Newton Method with Adaptive Step-Size for Under-Determined Systems of Equations

Boris T. Polyak Andrey A. Tremba

V.A. Trapeznikov Institute of Control Sciences RAS, Moscow, Russia
Profsoyuznaya, 65, 117997 Moscow, Russia
boris@ipu.ru, atremba@ipu.ru

Abstract

Newton method is a well-known tool for solving finite-dimensional systems of equations. Pure Newton-Raphson method has at least quadratic convergence rate, but its convergence radius is often limited. Damped version of Newton method uses smaller step-size with same direction, with larger convergence ball but linear convergence rate. We propose mixed step-size choice strategy, incorporating both quadratic convergence rate and wide (global in some cases) convergence radius. The method can be used in cases of under-determined equations and Banach-space equations. We present a modification of proposed method requiring no a-priori knowledge of problem constants as well. The method may be also used for solving a class of non-convex optimization problems.

1 Introduction

We aim to explore solvability conditions of generic equation

$$P(x) = 0, \tag{1}$$

with operator $P : \mathbb{R}^n \to \mathbb{R}^m$, $m \leq n$. Most results of this note are valid for Banach spaces as well, but we stick to the finite-dimensional setting for simplicity. Through the paper we assume existence of derivative P' of the operator in a ball $B_\rho = \{x \colon \|x - x_0\| \leq \rho\}$. We also assume that the derivative is Lipschitz with constant L in the ball:

$$\|P'(x_a) - P'(x_b)\| \leq L\|x_a - x_b\|, \ \ \forall x_a, x_b \in B_\rho. \tag{2}$$

Solvability of the equation (1) can be expressed in terms of conditions on a given point $x_0 : P(x_0) \neq 0$ solely, or in a set around condition point, e.g. B_ρ.

These additional conditions are related to non-degeneracy P, which may be expressed through its derivative P'. There are two types of conditions: point-wise

$$\|P'(x_0)^T h\|_* \geq \mu_0 \|h\|_*, \ \ \mu_0 > 0, \ \ \forall h \in \mathbb{R}^m, \tag{3}$$

and set-wise

$$\|P'(x)^T h\|_* \geq \mu \|h\|_*, \ \ \mu > 0, \ \ \forall h \in \mathbb{R}^m, \ \ \forall x \in B_\rho. \tag{4}$$

In: Yu. G. Evtushenko, M. Yu. Khachay, O. V. Khamisov, Yu. A. Kochetov, V.U. Malkova, M.A. Posypkin (eds.): Proceedings of the OPTIMA-2017 Conference, Petrovac, Montenegro, 02-Oct-2017, published at http://ceur-ws.org

Both conditions are generalization of notion "existence of inverse matrix", and valid for under-determined (also Banach) systems of equation. In order, this notion can be generalized for non-differentiable and infinite-dimensional spaces and is known as "metric regularity". Thus algorithms discussed in this note are applicable to Banach spaces as well.

For example, let norms in image and pre-image be Euclidean ones. Then Lipschitz constant of the derivative is coupled with spectral operator norm, and (3) simply states that the least singular value of matrix $P'(x_0)$ is positive: $\mu_0 = \sigma_m(P'(x_0)) > 0$. While the former property is easier to calculate, it leads to conservative, but elegant results. Most known one is Newton-Kantorovich theorem related to convergence of Newton method started at x_0.

Main goal of this note is to explore how convergence property of Newton method depend on "operational" ball size ρ. It appears that if we modify Newton method itself, then we can achieve considerable and provable improvement over known results.

1.1 Newton-Kantorovich and Mysovskikh Theorems

One of most practical generic algorithms for solving equations is Newton (Newton-Raphson) method, which uses idea of operator linearization. Let's write it down in slightly general than usual form, e.g. as in constraint step [Hager, 1993]:

$$\begin{aligned} z_k &= \arg \min_{P'(x_k)z=P(x_k)} \|z\|, \\ x_{k+1} &= x_k - z_k. \end{aligned} \tag{5}$$

Vector z_k is also called Newton direction. This scheme embrace both standard and under-determined systems. It has explicit expression in $n = m$ case for non-degenerate derivative as $z_k = (P'(x_k))^{-1}P(x_k)$. Also in $m \neq n$ Euclidean case it can be expressed via Moore-Penrose pseudo-inverse as $z_k = P'(x_k)^{\dagger}P(x_k)$ by [Ben-Israel, 1966].

It is known that if Newton method converges, then its convergence rate is quadratic $\|P(x_{k+1})\| \leq c_1\|P(x_k)\|^2$. Famous Newton-Kantorovich theorem impose semi-local conditions at one point x_0, ensuring that Newton method (5) converges [Kantorovich et al, 1982]. The theorem initially assumes existence of inverse operator $(P'(x))^{-1}$, which can be relaxed to Lipschitz condition (2). The following result is slight modification of [Robinson, 1972, Theorem 2].

Proposition 1. *Assume* (2) *and* (3) *hold. If* $\|P(x_0)\| \leq s$,

$$h = \frac{L}{\mu_0^2}s < \frac{1}{2}, \quad \textit{and } r_0 = \frac{1-\sqrt{1-2h}}{h}\frac{s}{\mu_0} \leq \rho$$

Then Newton method (5) *converges to a solution of* (1).

Neglectable difference is that Robinson used more general setup, and used norm of *right* inverse operator $\|(P'(x_0))^{-1}_{right}\|$ as a constant. It appears that the norm is equal to $1/\mu_0$. Review of [Galántai, 2000] includes other means for solution of under-determined equations, including generalized inverse, outer inverse of P', etc. The Proposition's statement is in same form as Newton-Kantorovich theorem, with following changes. Condition on inverse to derivative operator is replaced with condition (3), also condition on *second* derivative being replaced by (2). Kantorovich demonstrated that the theorem is unimprovable without further assumptions on operator P at points *other* than x_0. There are multiple results on how exactly the method converges, cf. historical review [Yamomoto, 2000].

Constant h can be improved, if we assume strong non-degeneracy not only at x_0, but on whole ball B_ρ, i.e. having assumed (4) instead of (3). It is described by another famous theorem by [Mysovskikh, 1949] (Newton-Mysovskikh theorem, see also [Kantorovich et al, 1982]), and updated by [Polyak, 1964] to Banach and non-exact cases. Applied to Newton method (5), an excerpt from [Polyak, 1964, Corollary 1] takes form:

Proposition 2. *Assume* (2) *and* (4) *hold. Let* $\|P(x_0)\| \leq s$. *Then if*

$$h = \frac{L}{\mu^2}s < 2 \quad \textit{and } r_1 = \frac{2\mu}{L}H_0\left(\frac{h}{2}\right) < \rho,$$

then Newton method (5) *converges to a solution* x_*, *and*

$$\|x_* - x_k\| \leq \frac{2\mu}{L}H_k\left(\frac{h}{2}\right).$$

Here used sum-of-double-exponentials function $H_k(\delta) = \sum_{\ell=k}^{\infty} \delta^{(2^\ell)}$ defined on $\delta \in [0,1)$. Below we also use inverse of the first function in the series: $\Delta : [0,\infty) \to [0,1)$, s.t. $\Delta(H_0(\delta)) \equiv \delta,\ \delta \in [0,1)$. All these functions are monotonically increasing and are easily computed with needed accuracy.

1.2 Convergence Radius

In the paper we study *convergence radius* of Newton method, i.e. conditions of type $\|P(x_0)\| \le s$, maximizing image radius s. This also determines a *convergence ball* in image space.

This is very useful in study of equation variability. Consider a function $g : \mathbb{R}^n \to \mathbb{R}^m$ and known solution $x_a : g(x_a) = 0$. The problem of interest is to study range of right-hand side y of equation

$$g(x) = y, \tag{6}$$

such that the equation still has a solution. Other problem is to estimate size of a small ball image $\{g(x) : x \in B_\varepsilon\}$. This problem is trivially converted to (1) by substitution and shift of coordinates $P(x) \equiv g(x) - y,\ \ x_0 \equiv x_a$ with $\|P(x_0)\| = \|y\|$. If equation (1) has solution (proven, say, by starting Newton method from x_0!) for all $\|P(x_0)\| \le s$, then equation (6) has a solution for any $y : \|y\| \le s$. Note that $g(\cdot)$ has exactly the same derivative as $P(\cdot)$.

In order to estimate convergence radius, conditions of Proposition 1 can be explicitly inverted.

Corollary 1. *Under conditions* (2), (3), *if*

$$\|P(x_0)\| \le \begin{cases} \dfrac{\mu_0^2}{2L}\left(1 - \left(1 - \dfrac{L}{\mu_0}\rho\right)^2\right), & \rho \le \dfrac{\mu_0}{L}, \\ \dfrac{\mu_0^2}{2L}, & \rho > \dfrac{\mu_0}{L}. \end{cases}$$

then Newton method (5) *converges to a solution of* (1).

Similarly, conditions of Newton-Mysovskikh theorem may be expressed in terms of convergence radius.

Corollary 2. *Under conditions* (2), (4), *if*

$$\|P(x_0)\| \le \frac{2\mu^2}{L}\Delta\left(\frac{L}{2\mu}\rho\right).$$

then Newton method (5) *converges to a solution of* (1).

In both cases theoretical applicability of *pure* (non-damped) Newton method is limited by a constant (maximal convergence radius is $\frac{\mu_0^2}{2L}$ or $\frac{2\mu^2}{L}$), independently on operational ball radius ρ. In main Section 2 we propose an algorithm, which increase convergence ball in same conditions as Proposition 2.

1.3 Damped Newton Method

Another very popular and practical choice is to make non-unit step-size in Newton direction, introducing damping parameter $\gamma_k \le 1$.

$$\begin{aligned} z_k &= \arg \min_{P'(x_k)z = P(x_k)} \|z\|, \\ x_{k+1} &= x_k - \gamma_k z_k. \end{aligned} \tag{7}$$

Newton method is tightly connected with equivalent minimization problem

$$\|P(x)\| \to \min_x,$$

and Newton direction is descent direction for the functional. Thus for appropriately small step-sizes γ_k follows $\|P(x_k - \gamma_k z_k)\| \le \|P(x_k)\|$, and $\{x_k\}$ is relaxation sequence for the functional. There are many strategies for the step-size γ, e.g. well-known Armijo backtracking rule $\gamma_i = \gamma_0 2^{-i}$, shrinkage $\gamma_i = \gamma_0 c_2^i,\ \ 0 < c_2 < 1$ or direct line-search [Ortega & Rheinboldt, 2000]. Typically damping step-size search terminates as soon following condition is satisfied for some constant c:

$$\|P(x_k - \gamma_k z_k)\| \le (1 - c\gamma_k)\|P(x_k)\|.$$

Damped Newton method may be globally converging, but its convergence rate is typically linear at first $\|P(x_{k+1})\| \le c_2\|P(x_k)\|$, but eventually leading to quadratic convergence rate. This was first demonstrated by [Pshenichnyi, 1970]. Unfortunately, explicit conditions for convergence radius and convergence rate were not studied. We also mention recent paper [Nesterov, 2007], where appropriate damping paired with modified Newton direction, obtained from an auxiliary regularized problem.

In the note we are to provide some simple and "natural" strategy for step-size choice. Then we prove both quadratic convergence rate of resulting algorithm, and larger convergence radius. Our basic method is using specific constants of a problem, which may be unknown a-priori. For overcoming this, we propose modification of the algorithm, which is adaptive to the constants. We demonstrate both methods on systems of quadratic equations, allowing simpler statements. It turned out that the same strategy was proposed by [Deuflhard, 2004], resulting in interesting "global" results, but convergence radius were still unexplored.

2 Choosing Step-Size for Newton Method

We consider vector equation (1) with $m \le n$, with initial point x_0 and (damped) Newton method (7). Key observation is following: under assumptions (2) and (4) norm of Newton direction vector is limited as $\|z_k\| \le \frac{1}{\mu}\|P(x_k)\|$ on $x \in B_\rho$. Then target functional $\|P(x)\|$, reflecting equation residual, has upper estimate

$$\|P(x_{k+1})\| \le \Big(|1-\gamma| + \frac{L}{2\mu^2}\gamma^2\|P(x_k)\|\Big)\|P(x_k)\|,$$

obtained by linearization due to differentiability on B_ρ. Optimal choice for minimizing the estimate is

$$\gamma_k = \min\Big\{1,\ \frac{\mu^2}{L\|P(x_k)\|}\Big\}. \tag{8}$$

It is one of two possible choices, other uses only Lipschitz constant L in step-size calculation [Polyak & Tremba].

The step-size depend on value $\|P(x_k)\|$, and if it is small enough, method becomes pure Newton method (5).

Theorem. *Let* (2) *and* (4) *hold. If*

$$\|P(x_0)\| \le \begin{cases} \dfrac{2\mu^2}{L}\Delta\Big(\dfrac{L}{2\mu}\rho\Big), & \rho \le 2\dfrac{\mu}{L}H_0\Big(\dfrac{1}{2}\Big), \\ \dfrac{\mu^2}{L}\left(1+\dfrac{1}{2}\Big\lfloor\dfrac{L\rho}{\mu} - 2H_0\Big(\dfrac{1}{2}\Big)\Big\rfloor\right), & \rho > 2\dfrac{\mu}{L}H_0\Big(\dfrac{1}{2}\Big), \end{cases} \tag{9}$$

then Newton algorithm (7) *with step-size* (8) *converges to a solution* x_* *of* (1)*. Also*

$$\|P(x_k)\| \le \begin{cases} \|P(x_0)\| - \dfrac{\mu^2}{2L}k, & k < k_{\max}, \\ \dfrac{2\mu^2}{L}2^{-(2^{(k-k_{\max})})}, & k \ge k_{\max}, \end{cases}$$

$$\|x_k - x_*\| \le \begin{cases} \dfrac{\mu}{L}(k_{\max} - k + 2H_0\Big(\dfrac{1}{2}\Big)), & k < k_{\max}, \\ \dfrac{2\mu}{L}H_{k-k_{\max}}\Big(\dfrac{1}{2}\Big), & k \ge k_{\max}. \end{cases} \tag{10}$$

where

$$k_{\max} = \max\Big\{0,\ \Big\lceil\frac{2L}{\mu^2}\|P(x_0)\|\Big\rceil - 2\Big\}. \tag{11}$$

This theorem statement is twofold. First, maximal number of non-unit step lengths is explicitly given. It allows to predict when linear convergence rate switches to quadratic one. Of course, total convergence rate still quadratic.

Second, the last part of (9) states potentially *unlimited* (or even global) convergence radius.

Sketch of proof. *Step length* (8) *results in non-increasing sequence of* $\|P(x_k)\|$ *(actually, monotonically decreasing whenever* $\|P(x_k)\| > 0$*). In first phase* $\gamma_k < 1$*, and* $\|P(x_{k+1})\| \le \|P(x_k)\| - \frac{\mu^2}{2L}$*, until threshold* $\|P(x_k)\| \le \frac{\mu^2}{L}$

met. This limits number of steps as (11). *In second phase,* $\gamma_k \equiv 1$, *and standard analysis of Newton steps is valid.*

Crucial part of proof is estimating distances $\|x_0 - x_k\|$ *(and eventually* $\|x_0 - x_*\|$*), which should guarantee sequence* $\{x_k\}$ *be* within *operational ball* B_ρ. *This distance implicitly depends (through* $k_{\max}$*) on initial condition as* (10). *Inverting this condition by limiting* $\|x_0 - x_*\| \le \rho$ *leads to convergence radius* (9).

Main drawback of the step length choice (8) is dependence on constants L, μ or its estimates. We propose a modification of the method, which omits this requirements.

2.1 Adaptive Constant Estimation

The idea is very simple. We are to notice that both constants L, μ enter step-size formulae (8) in combined fashion as μ^2/L, and thus may be estimated by backtracking. Let $\beta = \mu^2/L$ be "true" constant. Assume that we have some initial value β_0. There it can be either smaller than β, or greater.

If we start algorithm with $\beta_0 \le \beta$, then exist constants $\widehat{L} \ge L, \widehat{\mu} \le \mu$, such that $\widehat{\mu}^2/\widehat{L} = \beta_0$ and properties (2) and (4) hold for $\widehat{L}$ and $\widehat{\mu}$. Thus Theorem 2 is valid with respect to these constants.

Other case of $\beta_0 \ge \beta$ is trickier. Here we check whether $P(x_k - \gamma_k z_k)$ *sufficiently* decreases or not. If it is decreasing well enough, we accept this step-size, otherwise update $\beta_0 \to \beta_1 < \beta_0$. Let's demonstrate, how exactly it is done.

2.1.1 Algorithm (Constant-Adaptive Newton Method)

As mentioned, the algorithm takes as input initial estimate β_0 and scalar multiplier $0 < q < 1$. The algorithm is initialized with counter $k = 0$ and number $p_0 = \|P(x_0)\|$.

1. Solve
$$z_k = \arg \min_{P'(x_k)z = P(x_k)} \|z\|,$$
2. Evaluate
$$\gamma_k = \min\left\{1, \frac{\beta_k}{p_k}\right\},$$
$$p_{k+1} = \|P(x_k - \gamma_k z_k)\|.$$
3. If either
$$\gamma_k < 1 \text{ and } p_{k+1} < p_k - \frac{\beta_k}{2},$$
or
$$\gamma_k = 1 \text{ and } p_{k+1} < \frac{1}{2\beta_k} p_k,$$
holds, then go to Step 5. Otherwise
4. apply update rule $\beta_k \leftarrow q\beta_k$ and return to Step 2 without increasing counter.
5. Take
$$x_{k+1} = x_k - \gamma_k z_k,$$
set $\beta_{k+1} = \beta_k$, increase counter $k \leftarrow k + 1$, and go to Step 1.

Note that in contrast to known algorithm, we use backtracking search over *parameter* of the method, but not over *step length*.

Meanwhile the algorithm behaves nice in practice, its convergence analysis becomes complicated, and Theorem is not applicable directly.

2.2 Norm Variability

There is a specific property of under-determined systems of equations. Solution of linear constraint equations

$$P'(x_k)z = P(x_k)$$

in (5) (and (7)) is non-unique, in contrast to standard ($n = m$) system of equations. Thus choice of norm in pre-image space affects method's *trajectory* $\{x_k\}$. It appears that proper norm leads to very interesting results, e.g. solution sparsity. On the other hand, problem's constants L, μ do also depend on the norms and rarely could be calculated explicitly or estimated. Discussion on topic worth another article.

2.3 Connection with Non-convex Optimization Problems

Scalar problem (1) with $P : \mathbb{R}^n \to \mathbb{R}$ is not an optimization problem, but can be expressed as *non-differentiable* optimization problem $|P(x)| \to \min$. It is non-convex problem in most cases. Newton direction for Euclidean setup coincides with gradient or anti-gradient of $|P(\cdot)|$, and the method resembles a variant of gradient descent (and ascend) method for the function, differentiable everywhere except solution set $P(x) = 0$. Step-size (8) results in *quadratic* convergence rate for the problem under assumptions (2), (4). If operation ball B_ρ is the whole space, i.e. $\rho = \infty$, then the convergence is global.

3 Conclusions

We study convergence radius (range of $P(x_0)$) for Newton method applicability. For specifically chosen step-length strategy (8) we extended convergence radius. The strategy preserve quadratic convergence rate of the method. It is suitable both for standard and under-determined systems of equations (or Banach equation). To employ this strategy a-priori knowledge of problem constants is needed, and we proposed adaptive variant of the algorithm.

Acknowledgments

This work was supported by Russian Science Foundation, Project 16-11-10015.

References

[Ben-Israel, 1966] Ben-Israel, A. (1966). A Newton-Raphson method for the solution of systems of equations. *Journal of Mathematical Analysis and Applications, 15*, 243–252.

[Deuflhard, 2004] Deuflhard P. (2004). *Newton Methods for Nonlinear Problems. Affine Invariance and Adaptive Algorithms*. Berlin: Springer.

[Galántai, 2000] Galántai A. (2000). The theory of Newtons method. *Journal of Computational and Applied Mathematics, 124*, 25–44.

[Hager, 1993] Hager, W. W. (1993). Analysis and Implementation of a Dual Algorithm for Constrained Optimization. *Journal of Optimization Theory and Applications, 79(3)*, 427–462.

[Kantorovich et al, 1982] Kantorovich, L.V. & Akilov, G.P. (1982). *Functional Analysis*. 2nd ed. Oxford: Pergamon Press.

[Mysovskikh, 1949] Mysovskikh, I. (1949). On convergence of Newton's method. *Trudy Mat. Inst. Steklov, 28*, 145–147. (in Russian).

[Nesterov, 2007] Nesterov, Yu. (2007). Modified Gauss-Newton scheme with worst case guarantees for global performance. *Optimization Methods and Software 22(3)*, 469–483.

[Ortega & Rheinboldt, 2000] Ortega, J. M. & Rheinboldt, W. C. (2000). Iterative Solution of Nonlinear Equations in Several Variables. San Diego, California: Academic Press.

[Polyak, 1964] Polyak, B. T. (1964). Gradient methods for solving equations and inequalities. *USSR Computational Mathematics and Mathematical Phys. 4(6)*, 17–32.

[Polyak & Tremba] Polyak B. T. & Tremba A. A. (2018). Solving underdetermined nonlinear equations by Newton-like method Submitted to *Computational Optimization and Applications journal*.

[Pshenichnyi, 1970] Pshenichnyi, B. N. (1970). Newton's Method for the Solution of Systems of Equalities and Inequalities. *Mathematical Notes of the Academy of Sciences of the USSR, 8*, 827–830.

[Robinson, 1972] Robinson, S. M. (1972). Extension of Newton's Method to Nonlinear Functions with Values in a Cone. *Numerische Mathematik, 19*, 341–347.

[Yamomoto, 2000] Yamamoto, T. (2000). Historical developments in convergence analysis for Newton's and Newton-like methods. *Journal of Computational and Applied Mathematics, 124*, 1–23.