}' = g'(y)$. Then, for any $x \in Q$ we have:

$$\begin{aligned} \langle g(T), y-x\rangle &= \langle g(T) - \bar{g} - \bar{g}'(T-y), y-x\rangle + \langle \bar{g} + \bar{g}'(T-y), y-x\rangle \\ &\overset{(3.4)}{\leq} \tfrac{L}{2}r^2\|y-x\| + \langle \bar{g} + \bar{g}'(T-y), y-x\rangle \\ &= \tfrac{L}{2}r^2\|y-x\| + \langle \bar{g} + \bar{g}'(T-y), y-T\rangle + \langle \bar{g} + \bar{g}'(T-y), T-x\rangle \\ &\overset{(3.5)}{\leq} \tfrac{L}{2}r^2\|y-x\| + \langle \bar{g} + \bar{g}'(T-y), y-T\rangle + \tfrac{M}{2}r\langle B(T-y), x-T\rangle \\ &\leq \tfrac{L+M}{2}r^2\|y-x\| + \langle \bar{g} + \bar{g}'(T-y), y-T\rangle - \tfrac{M}{2}r^3 \\ &= \tfrac{L+M}{2}r^2\|y-x\| + \langle \bar{g} + \bar{g}'(T-y) - g(T), y-T\rangle \\ &\quad - \tfrac{M}{2}r^3 + \langle g(T), y-T\rangle \\ &\overset{(3.4)}{\leq} \tfrac{L+M}{2}r^2\|y-x\| - \tfrac{M-L}{2}r^3 + \langle g(T), y-T\rangle. \end{aligned}$$

Therefore, using this estimate (in the third line below), we obtain

$$\begin{aligned} & W_\beta(y, -\lambda g(T)) - \lambda\langle g(T), y-T\rangle \\ = \; & \max_{x\in Q}\{\lambda\langle g(T), y-x\rangle - \beta d(y,x) - \lambda\langle g(T), y-T\rangle\} \\ \leq \; & \max_{x\in Q}\{\lambda\tfrac{L+M}{2}r^2\|y-x\| - \lambda\tfrac{M-L}{2}r^3 - \tfrac{\beta}{3}\|y-x\|^3\} \\ \leq \; & \max_{\tau\geq 0}\{\lambda\tfrac{L+M}{2}r^2\tau - \tfrac{\beta}{3}\tau^3\} - \lambda\tfrac{M-L}{2}r^3 \\ = \; & r^3\cdot\left(\sqrt{\tfrac{\lambda^3(L+M)]^3}{18\beta}} - \lambda\tfrac{M-L}{2}\right). \end{aligned}$$

□

Denote $\kappa(M) = \frac{9(M-L)^2}{2(M+L)^3}$. Its maximal value is attained at $M = 5L$: $\kappa(5L) = \frac{1}{3L}$.

Corollary 1 *If $M > L$ and $0 < \lambda \leq \kappa(M) \cdot \beta$, then*

$$W_\beta(y, -\lambda g(T_M(y))) \;\; \leq \;\; \lambda \langle g(T_M(y)), y - T_M(y) \rangle. \tag{3.9}$$

4 Methods for constrained minimization

Consider the following minimization problem:

$$\min_x \{ f(x) : \; x \in Q \}, \tag{4.1}$$

where Q is a closed convex set and f is a convex function with Lipschitz continuous Hessian:

$$\| \nabla^2 f(x) - \nabla^2 f(y) \| \; \leq \; L \| x - y \|, \quad x, y \in Q. \tag{4.2}$$

Note that the operator $g(x) \stackrel{\text{def}}{=} \nabla f(x)$ satisfies conditions (3.2), (3.3). Thus, we can apply the results of Section 3 to the regularized Newton step $T = T_M(x)$ defined as a unique solution to the following variational inequality:

$$\langle \nabla f(x) + \nabla^2 f(x)(T - x) + \tfrac{1}{2} M \nabla_2 d(x, T), y - T \rangle \;\; \geq \;\; 0, \quad \forall y \in Q.$$

Note that now this point can be characterized in another way:

$$\begin{aligned} T_M(x) \;\; &= \;\; \arg\min_{y \in Q} \; \hat{f}_M(x, y), \\ \hat{f}_M(x, y) \;\; &\stackrel{\text{def}}{=} \;\; f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle + \tfrac{M}{6} \| y - x \|^3. \end{aligned} \tag{4.3}$$

In view of (4.2), for $M \geq L$ we have

$$f(y) \;\; \leq \;\; \hat{f}_M(x, y) \; \leq \; f(y) + \tfrac{L+M}{6} \| y - x \|^3, \quad x, y \in Q. \tag{4.4}$$

Therefore, $f(T_M(x)) \leq \hat{f}_M(x, T_M(X))$, and we obtain:

$$f(x) - f(T_M(x)) \;\; \geq \;\; f(x) - \hat{f}_M(x, T_M(X)) \; = \; \Delta_M(x) \;\; \stackrel{(3.6)}{\geq} \;\; \tfrac{M}{3} r_M^3(x). \tag{4.5}$$

Another important consequence of (4.4) is as follows:

$$f(T_M(x)) \;\; \leq \;\; \min_{y \in Q} \left\{ f(y) + \tfrac{L+M}{6} \| y - x \|^3 \right\} \; \leq \; f(x^*) + \tfrac{L+M}{6} \| x - x^* \|^3, \tag{4.6}$$

where x^* is an optimal solution to (4.1).

Consider now the following *regularized Newton's method*:

$$\boxed{\begin{array}{l} \text{Choose } x_0 \in Q \text{ and iterate} \\ \\ x_{k+1} \; = \; T_L(x_k), \quad k \geq 0. \end{array}} \tag{4.7}$$

Using the same arguments as in Theorem 1 in [7], we can prove the following statement.

Theorem 2 *Assume that the level sets of the problem (4.1) are bounded:*

$$\|x - x^*\| \leq D \quad \forall x \in Q : \; f(x) \leq f(x_0). \tag{4.8}$$

If the sequence $\{x_k\}_{k=1}^{\infty}$ *is generated by (4.7), then*

$$f(x_k) - f(x^*) \; \leq \; \frac{9LD^3}{(k+4)^2}, \quad k \geq 1. \tag{4.9}$$

Proof:
In view of (4.5), $f(x_{k+1}) \leq f(x_k)$, $k \geq 0$. Thus, $\|x_k - x^*\| \leq D$ for all $k \geq 0$. Further, in view of (4.6), we have

$$f(x_1) \; \leq \; f(x^*) + \tfrac{L}{3} D^3. \tag{4.10}$$

Consider now an arbitrary $k \geq 1$. Denote $x_k(\tau) = x^* + (1-\tau)(x_k - x^*)$. In view of the first inequality in (4.6), for any $\tau \in [0,1]$ we have

$$f(x_{k+1}) \;\; \leq \;\; f(x_k(\tau)) + \tau^3 \tfrac{L}{3} \|x_k - x^*\|^3 \; \leq \; f(x_k) - \tau(f(x_k) - f(x^*)) + \tau^3 \tfrac{LD^3}{3}.$$

The minimum of the right-hand side is attained for

$$\tau \;\; = \;\; \sqrt{\tfrac{f(x_k) - f(x^*)}{LD^3}} \; \leq \; \sqrt{\tfrac{f(x_1) - f(x^*)}{LD^3}} \overset{(4.10)}{<} 1.$$

Thus, for any $k \geq 1$ we have

$$f(x_{k+1}) \;\; \leq \;\; f(x_k(\tau)) - \tfrac{2}{3} \cdot \tfrac{(f(x_k) - f(x^*))^{3/2}}{\sqrt{LD^3}}. \tag{4.11}$$

Denote $\delta_k = f(x_k) - f(x^*)$. Then

$$\tfrac{1}{\sqrt{\delta_{k+1}}} - \tfrac{1}{\sqrt{\delta_k}} \;\; = \;\; \tfrac{\delta_k - \delta_{k+1}}{\sqrt{\delta_k \delta_{k+1}}(\sqrt{\delta_k} + \sqrt{\delta_{k+1}})} \overset{(4.11)}{\geq} \tfrac{2}{3\sqrt{LD^3}} \cdot \tfrac{\delta_k}{\sqrt{\delta_{k+1}}(\sqrt{\delta_k} + \sqrt{\delta_{k+1}})} \; \geq \; \tfrac{1}{3\sqrt{LD^3}}.$$

Thus, for any $k \geq 1$, we have

$$\tfrac{1}{\sqrt{\delta_k}} \;\; \geq \;\; \tfrac{1}{\sqrt{\delta_1}} + \tfrac{k-1}{3\sqrt{LD^3}} \overset{(4.10)}{\geq} \tfrac{1}{\sqrt{LD^3}} \cdot \left(\sqrt{3} + \tfrac{k-1}{3} \right) \; \geq \; \tfrac{k+4}{3\sqrt{LD^3}}.$$

□

Consider now an accelerated scheme. For simplicity, we assume that the constant L is known.

Choose some $x_0 \in Q$, and $s_0 = 0 \in E^*$. Define $\gamma = 27L$, $M = 5L$, and compute $x_1 = T_M(x_0)$.
For $k \geq 1$ **iterate:**
1. Update $s_k = s_{k-1} - \frac{k(k+1)}{2} \nabla f(x_k)$.
2. Compute $v_k = \pi_\gamma(x_0, s_k)$.
3. Select $y_k = \frac{k}{k+3} x_k + \frac{3}{k+3} v_k$.
4. Compute $x_{k+1} \; = \; T_M(y_k)$.

(4.12)

Define

$$\lambda_k = \tfrac{k(k+1)}{2}, \quad S_k = \tfrac{k(k+1)(k+2)}{6}, \quad k \geq 1.$$

Note that $S_k = \sum\limits_{i=1}^{k} \lambda_i$. Let us prove the following auxiliary result.

Lemma 4 *Define $\gamma = 27L$. Then, for any $k \geq 1$ we have*

$$\sum_{i=1}^{k} \lambda_i [f(x_i) + \langle \nabla f(x_i), x_0 - x_i \rangle] \quad \geq \quad S_k f(x_k) + W_\gamma(x_0, s_k). \tag{4.13}$$

Proof:
Note that for our choice of parameters, $\kappa(M) = \frac{9(M-L)^2}{2(M+L)^3} = \frac{1}{3L}$. Therefore, $\kappa(M)\gamma > 1$, and we have

$$\begin{aligned} S_1 f(x_1) + W_\gamma(x_0, s_1) \quad &= \quad f(x_1) + W_\gamma(x_0, -\nabla f(T_M(x_1))) \\ &\overset{(3.9)}{\leq} \quad \lambda_1 [f(x_1) + \langle \nabla f(x_1), x_0 - x_1 \rangle]. \end{aligned}$$

Thus, for $k = 1$, inequality (4.13) is valid.

Assume that (4.13) is true for some $k \geq 1$. Denote the left-hand side of this inequality by Σ_k. Then

$$\begin{aligned} \Sigma_{k+1} &= \Sigma_k + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_0 - x_{k+1} \rangle] \\ &\geq S_k f(x_k) + \lambda_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_0 - x_{k+1} \rangle] + W_\gamma(x_0, s_k) \\ &\geq S_{k+1} f(x_{k+1}) + \langle \nabla f(x_{k+1}), \lambda_{k+1} x_0 + S_k x_k - S_{k+1} x_{k+1} \rangle + W_\gamma(x_0, s_k) \\ &= S_{k+1} f(x_{k+1}) + \langle \nabla f(x_{k+1}), \lambda_{k+1}(x_0 - v_k) + S_{k+1}(y_k - x_{k+1}) \rangle + W_\gamma(x_0, s_k). \end{aligned} \tag{4.14}$$

Note that

$$\begin{aligned} W_\gamma(x_0, s_{k+1}) \quad &= \quad W_\gamma(x_0, s_k - \lambda_{k+1} \nabla f(x_{k+1})) \\ &\overset{(2.5)}{\leq} \quad W_\gamma(x_0, s_k) - \lambda_{k+1} \langle \nabla f(x_{k+1}), v_k - x_0 \rangle + W_{\frac{\gamma}{2}}(v_k, -\lambda_{k+1} \nabla f(x_{k+1})). \end{aligned}$$

Denote $\alpha_k = \frac{\lambda_{k+1}}{S_{k+1}} = \frac{3}{k+3}$. Then, in view of Step 3 in (4.12), we have

$$y_k = x_k + \alpha_k (v_k - x_k).$$

Hence, we can continue:

$$\begin{aligned} &W_\gamma(x_0, s_k) - \lambda_{k+1} \langle \nabla f(x_{k+1}), v_k - x_0 \rangle - W_\gamma(x_0, s_{k+1}) \\ &\geq -W_{\frac{\gamma}{2}}(v_k, -\lambda_{k+1} \nabla f(x_{k+1})) \overset{(2.7)}{\geq} -W_{\frac{\gamma}{2\alpha_k^3}}(y_k, -\tfrac{1}{\alpha_k} \lambda_{k+1} \nabla f(x_{k+1})) \\ &= -W_{\frac{\gamma}{2\alpha_k^3}}(y_k, -S_{k+1} \nabla f(x_{k+1})). \end{aligned} \tag{4.15}$$

Denote

$$\beta_k = \frac{\gamma}{2\alpha_k^3} = 27L \cdot \frac{(k+3)^3}{2 \cdot 3^3} = \frac{1}{2}L(k+3)^3.$$

Then $S_{k+1} \leq \frac{1}{6}(k+3)^3 = \kappa\beta_k$. Thus, conditions of Corollary 1 are satisfied and we can apply inequality (3.9):

$$-W_{\frac{\gamma}{2\alpha_k^3}}(y_k, -S_{k+1}\nabla f(x_{k+1})) \quad \geq \quad -S_{k+1}\langle \nabla f(x_{k+1}), y_k - x_{k+1} \rangle.$$

Using this estimate in (4.15), we obtain

$$W_\gamma(x_0, s_k) - \lambda_{k+1}\langle \nabla f(x_{k+1}), v_k - x_0 \rangle + S_{k+1}\langle \nabla f(x_{k+1}), y_k - x_{k+1} \rangle \geq W_\gamma(x_0, s_{k+1}).$$

Hence, in view of (4.14), we prove our inductive assumption for the next value $k+1$. □

Now we can easily estimate the rate of convergence of the accelerated process (4.12).

Theorem 3 *For any $k \geq 1$ we have*

$$f(x_k) - f(x^*) \ \leq \ \frac{54L\|x_0 - x^*\|^3}{k(k+1)(k+2)}. \tag{4.16}$$

Proof:
Indeed,

$$\sum_{i=1}^{k} \lambda_i [f(x_i) + \langle \nabla f(x_i), x_0 - x_i \rangle]$$

$$\stackrel{(4.13)}{\geq} S_k f(x_k) + \max_{x \in Q} \left\{ \langle s_k, x - x_0 \rangle - \tfrac{\gamma}{3}\|x - x_0\|^3 \right\}$$

$$\geq S_k f(x_k) + \sum_{i=1}^{k} \lambda_i \langle \nabla f(x_i), x_0 - x^* \rangle - \tfrac{\gamma}{3}\|x_0 - x^*\|^3.$$

Hence,

$$\begin{aligned} S_k f(x_k) \quad &\leq \quad \tfrac{\gamma}{3}\|x_0 - x^*\|^3 + \sum_{i=1}^{k} \lambda_i [f(x_i) + \langle \nabla f(x_i), x^* - x_i \rangle] \\ &\leq \quad \tfrac{\gamma}{3}\|x_0 - x^*\|^3 + S_k f(x^*). \end{aligned}$$

□

5 Second-order methods for variational inequalities

Let the nonlinear operator $g(x)$ satisfy conditions (3.2), (3.3). Consider the following *variational inequality* problem:

$$\text{Find } x^*: \quad \langle g(x^*), y - x^* \rangle \geq 0 \quad \forall y \in Q. \tag{5.1}$$

In order to measure quality of an approximate solution $y \in Q$ to this problem, we will employ the standard *restricted merit function*

$$f_D(x) = \max_{y \in Q} \{ \langle g(y), x - y \rangle : \ \|y - x_0\| \leq D \}. \tag{5.2}$$

It is easy to prove that for all $x \in \mathcal{F}_D$ this function is nonnegative. If $D \geq \|x_0 - x^*\|$, then $f_D(x^*) = 0$. Moreover, if $f_D(\hat{x}) = 0$ and $\|\hat{x} - x_0\| < D$, then $\hat{x}$ is a solution to (5.1) (see Lemma 1, [5]).

In order to form an approximate solution to (5.1), we often use the following *averaging procedure.*

Lemma 5 . *For a sequence of points $\{x_i\}_{i=1}^k \subset Q$ and a sequence of positive weights $\{\lambda_i\}_{i=1}^k$ define*

$$s_k = -\sum_{i=1}^k \lambda_i g(x_i), \quad S_k = \sum_{i=1}^k \lambda_i, \quad \hat{x}_k = \tfrac{1}{S_k} \sum_{i=1}^k \lambda_i x_i.$$

Then for any $\beta > 0$ we have

$$\begin{array}{rcl} f_D(\hat{x}_k) & \leq & \frac{1}{S_k} \left[\sum\limits_{i=1}^k \lambda_i \langle g(x_i), x_i - x_0 \rangle + \xi_D(s_k) \right] \\ & \overset{(2.4)}{\leq} & \frac{1}{S_k} \left[\frac{1}{3} \beta D^3 + \sum\limits_{i=1}^k \lambda_i \langle g(x_i), x_i - x_0 \rangle + W_\beta(x_0, s_k) \right]. \end{array} \tag{5.3}$$

In particular, if

$$\sum_{i=1}^k \lambda_i \langle g(x_i), x_0 - x_i \rangle \quad \geq \quad W_\beta(x_0, s_k), \tag{5.4}$$

then $f_D(\hat{x}_k) \leq \frac{\beta D^3}{3 S_k}$.

Proof:
Indeed,

$$\begin{array}{rcl} \sum\limits_{i=1}^k \lambda_i \langle g(x_i), x_i - x_0 \rangle + \xi_D(s_k) & \overset{(2.2)}{=} & \max\limits_{y \in Q} \left\{ \sum\limits_{i=1}^k \lambda_i \langle g(x_i), x_i - y \rangle : \ \|y - x_0\| \leq D \right\} \\ & \overset{(3.1)}{\geq} & \max\limits_{y \in Q} \left\{ \sum\limits_{i=1}^k \lambda_i \langle g(y), x_i - y \rangle : \ \|y - x_0\| \leq D \right\} \\ & = & S_k \cdot \max\limits_{y \in Q} \{ \langle g(y), \hat{x}_k - y \rangle : \ \|y - x_0\| \leq D \} \\ & = & S_k \cdot f_D(\hat{x}_k). \end{array}$$

□

Now we can estimate the rate of convergence of the following *dual* Newton's method (compare with (4.7)).

1. Choose $\beta = 6L$ and $M = 5L$.

Set $x_1 = T_M(x_0)$ and $s_0 = -g(x_1)$.

2. Iterate ($k \geq 1$):
$$\begin{aligned} s_k &= s_{k-1} - g(x_k), \\ v_k &= \pi_\beta(x_0, s_k), \\ x_{k+1} &= T_M(v_k). \end{aligned} \tag{5.5}$$

Theorem 4 *Let the operator $g(x)$ satisfy conditions (3.2), (3.3), and sequence $\{x_k\}_{k=1}^{\infty}$ be generated by (5.5). Then for the average points $\hat{x}_k = \frac{1}{k+1}\left[x_1 + \sum\limits_{i=1}^{k} x_i\right]$ we have*

$$f_D(\hat{x}_k) \;\leq\; \frac{2LD^3}{k+1}, \quad k \geq 1. \tag{5.6}$$

Proof:

Denote $\Delta_k = \langle g(x_1), x_1 - x_0\rangle + \sum\limits_{i=1}^{k} \langle g(x_i), x_i - x_0\rangle + W_\beta(x_0, s_k)$. In view of inequality (5.3), we need to prove that $\Delta_k \leq 0$ for all $k \geq 1$. Indeed, $s_1 = -2g(x_1)$, and, in view of our choice of parameters,

$$\kappa(M)\beta \geq 2. \tag{5.7}$$

Hence, applying inequality (3.9), we obtain

$$W_\beta(x_0, -2g(x_1)) \;\leq 2\langle g(x_1), x_0 - x_1\rangle.$$

Thus, $\Delta_1 \leq 0$. Assume now that $\Delta_k \leq 0$ for some $k \geq 1$. Then

$$\begin{aligned} \Delta_{k+1} &\leq \langle g(x_{k+1}), x_{k+1} - x_0\rangle + W_\beta(x_0, s_k - g(x_{k+1})) - W_\beta(x_0, s_k) \\ &\overset{(2.5)}{\leq} \langle g(x_{k+1}), x_{k+1} - \pi_\beta(x_0, s_k)\rangle + W_{\beta/2}(\pi_\beta(x_0, s_k), -g(x_{k+1})) \\ &\overset{(5.5)}{=} \langle g(x_{k+1}), x_{k+1} - v_k\rangle + W_{\beta/2}(v_k, -g(x_{k+1})). \end{aligned}$$

In view of (5.7) and relation (3.9), we conclude that the right-hand side of the latter inequality is non-positive. □

Assume now that the operator $g(x)$ is *strongly monotone*:

$$\langle g(x) - g(y), x - y\rangle \;\geq\; \mu\|x - y\|^2, \quad x, y \in Q. \tag{5.8}$$

For differentiable operators, this condition is equivalent to uniform nondegeneracy of the Jacobian $g'(x)$:

$$\langle g'(x)h, h\rangle \;\geq\; \mu\|h\|^2, \quad x \in Q,\; h \in E. \tag{5.9}$$

Lemma 6 *Assume that operator $g(x)$ satisfies (5.8) and $D \geq \|x_0 - x^*\|$. Then for any $x \in \mathcal{F}_D$ we have*

$$f_D(x) \;\geq\; \tfrac{1}{4}\mu\|x - x^*\|^2. \tag{5.10}$$

Proof:
Indeed, consider $y = \frac{1}{2}x + \frac{1}{2}x^* \in \mathcal{F}_D$. Then

$$\begin{aligned} f_D(x) &\overset{(5.2)}{\geq} \langle g(y), x - y\rangle \;=\; \langle g(y), y - x^*\rangle \\ &\overset{(5.8)}{\geq} \langle g(x^*), y - x^*\rangle + \mu\|y - x^*\|^2 \\ &\overset{(5.1)}{\geq} \mu\|y - x^*\|^2 \;=\; \tfrac{1}{4}\mu\|x - x^*\|^2. \end{aligned}$$

□

Thus, using (5.5) we can get close to a solution of the variational inequality with strongly monotone operator. Let us show that in a small neighborhood of the solution, the transformation $T_M(x)$ ensures a quadratic rate of convergence.

Theorem 5 *Let operator $g(x)$ satisfy conditions (3.3) and (5.8). Then for any $M \geq 0$ the process*

$$\boxed{x_{k+1} \;=\; T_M(x_k), \quad k \geq 0,} \tag{5.11}$$

converges quadratically:

$$\|x_{k+1} - x^*\| \;\leq\; \tfrac{L+M}{2\mu}\|x_k - x^*\|^2. \tag{5.12}$$

Proof:
Denote $r_k = \|x_k - x_{k+1}\|$, and $\rho_k = \|x_k - x^*\|$. Then

$$\begin{aligned} 0 &\overset{(5.1)}{\leq} \langle g(x^*), x_{k+1} - x^*\rangle \\ &= \langle g(x^*) - g(x_k) - g'(x_k)(x^* - x_k), x_{k+1} - x^*\rangle \\ &\quad + \langle g(x_k) + g'(x_k)(x^* - x_k), x_{k+1} - x^*\rangle \\ &\overset{(3.4)}{\leq} \tfrac{L}{2}\rho_k^2\rho_{k+1} + \langle g(x_k) + g'(x_k)(x_{k+1} - x_k) + g'(x_k)(x^* - x_{k+1}), x_{k+1} - x^*\rangle \\ &\overset{(3.5)}{\leq} \tfrac{L}{2}\rho_k^2\rho_{k+1} + \tfrac{M}{2}r_k\langle B(x_{k+1} - x_k), x^* - x_{k+1}\rangle - \langle g'(x_k)(x_{k+1} - x^*), x_{k+1} - x^*\rangle \\ &\overset{(5.9)}{\leq} \tfrac{L}{2}\rho_k^2\rho_{k+1} + \tfrac{M}{2}r_k\langle B(x_{k+1} - x_k), x^* - x_{k+1}\rangle - \mu\rho_{k+1}^2. \end{aligned}$$

Thus,

$$\mu\rho_{k+1}^2 \;\leq\; \tfrac{L}{2}\rho_k^2\rho_{k+1} + \tfrac{M}{2}r_k\langle B(x_{k+1} - x_k), x^* - x_{k+1}\rangle \;\leq\; \tfrac{L}{2}\rho_k^2\rho_{k+1} + \tfrac{M}{2}r_k^2\rho_{k+1},$$

which gives

$$\mu\rho_{k+1} \;\leq\; \tfrac{L}{2}\rho_k^2 + \tfrac{M}{2}r_k^2. \tag{5.13}$$

On the other hand, we can continue our main chain of inequalities as follows:

$$\begin{aligned} 0 \;&\leq\; \tfrac{L}{2}\rho_k^2\rho_{k+1} + \tfrac{M}{2}r_k\langle B(x_{k+1}-x_k), x^*-x_k+x_k-x_{k+1}\rangle - \mu\rho_{k+1}^2 \\ &\leq\; \tfrac{L}{2}\rho_k^2\rho_{k+1} + \tfrac{M}{2}r_k^2\rho_k - \tfrac{M}{2}r_k^3 - \mu\rho_{k+1}^2. \end{aligned}$$

Therefore, if $r_k \geq \rho_k$, then $\rho_{k+1} \leq \frac{L}{2\mu}\rho_k^2$. Otherwise, we get (5.12) from (5.13). □

Thus, the region of quadratic convergence of method (5.11) can be defined as

$$\mathcal{Q}_D \;=\; \left\{ x \in \mathcal{F}_D :\; \|x - x^*\| \leq \tfrac{2\mu}{L+M} \right\} \;\overset{(5.10)}{\supseteq}\; \left\{ x \in \mathcal{F}_D :\; f_D(x) \leq \tfrac{\mu^3}{(L+M)^2} \right\}.$$

Therefore, in view of (5.6), the analytical complexity of finding a point from $\mathcal{Q}_D$ by a single run of the method (5.5) is bounded by

$$O\left(\left[\tfrac{LD}{\mu}\right]^3\right)$$

iterations. However, the same goal can be achieved much more efficiently by a *restarting strategy*. Indeed, from (5.6) and (5.10), we see that this scheme halves the distance to the optimum in $O\left(\frac{LD}{\mu}\right)$ iterations. After that, we can restart the algorithm, taking the last point of the first stage as a starting point for the second one, etc. Since the length of a stage depends linearly on the initial distance to the optimum, the duration of any next stage will be twice smaller than that of the previous stage. Hence, the total number of iterations in all stages is of the order

$$O\left(\tfrac{LD}{\mu}\right). \tag{5.14}$$

Note that for optimization problems, in view of a much higher rate of convergence (4.16), the complexity bound drops to the level

$$O\left(\left[\tfrac{LD}{\mu}\right]^{1/3}\right) \tag{5.15}$$

iterations.

6 Discussion

At each iteration of the regularized Newton's method we need to solve the auxiliary problem (3.5). Let us analyze its complexity for the constrained optimization. In this case, the problem (3.5) can be written in the following form:

$$\min_{y \in Q} \left\{ \langle g, y-x\rangle + \tfrac{1}{2}\langle G(y-x), y-x\rangle + \tfrac{M}{6}\|y-x\|^3 \right\}, \tag{6.1}$$

where g is the gradient of the objective function at x and G is the Hessian. A non-standard feature of this problem consists in the presence of the *cubic* term in the objective function. Let us show that the problem (6.1) can be solved by a sequence of *quadratic* minimization problems.

Indeed, note that for any $r > 0$ we have

$$\frac{1}{3}r^3 \;=\; \max_{\tau \geq 0} \left[r^2\tau - \frac{2}{3}\tau^{3/2} \right].$$

Therefore, the problem (6.1) can be written in the dual form:

$$\min_{y \in Q} \max_{\tau \geq 0} \left\{ \langle g, y - x \rangle + \frac{1}{2}\langle G(y-x), y-x \rangle + \frac{M}{2} \left(\tau \|y - x\|^2 - \frac{2}{3}\tau^{3/2} \right) \right\}$$

$$= \max_{\tau \geq 0} \left(-\frac{M}{3}\tau^{3/2} + \underbrace{\min_{y \in Q} \left\{ \langle g, y - x \rangle + \frac{1}{2}\langle (G + \tau M)(y-x), y-x \rangle \right\}}_{\phi(\tau)} \right).$$

Note that the function $\phi(\tau)$ is defined by a *quadratic* minimization problem, which very often can be solved efficiently. On the upper level, we have a problem of maximizing a concave univariate function.

References

[1] Bennet A.A., Newton's method in general analysis, *Proc. Nat. Ac. Sci. USA*, 1916, **2**, No 10, 592–598.

[2] Conn A.B., Gould N.I.M., Toint Ph.L., *Trust Region Methods*, SIAM, Philadelphia, 2000.

[3] Dennis J.E., Jr., Schnabel R.B., *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, SIAM, Philadelphia, 1996.

[4] Kantorovich L.V., Functional analysis and applied mathematics, *Uspehi Matem. Nauk*, 1948, **3**, No 1, 89–185, (in Russian), translated as N.B.S. Report 1509, Washington D.C., 1952.

[5] Nesterov Yu., Dual extrapolation and its applications to solving variational inequalities and related problems. CORE Discussion Paper #2003/68, (2003). Accepted by *Mathematical Programming*.

[6] Nesterov Yu., *Introductory lectures on convex programming. A basic course*, Kluwer, Boston, 2004.

[7] Nesterov Yu., Accelerating the cubic regularization of Newton's method on convex problems. CORE Discussion Paper #2005/68, (2005).

[8] Nesterov Yu., Polyak B., Cubic regularization of Newton method and its global performance. CORE Discussion paper # 2003/41, (2003). Accepted by *Mathematical Programming*.

[9] Ortega J.M., Rheinboldt W.C., *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, NY, 1970.