Optimal step length for the maximal decrease of a self-concordant function by the Newton method

Anastasia Ivanova∗ Roland Hildebrand †

February 15, 2022

Abstract
In this paper we consider the problem of finding the optimal step length for the Newton method on the class of self-concordant functions, with the decrease in function value as criterion. We formulate this problem as an optimal control problem and use optimal control theory to solve it.

1 Introduction
This paper is devoted to the problem of finding the optimal step length of Newton’s method on the class of self-concordant functions, motivated by the appearance of this class in barrier methods for conic programming. Self-concordant functions were introduced by Yu. Nesterov and A. Nemirovsky [6] when studying the behavior of Newton’s method, as follows.

Definition 1.1. A convex $C^3$ function $f : D \to \mathbb{R}$ on a convex domain $D$ is called self-concordant if it satisfies the inequality

$$|f'''(x)[h,h,h]| \leq 2(f''(x)[h,h])^{3/2}$$

for all $x \in D$ and all tangent vectors $h$.

It is called strongly self-concordant if in addition $\lim_{x \to \partial D} f(x) = +\infty$.

Step lengths for the damped Newton method were also considered in [1, 8, 5].

In this paper we find the optimal step length of Newton’s method with respect to the decrease of the function value. This criterion was considered in [6, Theorem 2.2.1], where the decrease has been lower bounded by an explicit function of the step length $\gamma_k$ and the Newton decrement $\rho_k$. The same bound has been derived in [3] in a more general context. In the latter paper it is shown that the step length $\gamma_k = \frac{1}{1+\rho_k}$ maximizes this lower bound. The same expression for the step length is also proposed in [6, Theorem 2.2.3] for larger values of the decrement. In the present paper we show by employing optimal control theory that this step length is optimal.

Optimal control theory has already been used in [4] to find an optimal step-length $\gamma^*$ which minimizes the worst-case Newton decrement in the next iteration.

The remainder of the paper is structured as follows. In Section 2 we describe the problem statement. In Section 3 we rephrase the problem as an optimal control problem and solve it analytically.

2 Problem statement
We consider Newton’s method with a damped step

$$x_{k+1} = x_k - \gamma_k (F''(x_k))^{-1} F'(x_k).$$ (2)
where \( \gamma_k \in (0, 1] \) is the step-size. Since Newton’s method is affinely invariant, i.e., a sequence of iterations on a given function transforms into itself under an affine transformation of coordinates, it is natural to study the behavior of the method on a class of functions that is also affinely invariant. This leads to the self-concordant functions which naturally arise as an affinely invariant analogue of functions with a Lipschitz continuous Hessian, and hence is well suited for an analysis of the behaviour of Newton’s method.

Namely \( \mathbb{E} \), the state at iteration \( k \) is described by a single scalar, the Newton decrement

\[
\rho_k = \| F'(x_k) \|_{F''(x_k)} = \sqrt{(F'(x_k))\top (F''(x_k))^{-1} F'(x_k)}.
\]

We consider the problem of finding a step length \( \gamma_k \) which maximizes the decrease \( F(x_{k+1}) - F(x_k) \) of the function value in the worst case realization of the function \( F(\cdot) \). So, we firstly need for given step length and given decrement to find the worst realization of the function giving the minimal decrease, and then to maximize this progress over the value of the step length, yielding the optimal step length as a function of the decrement. This leads to the following optimization problem:

\[
\max_{\gamma_k} \min_{F \in \mathcal{S}} \left( F(x_k) - F(x_{k+1}) \right),
\]

where \( \gamma_k \) is the step length, \( x_{k+1} \) is given by \( 2 \), the decrement \( \| F'(x_k) \|_{F''(x_k)} \) is fixed to some value \( \rho_k \), and \( \mathcal{S} \) is the class of functions satisfying \( 1 \).

3 Solution using optimal control theory

In this section we describe the solution of problem \( 4 \).

We consider a single iteration of the Newton method. Let the end point \( x_{k+1} \) be given by \( 2 \) and consider the line segment between \( x_k \) and \( x_{k+1} \). We study the evolution of the values of the function and its derivatives along this segment. This distinguishes our approach from the approach in \( 2 \), where \( n \) iterations and only the values of the function and its derivatives at the points \( x_1, \ldots, x_n \), i.e., a finite dimensional object, are considered. In contrast to this we consider an infinite dimensional object. The suitable apparatus to solve this problem is optimal control theory. We start with formulating our problem as an optimal control one.

3.1 Optimal control problem statement

For simplicity we denote \( \gamma := \gamma_k \) and \( \rho := \rho_k \). Parameterize the line segment between the current iterate \( x_k \) and the next one \( x_{k+1} \) affinely by a variable \( s \in [0, T] \), such that

\[
x(0) = x_k, \quad x(s) = x_k + sh, \quad x(T) = x_{k+1} = x_k + Th,
\]

where \( h \) is a direction vector along the segment between \( x_k \) and \( x_{k+1} \). Later we shall impose a normalization condition on the vector \( h \), thereby determining the value of \( T \).

Let us investigate the evolution of the gradient and Hessian of \( F \) along the segment. Since our goal is to find a realization of the function \( F(\cdot) \), which minimizes the decrease of the function value, we represent the cost function as follows:

\[
F(x_k) - F(x_{k+1}) = - \int_0^T \langle F'(x_k + sh), h \rangle \, ds.
\]

We get the maximization problem

\[
\int_0^T \langle F'(x_k + sh), h \rangle \, ds \to \max.
\]

For simplicity we introduce the function \( g(s) = \langle F'(x_k + sh), h \rangle \), then

\[
\xi(s) := \frac{dg(s)}{ds} = \langle F''(x_k + sh)h, h \rangle.
\]
From (1) we get
\[ \left| \frac{d\xi(s)}{ds} \right| = |F'''(x_k + sh)[h, h, h]| \leq 2(F'''(x_k + sh)[h, h])^{3/2} = 2\xi(s)^{3/2}, \]
or equivalently
\[ \frac{d\xi(s)}{ds} = 2u\xi(s)^{3/2}, \quad (6) \]
where \( u \in [-1, 1] \). Thus, \( u \) can be interpreted as a control, \( U = [-1, 1] \) as the set of admissible controls, and (6) as a controlled system, where \( \xi(s) \) is a positive scalar function. Introducing the function \( w(s) := \sqrt{\xi(s)} \), we get from (6) that
\[ \frac{dw^2(s)}{ds} = 2uw(s)^3 \Rightarrow \frac{dw(s)}{ds} = uw(s)^2. \]

To impose a normalization condition on \( h \), note that
\[ ||h||_{F''(x_k)} := \sqrt{F''(x_k)[h, h]} = w(0), \]
since \( x_k \) corresponds to the value \( s = 0 \). Moreover, from (2) and (5) we get
\[ x_{k+1} - x_k = Th = -\gamma(F''(x_k))^{-1}F'(x_k). \quad (7) \]
Then multiplying by \( F'(x_k) \) and using (3), we get
\[ T\langle F'(x_k), h \rangle = -\gamma\rho^2. \quad (8) \]
Moreover, multiplying (7) by \( F''(x_k) \) and \( h \), we obtain
\[ TF''(x_k)[h, h] = -\gamma\langle F'(x_k), h \rangle. \quad (9) \]
Substituting this in (8), we get
\[ T^2F''(x_k)[h, h] = \gamma^2\rho^2. \]
Choosing the normalization for \( h \) such that \( ||h||_{F''(x_k)} = 1 \), we obtain that \( T = \gamma\rho \). Substituting this into (9), we get that at \( s = 0 \)
\[ w(0) = 1, \quad g(0) = -\rho. \]

Finally, we get the following optimal control problem
\[ \int_0^{\gamma\rho} g(s)ds \rightarrow \max, \]
\[ \frac{dg}{ds} = w^2, \quad \frac{dw}{ds} = uw^2, \]
\[ w(0) = 1, \quad g(0) = -\rho \]
with control \( u \in [-1, 1] \). Introducing a new variable \( t \), such that
\[ t = -||x_{k+1} - x||_{F''(x_k)} = -\sqrt{\xi(s)[h, h]} \cdot (T - s)^2 = -w \cdot (T - s), \]
we get
\[ \frac{dt}{ds} = -w^2u \cdot (T - s) + w = w \cdot (ut + 1), \]
\[ \frac{dg}{dt} = \frac{dg}{ds} \cdot \frac{ds}{dt} = w \cdot 1 + ut, \]
\[ \frac{dw}{dt} = \frac{dw}{ds} \cdot \frac{ds}{dt} = uw \cdot 1 + ut. \]
and \( t \in [−ργ, 0] \). Denoting \( z = \frac{4}{w} \), we obtain
\[
\frac{dz}{dt} = \frac{dg}{w(1 + ut)} = \frac{w}{1 + ut} \cdot \frac{w - \frac{2}{w} \cdot w^2}{1 + ut} = \frac{1 - zu}{1 + ut},
\]
\[
g ds = \frac{g}{w(1 + ut)} dt = \frac{z}{1 + ut} dt.
\]

So, we can rewrite the optimal control problem as follows
\[
\int_{−γρ}^{0} \frac{z}{1 + ut} dt \rightarrow \max, \quad (10)
\]
\[
\dot{z} = \frac{1 - uz}{1 + ut},
\]
\[
z(−γρ) = −\rho,
\]
where \( u \in U = [−1, 1] \) and \( z \in \mathbb{R} \).

### 3.2 Solution of the problem

In this section we solve problem (10).

Firstly, according to Pontryagin’s maximum principle [7], we get the following Hamiltonian for the optimal control problem (10)
\[
\mathcal{H} = \frac{z}{1 + ut} + \psi \frac{1 - uz}{1 + ut},
\]
where \( \psi \in \mathbb{R} \) is the adjoint variable to \( z \). The dynamics of the adjoint variable is given by
\[
\dot{\psi} = -\frac{\partial \mathcal{H}}{\partial z} = \frac{uw - 1}{1 + ut}.
\]

If the control is bang-bang, i.e., \( u \) is piece-wise constant with values in \( \{-1, 1\} \), then the dynamics of the primal and adjoint variables can be integrated explicitly. For the primal variable \( z \) we get
\[
-\log |1 - uz| + C' = \log |1 + ut| \quad \Rightarrow \quad z(t) = \frac{C + t}{1 + ut},
\]
where \( C = (1 \pm e^{C'})u \) is some constant. For the adjoint variable \( \psi \) we get the solutions
\[
u \log |u\psi - 1| + C'' = \log |1 + ut| \quad \Rightarrow \quad \psi(t) = C_1(tu + 1) + \frac{1}{u},
\]
where \( C_1 = \pm e^{-C''}u \) is some constant. The transversality condition at the end-point gives \( \psi(0) = 0 \), hence \( C_1 = -\frac{1}{u} \) and
\[
\psi(t) = -t
\]
for all \( t \) from the last switching point to \( t = 0 \).

Our next step is to determine the optimal control \( u \). Maximizing the Hamiltonian over the variable \( u \) we get
\[
u = \begin{cases} 
1, & \text{if } \psi z + tz + t\psi < 0, \\
-1, & \text{if } \psi z + tz + t\psi > 0.
\end{cases}
\]

For \( t \) sufficiently close to 0 we get
\[
\psi z + tz + t\psi = -t \cdot \frac{C + t}{1 + ut} + \left( \frac{C + t}{1 + ut} - t \right) \cdot t = -t^2 < 0,
\]
and the control \( u = 1 \) is optimal. But the expression \(-t^2\) remains negative for all negative \( t \) up to the starting point \( t = γρ \). Hence the control \( u = 1 \) and above expressions for \( z(t), \psi(t) \) are valid over the whole interval.
Using the boundary conditions for $z(t)$, we obtain that

$$z(-\gamma \rho) = \frac{C - \gamma \rho}{1 - \gamma \rho} = -\rho \implies C = -\rho + \gamma \rho^2 + \gamma \rho.$$ 

Therefore

$$z(t) = \frac{-\rho + \gamma \rho^2 + \gamma \rho + t}{1 + t}.$$ 

Substituting this value into the objective function, we obtain

$$- \int_{-\gamma \rho}^{0} \frac{z(t)}{1 + t} \, dt = - \int_{-\gamma \rho}^{0} \frac{(-\rho + \gamma \rho^2 + \gamma \rho - 1) + (1 + t)}{(1 + t)^2} \, dt$$

$$= (-\rho + \gamma \rho^2 + \gamma \rho - 1) \left(1 - \frac{1}{1 - \gamma \rho}\right) + \log(1 - \gamma \rho)$$

$$= (1 + \rho)\gamma \rho + \log(1 - \gamma \rho) =: f(\gamma).$$

To find the optimal step length we need to maximize the function $f$ over $\gamma$. The first order optimality condition gives

$$\frac{\partial f}{\partial \gamma} = \frac{\rho^2 (-\gamma \rho - \gamma + 1)}{1 - \rho \gamma} = 0 \implies \gamma^* = \frac{1}{1 + \rho}.$$ 

Since $f''(\gamma) = -\frac{\rho^2}{(1 - \gamma \rho)^2} < 0$, we get that $\gamma^*$ is a maximum.

Thus we can solve this problem analytically. The same result was proposed in [6], and in [3] it is shown that this step-size maximizes a lower bound on the decrease of the function value. Here we have proved that this step length is actually optimal for this criterion.

References

[1] O. P. Burdakov. Some globally convergent modifications of Newton’s method for solving systems of nonlinear equations. Doklady Akademii Nauk, 254(3):521–523, 1980.

[2] E. De Klerk, F. Glineur, and A. B. Taylor. Worst-case convergence analysis of inexact gradient and Newton methods through semidefinite programming performance estimation. SIAM Journal on Optimization, 30(3):2053–2082, 2020.

[3] W. Gao and D. Goldfarb. Quasi-Newton methods: superlinear convergence without line searches for self-concordant functions. Optimization Methods and Software, 34(1):194–217, 2019.

[4] R. Hildebrand. Optimal step length for the Newton method: Case of self-concordant functions. Math. Methods. Oper. Res., 94:253–279, 2021.

[5] Y. Nesterov. Lectures on Convex Optimization, volume 137 of Springer Optimization and its Applications. Springer, 2018.

[6] Y. Nesterov and A. Nemirovskii. Interior-point polynomial algorithms in convex programming, volume 13. SIAM, 1994.

[7] L. Pontryagin, V. Boltyanskii, R. Gamkrelidze, and E. Mischchenko. The Mathematical Theory of Optimal Processes. Wiley, New York, London, 1962.

[8] D. Ralph. Global convergence of damped Newton’s method for nonsmooth equations via the path search. Mathematics of Operations Research, 19(2):352–389, 1994.