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

Anastasia Ivanova, Roland Hildebrand

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

Anastasia Ivanova* Roland Hildebrand †

October 19, 2023

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

In this paper we consider Newton's method with a damped step, producing iterations according to

$$x_{k+1} = x_k - \gamma_k (F''(x_k))^{-1} F'(x_k), \tag{1}$$

where $\gamma_k \in (0, 1]$ is the step-size and $\gamma_k = 1$ corresponds to a full step.

Newton's method is affinely invariant in the following sense. Let $A : x \mapsto y$ be an affine coordinate transformation, and set $y_0 = A(x_0)$. Fix a sequence of step lengths γ_k, $k \in \mathbb{N}$. Then produce a sequence of iterates x_k according to (1) with initial point x_0, and a sequence y_k with the same step lengths, but with initial point y_0 and computed in the coordinate system y. Then $y_k = A(x_k)$ for all $k \in \mathbb{N}$.

Thus it is natural to study the behavior of the method on a class of functions that is also affinely invariant, i.e., such that membership in the class does not change under affine coordinate transformations. 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. Self-concordant functions were introduced by Yu. Nesterov and A. Nemirovsky [9] 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

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

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

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

The authors in [9] describe the state at iteration k by a single scalar, the *Newton decrement*

$$\rho_k = ||F'(x_k)||_{F''(x_k)} := \sqrt{(F'(x_k))^\top (F''(x_k))^{-1} F'(x_k)}. \tag{3}$$

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, in particular, when solving linear programs, second-order cone programs, and semi-definite programs. Step lengths for the damped Newton method were also considered in [1, 11, 8].

*Univ. Grenoble Alpes, LJK, 38000 Grenoble, France; HSE University, Moscow, Russian Federation (anastasia.ivanova@univ-grenoble-alpes.fr).

†Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France; (roland.hildebrand@univ-grenoble-alpes.fr).

Step lengths for the damped Newton method were also considered in [1, 11, 8]. The behaviour of the Newton decrement and the function value under specific step sizes has been studied in [12, Section 2.2]. In [2, Corollary 6.1] the decrease of the distance from the optimum and of the norm of the gradient, both in the local metric of the initial point, have been bounded for self-concordant functions if the initial point is close enough to the minimum. The same bound has been obtained in [2, Corollary 6.3] for the local metric of the minimum. A bound on the decrease in function value can be derived from [2, Theorem 5.3], however, it depends on the difference between the current and optimal function values. In that paper an inexact Newton step can be taken, and the bound depends on the error. The methods used in [2] rely on semi-definite programming (see also [3]) and are completely different from those employed here.

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 [9, Theorem 2.2.1], where the decrease has been lower bounded by an explicit function of the step length γ_k and the Newton decrement ρ_k. The same bound has been derived in [4] 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 [9, Theorem 2.2.3] for larger values of the decrement. While in [4], and implicitly in [9], the step length has been obtained as the maximizer of a bound, in the present paper we show by employing optimal control theory that this step length is actually optimizing the function value itself. It turns out, however, that no further improvement over the results in the mentioned papers occurs, despite the use of the exact criterion.

The idea to use optimal control for the worst-case analysis of first order methods has been developed by Laurent Lessard and co-authors in [7], where they use this technique to derive numerical upper bounds on convergence rates for the Gradient method, the Heavy-ball method, Nesterov's accelerated method, and related variants by solving small, simple semidefinite programming problems. The same technique was also considered in [13, 6].

Optimal control theory has already been used in [5] to find an optimal step-length γ^* for the Newton method on self-concordant functions. However, a different strategy has been adopted there. Instead of the worst-case function value, as in the present work, the worst-case Newton decrement in the next iteration is minimized. This criterion is more in line with the philosophy of interior-point methods as presented in [9], but it has the drawback that if the decrement is larger than 1, no progress can be guaranteed at all. Also, the optimal value of the step length turns out to be not expressible in closed form in general. The criterion used in the present paper, on the contrary, can be strictly improved at each step, no matter how far we are from the optimum at the current iteration, and the value of the optimal step length is a simple analytic function of the data available at the current iteration.

In this paper consider the problem of finding a step length γ_k which maximizes the decrease $F(x_k) - F(x_{k+1})$ 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:

$$\max_{\gamma_k} \min_{F \in \mathcal{S}} \left(F(x_k) - F(x_{k+1}) \right), \tag{4}$$

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

2 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 (1) 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.

2.1 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

$$x(0) = x_k, \quad x(s) = x_k + sh, \quad x(T) = x_{k+1} = x_k + Th, \tag{5}$$

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 (2) we get

$$\left|\frac{d\xi(s)}{ds}\right| = |F'''(x_k + sh)[h, h, h]| \le 2(F''(x_k + sh)[h, h])^{3/2} = 2\xi(s)^{3/2},$$

or equivalently

$$\frac{d\xi(s)}{ds} = 2u\xi(s)^{3/2}, \tag{6}$$

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

$$\frac{dw^2(s)}{ds} = 2uw(s)^3 \quad \Rightarrow \quad \frac{dw(s)}{ds} = uw(s)^2.$$

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 (1) and (5) we get

$$x_{k+1} - x_k = Th = -\gamma(F''(x_k))^{-1}F'(x_k). \tag{7}$$

Then taking the inner product with $F'(x_k)$ and using (3), we get

$$T\langle F'(x_k), h\rangle = -\gamma\rho^2. \tag{8}$$

Moreover, taking the inner product of (7) with $F''(x_k)$ and h , we obtain

$$TF''(x_k)[h, h] = -\gamma\langle F'(x_k), h\rangle. \tag{9}$$

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

$$\begin{aligned} &\int_0^{\gamma\rho} g(s)ds \ \to \ \max, \\ &\frac{dg}{ds} = w^2, \quad \frac{dw}{ds} = uw^2, \\ &w(0) = 1, \quad g(0) = -\rho \end{aligned}$$

with control $u \in [-1, 1]$. Introducing a new variable t, such that

$$t = -||x_{k+1} - x||_{F''(x)} = -\sqrt{\xi(s)[h, h] \cdot (T - s)^2} = -w \cdot (T - s),$$

we get

$$\begin{aligned} \frac{dt}{ds} &= -w^2 u \cdot (T - s) + w = w \cdot (ut + 1), \\ \frac{dg}{dt} &= \frac{dg}{ds} \cdot \frac{ds}{dt} = \frac{w}{1 + ut}, \\ \frac{dw}{dt} &= \frac{dw}{ds} \cdot \frac{ds}{dt} = \frac{wu}{1 + ut}, \end{aligned}$$

and $t \in [-\rho\gamma, 0]$. Denoting $z = \frac{g}{w}$, we obtain

$$\begin{aligned} \frac{dz}{dt} &= \frac{d(g/w)}{dt} = \frac{\frac{w}{1+ut} \cdot w - \frac{g}{w} \cdot \frac{w^2 u}{1+ut}}{w^2} = \frac{1 - zu}{1 + ut}, \\ g\, ds &= \frac{g}{w(1 + ut)} dt = \frac{z}{1 + ut} dt. \end{aligned}$$

So, we can rewrite the optimal control problem as follows

$$\begin{aligned} &\int_{-\gamma\rho}^0 \frac{z}{1 + ut} dt \ \to \ \max, \\ &\dot{z} = \frac{1 - uz}{1 + ut}, \\ &z(-\gamma\rho) = -\rho, \end{aligned} \tag{10}$$

where $u \in U = [-1, 1]$ and $z \in \mathbb{R}$.

2.2 Solution of the problem

In this section we solve problem (10).

Firstly, according to Pontryagin's maximum principle [10], 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

$$\dot{\psi} = -\frac{\partial \mathcal{H}}{\partial z} = \frac{u\psi - 1}{1 + ut}.$$

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

$$-u \log|1 - uz| + C' = u \log|1 + ut| \quad \Rightarrow \quad z(t) = \frac{C + t}{1 + ut},$$

where $C = (1 \pm e^{C'u})u$ is some constant. For the adjoint variable ψ we get the solutions

$$u \log|u\psi - 1| + C'' = u \log|1 + ut| \quad \Rightarrow \quad \psi(t) = C_1(tu + 1) + \frac{1}{u},$$

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 $t = -\gamma\rho$. 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

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

Therefore

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

Substituting this value into the objective function, we obtain

$$\begin{aligned} -\int_{-\gamma\rho}^{0} \frac{z(t)}{1 + t}\, dt &= -\int_{-\gamma\rho}^{0} \frac{(-\rho + \gamma\rho^2 + \gamma\rho - 1) + (1 + t)}{(1 + t)^2}\, dt \\ &= (-\rho + \gamma\rho^2 + \gamma\rho - 1)\left(1 - \frac{1}{1 - \gamma\rho}\right) + \log(1 - \gamma\rho) \\ &= (1 + \rho)\gamma\rho + \log(1 - \gamma\rho) =: f(\gamma). \end{aligned}$$

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

$$\frac{\partial f}{\partial \gamma} = \frac{\rho^2(-\gamma\rho - \gamma + 1)}{1 - \rho\gamma} = 0 \quad \Rightarrow \quad \gamma^\star = \frac{1}{1 + \rho}.$$

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

Thus we can solve this problem analytically. The same result was proposed in [9], and in [4] 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] Y. Drori and M. Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Math. Program.*, 145(1–2):451–482, 2014.

[4] 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.

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

[6] B. Hu and L. Lessard. Control interpretations for first-order optimization methods. In *2017 American Control Conference (ACC)*, pages 3114–3119. IEEE, 2017.

[7] L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

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

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

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

[11] 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.

[12] J. Renegar. *A Mathematical View of Interior-point Methods in Convex Optimization.* MPS-SIAM Series on Optimization. SIAM, MPS, 2001.

[13] A. Taylor, B. Van Scoy, and L. Lessard. Lyapunov functions for first-order methods: Tight automated convergence guarantees. In *International Conference on Machine Learning*, pages 4897–4906. PMLR, 2018.