Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Backtracking New Q-Newton's Method

Tuyen Trung Truong

tuyentt@math.uio.no

University of Oslo

Research Article

Keywords: Backtracking line search, Basin of attraction, Convergence guarantee, Dynamical systems, Lojasiewicz gradient inequality, Optimization, Newton's method, Random processes, Rate of convergence, Root finding, Saddle points

Posted Date: March 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9111824/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

BACKTRACKING NEW Q-NEWTON'S METHOD

TUYEN TRUNG TRUONG

To Professor Eric Bedford, who guided me to Dynamical Systems and computer's assistance in research.

Abstract. In this paper, we introduce a new Newton-type method - named Backtracking New Q-Newton's method (BNQN) - which retains the same quick local rate of convergence while having much stronger global convergence guarantees. It is a combination of a previously designed algorithm (New Q-Newton's method) and Armijo's Backtracking line search. The main contribution of this paper is as follows: We prove that this new algorithm, applied to a general C^3 cost function f, has the following good properties: For a sequence $\{x_n\}$ constructed by this algorithm from a random initial point x_0:

- Finding critical points: Any cluster point of the sequence is a critical point of f.

- Fast convergence rate: Near any non-degenerate local minimum, it has local quadratic convergence rate.

- Avoidance of saddle points: If x_0 is outside an exceptional set of Lebesgue measure 0, then $\{x_n\}$ cannot converge to a saddle points.

BNQN is the first variant of Newton's method in the literature satisfying all these 3 properties.

We also prove some further results, and propose some other variants. We illustrate the main results with both theoretical and experimental examples.

Information

Author: Tuyen Trung Truong. Affiliation: Mathematics Department, University of Oslo, Norway. Email: tuyentt@math.uio.no. ORCID: 0000-0001-9103-0923

Statements and Declarations

Competing interests: none. Data availability: not applicable.

1. Introduction

Solving efficiently systems of equations is an important origin of mathematics and is still a very active research, in particularly in fields such as Complex Variables, Algebraic Geometry, Computer Algebra, Dynamical Systems, and Optimization. In particular, root finding methods

Date: March 13, 2026.

Key words and phrases. Backtracking line search, Basin of attraction, Convergence guarantee, Dynamical systems, Lojasiewicz gradient inequality, Optimization, Newton's method, Random processes, Rate of convergence, Root finding, Saddle points.

2020 Mathematics Subject Classification. 37N40, 49M15, 49M37, 65Exx, 65Hxx, 65K05, 90C26.

for polynomials and holomorphic functions in one complex variables is very extensively studied in Complex Dynamics, and a good reference on Newton's method and some of its well known variants is [6]. In general, many practical questions can also be reduced to solving systems of equations. One can treat solving a system of equations $F(x) = 0$ as an optimization problem by looking at global minima of the associated scalar function $f(x) = ||F(x)||^2/2$. Vice versa, root finding methods can help with finding critical points of a function f, which satisfies a system of equation $\nabla f(x) = 0$. Thus, solving systems of equations and Optimization go hand in hand, and usually a method for root finding has a corresponding version in Optimization. This sets the narrative of this paper, where we seek to define new methods in Optimization and root finding with both strong theoretical guarantees and good experimental performances, and having connections with other subjects.

When finding minima of a function $f : \mathbb{R} \to \mathbb{R}$, the first step could be to look for its critical points $f(x) = 0$. Applying Newton's method to the equation $f'(x) = 0$, one obtains the scheme $x_{n+1} = x_n - (f"(x_n))^{-1} f'(x_n)$. In higher dimensions, one can proceed as follows. Let $f : \mathbb{R}^m \to \mathbb{R}$ be a C^2 function, with gradient $\nabla f(x)$ and Hessian $\nabla^2 f(x)$. Newton's method for optimization $x_{n+1} = x_n - (\nabla^2 f(x_n))^{-1} \nabla f(x_n)$ (if the Hessian is **invertible**) is a very popular iterative method for solving systems of equations and for optimization in general. One attractive feature of this method is that if it **converges** then it usually converges very fast, with the rate of convergence being quadratic, which is generally faster than that of Gradient descent (GD) methods. We recall that if $\{x_n\} \subset \mathbb{R}^m$ converges to x_∞, and so that $\|x_{n+1} - x_\infty\| = O(\|x_n - x_\infty\|^\epsilon)$, then ϵ is the rate of convergence of the given sequence. If $\epsilon = 1$ then we have linear rate of convergence, while if $\epsilon = 2$ then we have quadratic rate of convergence. The study of Newton's method, and methods to find solutions to systems of equations in general, has been one major force for many advancements (besides the realistic applications) in pure mathematics and numerical methods. Being at the interface between both theoretical and applied worlds, the question of finding methods with good characters in both theory and practice is both extremely interesting and challenging. This paper presents a promising candidate for such methods, which is a modification of Newton's method inherent the good characteristics of Newton's method while simultaneously resolving some major disadvantages of Newton's method.

However, there is no guarantee that Newton's method will converge, and it is problematic near points where the Hessian is not invertible. Moreover, it cannot avoid non-degenerate saddle points. Recall that a non-degenerate **saddle point** is a point x^* which is a non-degenerate critical point of f (that is $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is invertible) so that the Hessian has at least one **negative eigenvalue**. (Note that this definition allows also local maxima.)

Convention: From now on, for ease of presentation, by saddle points we mean non-degenerate saddle points, unless otherwise specified. This case already covers many interesting applications, for example to Morse functions (which is expected generically).

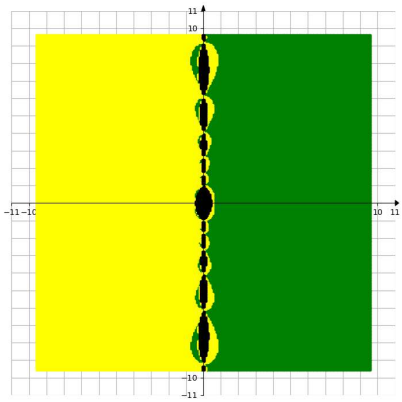


FIGURE 1. Basins of attraction for finding roots of a polynomial of degree 2, more precisely $F(z) = z^2 - 1$, for Newton's method for Optimization (applied to find minima of the function $f(x, y) = |g(x + iy)|^2/2$). Initial points of the same color (green or yellow) produce sequences converging to the same root of the polynomial, while initial points of black color are those which produce sequences converging to the point 0 (the unique critical point of $F(z)$, and a saddle point of $f(x, y)$).

Remark: If the cost function f is convex, then it has no saddle points, since any critical point is also a global minimum. Therefore, the question of avoiding saddle points is only meaningful for non-convex cost functions.

Saddle points are dominant and hence problematic in higher dimensions, see [7]. Figure 1 illustrates a simple case of applying Newton's method to minimise the function $f(x, y) = |F(x + iy)|^2/2$, where $F(z) = z^2 - 1$. The function f has 3 critical points: 2 global minima (the roots $1, -1$ of F), and a saddle point (the critical point 0 of F).

Another serious issue with Newton's method is the existence of attracting cycles of non-critical points, even for simple cost functions. For example, consider the simple polynomial of degree 4, $f(x) = (x^2 - 1)(x^2 + A)$. For some special values of A (such as $A = (29 - \sqrt{720})/11$), Newton's method applied to find roots of f will have an attracting 2-cycle, none of them is a critical point of $\|f\|^2$. For polynomials $P(z)$ in 1 complex variable z, it is recently shown in [28] that random Damped Newton's method $x_{n+1} = x_n - \delta_n P(x_n)/P'(x_n)$, where δ_n is a random complex number and where the initial point x_0 is randomly chosen, will converge to a root of $P(z)$. It is, however, unknown how good random Damped Newton's method is for non-polynomial cost functions and in higher dimensions, and the extensive experiments performed by the author and coauthors in [33] illustrates that it does not fare well in the general setting.

Newton's method, and other iterative methods in solving systems of equations and optimization, can be studied in the general setting of discrete Dynamical Systems. There are many modifications of Newton's method. A short review will be presented in Section 2. As far as we know, none of the known methods answer in the affirmative the following question:

Question: Is there a Newton-type method which simultaneously have the following four good properties:

1. Near non-degenerate local minima, it has quadratic rate of convergence.
2. All of the cluster points of sequences constructed by its are critical points of the cost function.
3. It has global convergence guarantee (to local minima) for large classes of cost functions.
4. It is easy to implement and use on computers, and behave stably with respect to its parameters.

Importantly, while commonly in the Optimization literature, most results require strong conditions on the cost function (like convexity or the gradient being Lipschitz continuous), in this paper we look for methods which are applicable to more general cost functions and hence require less hurdles in both theory and practice. For example, the well known Riemann xi function does not satisfy those mentioned assumptions, but satisfies the Lojasiewicz inequality (to be recalled later) under which our new methods have strong global convergence guarantees. Likewise, we are also not aware of any current Newton-type methods in equation solving which have global convergence assurance for functions like Riemann xi function.

This paper shows that a modification (Backtracking New Q-Newton's method, or BNQN) of a very new version of Newton's method, so-called New Q-Newton's method, introduced by the author and collaborators in [33], answers affirmatively the above Question. The method is also straight forward to be implemented from the pseudo-code, and very flexible with respect to its parameters. Simplified versions of its also work well. For the sake of comprehension, here we describe the essence of these new algorithms, and present the details of the algorithms in the next section.

We start with New Q-Newton's method. The main idea is that if we fix $m+1$ distinct numbers $\delta_0, \dots, \delta_m$, and fix $\tau > 0$, then for every x so that $\nabla f(x) \neq 0$, there is one j so that $\nabla^2 f(x) + \delta_j \|\nabla f(x)\|^\tau Id$ is invertible. We choose such a j, and change the negative eigenvalues of the matrix to their negative, before taking its inverse in the update rule. If $\delta_0, \dots, \delta_m$ are chosen randomly, then [33] shows that New Q-Newton's method has local stable manifolds near saddle points. Moreover, near non-degenerate local minima, New Q-Newton's method has quadratic rate of convergence. However, it is unknown whether New Q-Newton's method has good convergence guarantees.

In this paper, we observe that the matrix in New Q-Newton's method is semi-positive and hence one can incorporate Armijo's Backtracking line search [1] into it. We call the new algorithm Backtracking New Q-Newton's method. It preserves the good properties of New Q-Newton's method, while also resolves the convergence issue. We also show the global avoidance of saddle points for both New Q-Newton's method and BNQN. Moreover, a family of such modifications can be

introduced, some of them having the flavours of quasi-Newton's methods, with strong theoretical guarantees. These more general modifications will be presented in the last section of this paper. We conclude this introduction with some main properties of Backtracking New Q-Newton's method.

We recall that a function f is Morse if all of its critical points are non-degenerate (i.e. if $\nabla f(x_0) = 0$, then $\nabla^2 f(x_0)$ is invertible). Such a function has at most countably many critical points. By transversality results, Morse functions are dense in the set of functions.

Theorem 1.1. *Let $f : \mathbb{R}^m \to \mathbb{R}$ be a C^3 function. Let $\{x_n\}$ be a sequence constructed by the Backtracking New Q-Newton's method, where $0 < \tau < 1$.*

1) (Descent property) $f(x_{n+1}) \leq f(x_n)$ for all n.

2) If x_∞ is a **cluster point** *of $\{x_n\}$, then $\nabla f(x_\infty) = 0$. That is, x_∞ is a* **critical point** *of f.*

3) If f has at most countably many critical points, then either $\lim_{n\to\infty} \|x_n\| = \infty$ or $\{x_n\}$ converges to a critical point of f. In the latter case, if the initial point x_0 is randomly chosen, then the limit point cannot be a saddle point, and if f is moreover a **Morse function** *then the limit point is a local minimum.*

4) There is a set $\mathcal{A} \subset \mathbb{R}^m$ of Lebesgue measure 0, so that if $x_0 \notin \mathcal{A}$, and x_n converges to x_∞, then x_∞ cannot be a **saddle point** *of f. Hence, if $\nabla^2 f(x_\infty)$ is invertible, then x_∞ must be a local minimum.*

5) If x_n converges to x_∞ which is a non-degenerate critical point of f, then the rate of convergence is at least **linear**.

If moreover, x_∞ is a local minimum, then the rate of convergence is quadratic.

6) If x'_∞ is a non-degenerate local minimum of f, then for initial points x'_0 close enough to x'_∞, the sequence $\{x'_n\}$ constructed by the Backtracking New Q-Newton's method will converge to x'_∞.

Note that the condition for a function to be Morse is a generic condition. Hence, for a generic function, and for a randomly chosen initial point x_0, our algorithm Backtracking New Q-Newton's method either converges to a local minimum or diverges to infinity. Another important class of functions, which included real analytic functions, has many realistic applications (such as in Deep Learning). The next main theorem describes convergent guarantees of Backtracking New Q-Newton's method for this class of functions. To this end, we first recall about Lojasiewicz gradient inequality and Lojasiewicz exponent.

Lojasiewicz gradient inequality. A function f satisfies Lojasiewicz gradient inequality at a point x^* if there is a small neighbourhood U of x^*, a constant $0 < \mu < 1$ and a constant $C > 0$ so that for all $x, y \in U$ we have

$$|f(x) - f(y)|^\mu \leq C\|\nabla f(x)\|.$$

We will need the following quantity in statements of the next results:

Definition (Lojasiewicz exponent): Assume that $f : \mathbb{R}^m \to \mathbb{R}$ has the Lojasiewicz gradient inequality near its critical points. Then at each critical point x^* of f, we define

$\mu(x^*) := \inf\{\mu :$ there is an open neighbourhood U of x^* and a constant $C > 0$ so that for all $x, y \in U$ we have $|f(x) - f(y)|^\mu \leq C\|\nabla f(x)\|\}$.

We say that the gradient ∇f satisfies Lojasiewicz gradient inequality at a point x^*, if the function $F(x, y) = \langle \nabla f(x), y\rangle : \mathbb{R}^{2m} \to \mathbb{R}$ satisfies Lojasiewicz gradient inequality. By Lojasiewicz' theorem, if f is real analytic (and hence $F(x, y)$ is also real analytic), then f and its gradient satisfy Lojasiewicz gradient inequality. Hence, the next theorem can be applied to **quickly** finding roots of systems of (real or complex) analytic equations.

Theorem 1.2. *Assume that $f : \mathbb{R}^m \to \mathbb{R}$ satisfies the Lojasiewicz gradient inequality. Let $\{x_n\}$ be a sequence constructed by Backtracking New Q-Newton's method. Assume also that $0 < \tau \leq 1$.*

1) Assume that f has at most countably many critical points, and ∇f satisfies the Lojasiewicz gradient inequality. Then either $\lim_{n\to\infty} \|x_n\| = \infty$ or $\{x_n\}$ converges to a critical point of f.

2) Assume that for all critical points x^* of f, we have $\mu(x^*) \times (1 + \tau) < 1$. For $\tau = 1$, assume moreover that ∇f also satisfies the Lojasiewicz gradient inequality. Then either $\lim_{n\to\infty} \|x_n\| = \infty$ or $\{x_n\}$ converges to a critical point of f.

The result in part 1 is new only when $\mu = 1$ (the remaining case is covered by Theorem 1.1). In part 2, we do not require that f has only countably many critical points. In part 2), it is allowed that $\sup_{\nabla f(x^*)=0} \mu(x^*) \times (1 + \tau) = 1$, provided the supremum is not attained at any critical point x^*. On the other hand, since in general $\mu(x^*) \geq 1/2$, it follows that if $\mu(x^*)(1 + \tau) < 1$ then we should have $\tau < 1$.

In [33], the case where $f : \mathbb{R}^2 \to \mathbb{R}$ is defined as $f = u^2 + v^2$, where $u =$ the real part of g and $v =$ the imaginary part of g, g is a univariate holomorphic (more generally, meromorphic) function and τ can be ≥ 1, was treated. A generalisation of this result, which is a bit complicated to state here, will be presented in Theorem 4.5.

We recall that Smale's 17th problem [27] concerns designing efficient algorithms to **find a root of** generic systems of polynomial equations $F = 0$ in complex variables. Combining Newton's method with homotopy method [26], and randomness it has been solved in [5][8]. However, there is a trade off: in this algorithm, one needs to construct a new system of polynomial equations $G = 0$ depending on all coefficients of the system F, and hence the algorithm is very expensive. When one wants to find a new root, one needs to construct a new system $G' = 0$, which is again expensive (and one is not guaranteed to avoid the root already found). As such, it is extremely difficult to implement an efficient version of this theoretical result. On the practical side, an associated software for homotopy continuation for systems of polynomial equations in complex variables has been implemented in Bertini [4]. However, homotopy continuation methods have several disadvantages. One serious difficulty with homotopy continuation methods is that the curves

to trace may have very high curvatures (this is apparent already when one applies to finding roots of a polynomial of high degree in 1 complex variable), which increase both the difficulty in tracking the paths and in avoiding failure. Moreover, the more general and useful question of finding efficient algorithms to solve systems of polynomials in real variables (i.e. over $\mathbf{R}$), or over quaternions, or finding efficient algorithms to solve systems of transcendental equations in complex variables, is still largely open. That Backtracking New Q-Newton's method has a strong convergence guarantee in general optimization problems, or to solving certain systems of real analytic functions in Theorem 1.2 and more in Section 4, as well as in finding roots of meromorphic functions in 1 complex variable (as will be discussed more in the paragraph "Organization of the paper" and in Section 5), suggests that it is a good place to start in the search for such efficient algorithms in these general settings. In ongoing work, we are exploring further the application of BNQN in solving systems of polynomial equations in higher dimensions, and it seems that phenomena observed in complex dimension 1 (e.g. similarity between basins of attraction for BNQN and Voronoi's diagrams [11] and Theorem 3.3 in [33]) also occur in higher dimensions. BNQN can also be combined with homotopy continuation, replacing the use of Newton's method in unstable situations.

Takayuki Watanabe informed us of interesting experiments involve a system of equations in 2 complex variables $z_1 + z_1z_2 = z_1^2 - z_2^2 = 0$. $(0,0)$ is a multiple root of this system. The interesting fact about this system is that $(0,0)$ is **not** an attractor for the dynamics of Newton's method, but it seems to be an attractor for BNQN. Please see Section 5.3 for illustrations. Moreover, it seems from the experiments that for a random choice of initial point x_0, then BNQN applied to solve this system of equation will converge to one of the 3 roots $(0,0)$, $(-1,-1)$ and $(1,-1)$ of the system. Using the theoretical results developed so far, we show that this is indeed the case. The proof of this result can be adapted to many other cases of interest.

Theorem 1.3. *Let $h(z_1,z_2) = (z_1 + z_1z_2, z_1^2 - z_2^2)$, and $F : \mathbf{R}^4 \to \mathbf{R}$ defined by $F(y_1,y_2,y_3,y_4) = ||h(y_1 + iy_2, y_3 + iy_4)||^2$. There is an exceptional set $\mathcal{E} \subset \mathbf{R}^4$ of Lebesgue measure 0 so that if we apply BNQN to the function F and an initial point $x_0 \in \mathbf{R}^4 \backslash \mathcal{E}$, then the constructed sequence converges, and the limit point is one of the 3 roots $(0,0)$, $(-1,-1)$ and $(1,-1)$.*

Python code: One practical usefulness of the algorithms developed in this paper is that they are very easy to implement, literally as the pseudo code (and even simplified implementations work pretty well). A python code is available here [38]. In this implementation, we do not need the precise values of the gradient and Hessian matrix of a function, but their approximations from the Python package numdiftools [36] are enough.

Organisation of the paper: The remaining of this paper is organised as follows. In Section 3, we recall about New Q-Newton's method method, and define the new algorithm Backtracking New Q-Newton's method. In Section 4, we present the proofs of Theorems 1.1 and 1.2, as well as some extensions to solving systems of equations (including some new algorithms resulting from

a combination between ideas of Backtracking New Q-Newton's method and that of Regularized Newton's method and Levenberg-Marquardt's method). This is followed by Section 5, where we present some experimental results comparing the long-term global behaviour of our methods with well known ones (such as BFGS, Levenberg-Marquard's algorithm, Regularized Newton's method and (Adaptive) Cubic Regularization of Newton's method) on some polynomial and non-polynomial systems of equations in several variables. We display a couple of experiments, concerning finding roots of some polynomial systems and of the Riemann xi function, and refer to the arXiv version of this paper arXiv:2209.05378, as well as some more recent works [11] [29], for more extensive experiments. In Section 5, we also prove Theorem 1.3. In the appendix, we describe some further extensions.

Further applications. We note that a recent paper [12] proved that indeed the picture one obtains when applying Backtracking New Q-Newton's method to find roots of a polynomial of degree 2, which looks similar to that produced by the classical Schröder's theorem [24], is indeed correct. This, and the results in [11] illustrate that Backtracking New Q-Newton's method has a close relation to Voronoi's diagrams and captures better the geometric configuration of the roots than Newton's method. These facts are rather unexpected, given that Backtracking New Q-Newton's method depends on many seemingly random factors (its parameters, as well as the use of Armijo's Backtracking line search), and can be applied to the Riemann hypothesis [29]. Several theoretical explanations of these facts are very recently given in [10], when BNQN (an iterative method) is shown to have striking similarities to Newton's method and the well known Poincaré-Bendixon theorem for differentiable real dynamical systems on the complex plane.

Remark: This paper combines the author's 3 previous ones: arXiv:2108.10249, arXiv: 2109.11395, and arXiv:2110.07403, with some slight revisions, and supersedes these papers. The proof of avoidance of saddle point for BNQN uses the Euclidean part of the paper: arXiv:2008.11091, and supersedes that part. Extensions to Riemannian manifolds and Banach spaces, where a simplified version of BNQN is proposed, can be found in [30].

Acknowledgments. The author is partially supported by Young Research Talents grant 300814 from Research Council of Norway. The author thanks Thu Hien To and Vu Dinh for helping with experiments, thanks Jonathan Hauenstein for explaining thoroughly many aspects of homotopy continuation and Bertini software, and thanks Cinzia Bisi and Stefano Luzzatto, as well as Mohammad Alkousa, for inspiring discussions on the subject. The recent event in Kongsberg, which happened when part of this work was being done, has been influential to the author.

2. A short review of some variants of Newton's method

There are so many variants of Newton's method (in both root finding and optimization) that one cannot even mention all their names in a research paper which is not for review purpose.

We refer interested readers to some comprehensive surveys such as [14][15]. This paper treats only discrete optimization/root finding algorithms, and not the flow methods like Newton's flow. Besides experiments here, see [31, 33] and additional experiments in this paper. Let us also mention that there is a close relation between Backtracking New Q-Newton's method and Newton's flows see [11].

As mentioned before, in the literature the focus is to local convergence or global convergence but under strong assumptions, which make it inapplicable to interesting cases like the Riemann xi function. In this section, we concentrate on reviewing some methods which are closest to our method and which have good global behaviour (e.g. global convergence guarantees or global avoidance of saddle points).

Among previous second order methods which can avoid saddle points, there is one in [23] which is most closely related to our method which we now provide some detailed analysis. **First**, consider the algorithms. The method in [23] roughly works as follows: they choose a number $\kappa > 0$ and change the eigenvalues λ's of the Hessian matrix $\nabla^2 f(x_n)$ to $\max\{|\lambda|, \kappa\}$, then do Armijo's Backtracking line search (please see the appendix for Armijo's Backtracking line search). Our method BNQN is different in that we replace the Hessian matrix $\nabla^2 f(x_n)$ by $A_n = \nabla^2 f(x_n) + \delta_j ||\nabla f(x_n)||^{1+\alpha} Id$ (whose eigenvalues all have absolute values bigger than an appropriate multiple of $||\nabla f(x_n)||^{1+\alpha}$), change eigenvalues λ's of A_n to $|\lambda|$, and then apply Armijo's Backtracking line search. Hence, the 2 methods have roughly the same complexity, while BNQN is closer to the original Newton's method near critical points of f. In particular, BNQN has quadratic rate of convergence near non-degenerate local minima of f, while it is easy to construct examples (for example $f(x) = \frac{\kappa_0}{2} x^2$ in 1 real variable, where $0 < \kappa_0 < \kappa$) where the method in [23] has slower rate of convergence. Indeed, the method in [23] is somehow in between Newton's method and Gradient descent. **Second**, consider the scope of applications. The theoretical results in [23] require that the gradient ∇f and Hessian matrix $\nabla^2 f$ are Lipschitz continuous, and also f has only a finite number of critical points. While these conditions allow to prove explicit estimates for the method in [23], they are not applicable for many natural classes of functions (for example, even polynomials of degree 3 do not satisfy these conditions). For BNQN method, these conditions are not needed, and the global convergence and avoidance of saddle points for BNQN can be applied for example to functions which have at most countably many critical points (which is a generic condition) and even to functions which are very transcendental (for example, to find roots of meromoprhic functions and in particular the Riemann xi function, see [33, 12, 11]). A new equivalence of and approach to the Riemann hypothesis is recently obtained in [29].

A well known variant of Newton's method is Levenberg-Marquardt algorithm [16][20], very popular for using in the least square fit problem where the cost function f is a sum of squares of real functions: $f(x) = \frac{1}{2}[f_1(x)^2 + \ldots + f_N(x)^2]$, and hence is helpful for equations solving. Let

$F(x) = (f_1(x), \dots, f_N(x))$, and $JF(x)$ the Jacobian of F. Then the update rule for Levenberg-Marquardt algorithm is:

$$z_{n+1} = z_n - [JF(z_n)^T JF(z_n) + \lambda_n Id]^{-1} JF(z_n)^T . F(z_n), \tag{1}$$

where A^T is the transpose of a matrix A, $\lambda_n > 0$ is an appropriate constant. A usual choice for λ_n is $\lambda_n = c\|F(z_n)\|^\tau$ for constants $c, \tau > 0$. (One can also interpolate λ_n between $\|F(z_n)\|^{\tau_1}$ and $\|JF(z_n)^T.F(z_n)\|^{\tau_2}$, see [3].) If in (1) one chooses $\lambda_n = 0$, then one obtains the classical Newton's method applied directly to solve the system of equations $F = 0$. Hence, Levenberg-Marquardt algorithm can be viewed as a correction of Newton's method for solving systems of equations, treating the case $JF(z_n)$ not invertible.

For a general cost function f, an analog (or extension) of Levenberg-Marquardt algorithm is the so called Regularized Newton method [35][34][25]. It replaces the $\nabla^2 f(z_n)$ in Newton's method by $\nabla^2 f(z_n) + [c_1 \max\{0, -\lambda_{min}(\nabla^2 f(z_n))\} + \lambda_n] Id$, where $c_1 > 1$ is a constant, $\lambda_{min}(A)$ is the smallest eigenvalue of a real symmetric matrix A, and $\lambda_n > 0$ is appropriately chosen. A usual choice for λ_n here is $c_2\|\nabla f(z_n)\|^\tau$ for constants $c_2, \tau > 0$.

Research on these algorithms concentrate only on local convergence (and rate of convergence) near local minima, without much addressing about the global convergence (that is, research on if cluster points of the constructed sequence are all critical points of the cost function, and if the constructed sequence actually converges) or saddle point avoidance. The design of these methods can be a reason for the fact that no result on avoidance of saddle points is available for them (for a more detailed analysis of this issue, see Section 3). In the case of Levenberg-Marquardt method, this can be explained by the fact that the matrix $JF(z_n)^T JF(z_n)$ is not the whole of the Hessian of the function $f(x) = \frac{1}{2}\|F(x)\|^2$, and hence may not be able to control near saddle points of $f(x)$. In the case of Regularized Newton's method, there are two factors which may contribute to the difficulty when dealing theoretically with this method. First, the update rule in Regularized Newton's method may not be a C^1 map near saddle points where $\lambda_{min}(\nabla^2 f)$ has multiplicity > 1. Second, the choice of the parameters in this method is not flexible enough. Also, on various examples, these two methods converge only when one adds a line search (e.g. Armijo's Backtracking line search) to them.

Indeed, the above analysis also applies to many other well known variants of Newton's method in the literature. As far as we know, currently there is no variant which has theoretical guarantees for both convergence and avoidance of saddle points, while keeping the fast rate of convergence, and has an easy and flexible implementation with good performance. For example, see here for the discussion of difficulty with implementing the well known Cubic Regularization method [37]. For a comprehensive test of several variants of Newton's method, on various test problems, the readers can see [33].

The precursor to the methods presented in this paper is New Q-Newton's method [33], which we will recall now. Let $A : \mathbb{R}^m \to \mathbb{R}^m$ be an invertible **symmetric** square matrix. In particular, it is diagonalisable. Let V^+ be the vector space generated by eigenvectors of positive eigenvalues of A, and V^- the vector space generated by eigenvectors of negative eigenvalues of A. Then $pr_{A,+}$ is the orthogonal projection from $\mathbb{R}^m$ to V^+, and $pr_{A,-}$ is the orthogonal projection from $\mathbb{R}^m$ to V^-. As usual, Id means the $m \times m$ identity matrix.

Algorithm 1: New Q-Newton's method

Result: Find a minimum of $f : \mathbb{R}^m \to \mathbb{R}$

Given: $\{\delta_0, \delta_1, \dots, \delta_m\} \subset \mathbb{R}$ (chosen **randomly**) and $\alpha > 0$;

Initialization: $x_0 \in \mathbb{R}^m$;

for $k = 0, 1, 2 \dots$ **do**

  $j = 0$

  if $\|\nabla f(x_k)\| \neq 0$ **then**

    while $\det(\nabla^2 f(x_k) + \delta_j \|\nabla f(x_k)\|^{1+\alpha} Id) = 0$ **do**

      $j = j + 1$

    end

  end

  $A_k := \nabla^2 f(x_k) + \delta_j \|\nabla f(x_k)\|^{1+\alpha} Id$

  $v_k := A_k^{-1} \nabla f(x_k) = pr_{A_k,+}(v_k) + pr_{A_k,-}(v_k)$

  $w_k := pr_{A_k,+}(v_k) - pr_{A_k,-}(v_k)$

  $x_{k+1} := x_k - w_k$

end

Experimentally, on small scale problems, New Q-Newton's method works quite competitive against the methods mentioned above, see [33]. However, while it can avoid saddle points, it does not have a descent property, and an open question in [33] is whether it has good convergence guarantees.

3. Algorithms

We now define a new algorithm, Backtracking New Q-Newton's method, incorporating Armijo's Backtracking line search into New Q-Newton's method, which resolves this convergence issue. Backtracking New Q-Newton's method uses hyperparameters in a more sophisticated manner than that of New Q-Newton's method, and we need some notations. For a symmetric, square real matrix A, we define:

$sp(A)$ = the maximum among $|\lambda|$'s, where λ runs in the set of eigenvalues of A, this is usually called the spectral radius in the Linear Algebra literature;

and

$minsp(A)$ = the minimum among $|\lambda|$'s, where λ runs in the set of eigenvalues of A, this number is non-zero precisely when A is invertible.

One can easily check the following more familiar formulas: $sp(A) = \max_{\|e\|=1} \|Ae\|$ and $minsp(A) = \min_{\|e\|=1} \|Ae\|$, using for example the fact that A is diagonalisable.

Algorithm 2: Backtracking New Q-Newton's method

Result: Find a minimum of $f : \mathbb{R}^m \to \mathbb{R}$

Given: $\{\delta_0, \delta_1, \dots, \delta_m\} \subset \mathbb{R}$ (chosen **randomly**), $0 < \tau$ and $0 < \gamma_0 < 1$;

Initialization: $x_0 \in \mathbb{R}^m$;

$\kappa := \frac{1}{2} \min_{i \neq j} |\delta_i - \delta_j|$;

for $k = 0, 1, 2 \dots$ **do**

  $j = 0$

  if $\|\nabla f(x_k)\| \neq 0$ **then**

    while $minsp(\nabla^2 f(x_k) + \delta_j \|\nabla f(x_k)\|^\tau Id) < \kappa \|\nabla f(x_k)\|^\tau$ **do**

      $j = j + 1$

    end

  end

  $A_k := \nabla^2 f(x_k) + \delta_j \|\nabla f(x_k)\|^\tau Id$

  $v_k := A_k^{-1} \nabla f(x_k) = pr_{A_k,+}(v_k) + pr_{A_k,-}(v_k)$

  $w_k := pr_{A_k,+}(v_k) - pr_{A_k,-}(v_k)$

  $\widehat{w_k} := w_k / \max\{1, \|w_k\|\}$

  $\gamma := \gamma_0$

  if $\|\nabla f(x_k)\| \neq 0$ **then**

    while $f(x_k - \gamma \widehat{w_k}) - f(x_k) > -\gamma \langle \widehat{w_k}, \nabla f(x_k) \rangle / 3$ **do**

      $\gamma = \gamma / 3$

    end

  end

  $x_{k+1} := x_k - \gamma \widehat{w_k}$

end

In the case f has compact sublevels, one can choose $\hat{w}_k = w_k$ in Backtracking New Q-Newton's method (see the discussion after the proof of Theorem 1.1 in Section 3).

In the experiments we also explore a simplified version of Backtracking New Q-Newton's method, named Simplified Backtracking New Q-Newton's method, where we replace $pr_{A_k,-}(v_k)$ by its projection to a one-dimensional subvector space corresponding to the smallest negative eigenvalue. Experiments show that this simplified version works similarly to Backtracking New Q-Newton's method. One can also consider the whole eigenspace corresponding to the smallest negative eigenvalue, instead of just one eigenvector corresponding to that eigenvalue.

Algorithm 3: Simplified Backtracking New Q-Newton's method

Result: Find a minimum of $f : \mathbb{R}^m \to \mathbb{R}$

Given: $\{\delta_0, \delta_1, \dots, \delta_m\} \subset \mathbb{R}$ (chosen **randomly**), $0 < \tau$ and $0 < \gamma_0 < 1$;

Initialization: $x_0 \in \mathbb{R}^m$;

$\kappa := \frac{1}{2} \min_{i \neq j} |\delta_i - \delta_j|$;

for $k = 0, 1, 2 \dots$ **do**

  $j = 0$

  if $\|\nabla f(x_k)\| \neq 0$ **then**

    while $minsp(\nabla^2 f(x_k) + \delta_j \|\nabla f(x_k)\|^\tau Id) < \kappa \|\nabla f(x_k)\|^\tau$ **do**

      $j = j + 1$

    end

  end

  $A_k := \nabla^2 f(x_k) + \delta_j \|\nabla f(x_k)\|^\tau Id$

  $v_k := A_k^{-1} \nabla f(x_k) = pr_{A_k,+}(v_k) + pr_{A_k,-}(v_k)$

  $minVec := 0$

  $minVal :=$ the smallest eigenvalue of A_k

  if $minVal < 0$ **then**

    minVec:= one unit eigenvector of A_k with eigenvalue minVal

  end

  $w_k := pr_{A_k,+}(v_k) - < v_k, minVec > minVec$

  $\widehat{w_k} := w_k / \max\{1, \|w_k\|\}$

  $\gamma := \gamma_0$

  if $\|\nabla f(x_k)\| \neq 0$ **then**

    while $f(x_k - \gamma \widehat{w_k}) - f(x_k) > -\gamma \langle \widehat{w_k}, \nabla f(x_k) \rangle / 3$ **do**

      $\gamma = \gamma / 3$

    end

  end

  $x_{k+1} := x_k - \gamma \widehat{w_k}$

end

4. Theoretical results

We present in this section the proofs of Theorems 1.1 and 1.2. We also discuss solving a system of equations, and extend a result in [33] concerning roots of a meromorphic function in 1 complex variable to higher dimensions.

4.1. **Proofs of Theorems 1.1 and 1.2.** We start with auxilliary lemmas.

Lemma 4.1. *Let $\delta_0, \dots, \delta_m$ be distinct real numbers. Define*

$$\kappa := \frac{1}{2} \min_{i \neq j} |\delta_i - \delta_j|.$$

Let A be an arbitrary symmetric $m \times m$ matrix, and ϵ an arbitrary real number. Then there is $j \in \{0, \dots, m\}$ so that $minsp(A + \delta_j \epsilon Id) \geq |\kappa \epsilon|$.

Proof. Let $\lambda_1, \dots, \lambda_m$ be the eigenvalues of A. By definition

$$minsp(A + \delta_j \epsilon Id) = \min_{i=1,\dots,m} |\lambda_i + \delta_j \epsilon|.$$

The RHS can be interpreted as the minimum of the distances between the points $-\lambda_i$ and $\delta_j \epsilon$. The set $\{\delta_0 \epsilon, \delta_1 \epsilon, \dots, \delta_m \epsilon\}$ consists of $m+1$ distinct points on the real line, the distance between any 2 of them is $\geq 2|\kappa \epsilon|$. Hence, by the pigeon hole principle, there is at least one j so that the distance from $-\lambda_i$ to $\delta_j \epsilon$ is $\geq |\kappa \epsilon|$ for all $i = 1, \dots, m$, which is what needed. □

The next lemma considers quantities such as $A_k, v_k, w_k, \dots$ which appear in Algorithm 2 for Backtracking New Q-Newton's method. We introduce hence global notations for them for the ease of exposition. Fix a sequence $\delta_0, \delta_1, \dots, \delta_m$ of real numbers. Define $\kappa := \min_{i \neq j} |\delta_i - \delta_j|$. Fix also a number $\alpha > 0$, and a C^2 function $f : \mathbb{R}^m \to \mathbb{R}$. For $x \in \mathbb{R}^m$ so that $\nabla f(x) \neq 0$, we define:

$\delta(x) := \delta_j$, where j is the smallest index for which $minsp(\nabla^2 f(x) + \delta_j \|\nabla f(x)\|^{1+\tau}) \geq \kappa \|\nabla f(x)\|^{(\tau)}$; (see Lemma 4.1).

$A(x) := \nabla^2 f(x) + \delta(x) \|\nabla f(x)\|^{\tau}$;

$v(x) := A(x)^{-1}.\nabla f(x)$;

$w(x) := pr_{A(x),+} v(x) - pr_{A(x),-} v(x)$;

and

$\widehat{w(x)} := w(x)/\max\{1, \|w(x)\|\}$.

Lemma 4.2. *Let f be a C^2 function and $x \in \mathbb{R}^m$ for which $\nabla f(x) \neq 0$. We have:*

1) $\|\nabla f(x)\|/sp(A(x)) \leq \|w(x)\| \leq \|\nabla f(x)\|/minsp(A(x))$.

2) (Descent direction)

$$\|\nabla f(x)\|^2/sp(A(x)) \leq \langle w(x), \nabla f(x) \rangle \leq \|\nabla f(x)\|^2/minsp(A(x)),$$
$$minsp(A(x)) \|w(x)\|^2 \leq \langle w(x), \nabla f(x) \rangle \leq sp(A(x)) \|w(x)\|^2.$$

3) In particular, Armijo's condition

$$f(x - \gamma w(x)) - f(x) \leq -\gamma \langle w(x), \nabla f(x) \rangle / 3,$$

is satisfied for all $\gamma > 0$ small enough.

Proof. We denote by $e_1, \ldots, e_m$ an orthonormal basis of eigenvectors of $A(x)$, and let λ_i be the corresponding eigenvalue of e_i. If we write:

$$\nabla f(x) = \sum_{i=1}^{m} a_i e_i = \sum_{\lambda_i > 0} a_i e_i + \sum_{\lambda_i < 0} a_i e_i,$$

then by definition

$$\begin{aligned} v(x) &= \sum_{\lambda_i > 0} a_i e_i/\lambda_i + \sum_{\lambda_i < 0} a_i e_i/\lambda_i = \sum_{\lambda_i > 0} a_i e_i/|\lambda_i| - \sum_{\lambda_i < 0} a_i e_i/|\lambda_i|, \\ w(x) &= \sum_{\lambda_i > 0} a_i e_i/|\lambda_i| + \sum_{\lambda_i < 0} a_i e_i/|\lambda_i| = \sum_{i=1}^{m} a_i e_i/|\lambda_i|. \end{aligned}$$

Hence, by calculation

$$\begin{aligned} \langle w(x), \nabla f(x) \rangle &= \sum_{i=1}^{m} a_i^2/|\lambda_i|, \\ \|\nabla f(x)\|^2 &= \sum_{i=1}^{m} a_i^2, \\ \|w(x)\|^2 &= \sum_{i=1}^{m} a_i^2/|\lambda_i|^2. \end{aligned}$$

From this, we immediately obtain 1) and 2).

3) By Taylor's expansion, for γ small enough:

$$\begin{aligned} f(x - \gamma w(x)) - f(x) &= -\gamma \langle w(x), \nabla f(x) \rangle + \frac{\gamma^2}{2} \langle \nabla^2 f(x) w(x), w(x) \rangle + o(\|\gamma w(x)\|^2) \\ &= -\gamma \langle w(x), \nabla f(x) \rangle + \gamma^2 O(\langle w(x), \nabla f(x) \rangle). \end{aligned}$$

The last equality follows from 2), where we know that $\|w(x)\|^2 \sim \langle w(x), \nabla f(x) \rangle$. Hence, when γ is small enough we obtain 3). □

Fix $0 < \gamma_0 < 1$. The above lemma shows that, for x with $\nabla f(x) \neq 0$, the following quantity (Armijo's step-size) is a well-defined positive number:

$\gamma(x) :=$ the largest number γ among the sequence $\{\gamma_0, \gamma_0/3, \gamma_0/3^2, \ldots, \gamma_0/3^j, \ldots\}$ for which

$$f(x - \gamma w(x)) - f(x) \leq -\gamma \langle w(x), \nabla f(x) \rangle / 3.$$

The next lemma explores properties of this quantity.

Lemma 4.3. *Let $\mathcal{B} \subset \mathbb{R}^m$ be a compact set so that $\epsilon = \inf_{x \in \mathcal{B}} \|\nabla f(x)\| > 0$. Then*

1) There is $\tau > 0$ for which:

$$\begin{aligned} \sup_{x\in\mathcal{B}} sp(A(x)) &\leq \tau, \\ \inf_{x\in\mathcal{B}} minsp(A(x)) &\geq 1/\tau. \end{aligned}$$

2) $\inf_{x\in\mathcal{B}} \gamma(x) > 0$.

Proof. 1) Since both $\nabla^2 f(x)$ and $\nabla f(x)$ have bounded norms on the compact set $\mathcal{B}$, and δ_j belongs to a finite set, it follows easily that $\sup_{x\in\mathcal{B}} sp(A(x)) < \infty$. (For example, this is seen by using that for a real symmetric matrix A, we have $sp(A) = \max_{\|v\|=1} |\langle Av, v\rangle|$.)

By Lemma 4.1, $\inf_{x\in\mathcal{B}} minsp(A(x)) \geq \kappa\epsilon^{1+m}$.

2) From 1) and Lemma 4.2, there is $\tau > 0$ so that for all $x \in \mathcal{B}$ we have

$$\begin{aligned} \|\nabla f(x)\|^2/\tau \leq \langle w(x), \nabla f(x)\rangle \leq \tau\|\nabla f(x)\|^2, \\ \|w(x)\|^2/\tau \leq \langle w(x), \nabla f(x)\rangle \leq \tau\|w(x)\|^2. \end{aligned}$$

From this, using Taylor's expansion as in the proof of part 3) of Lemma 4.2 and the assumption that f is C^2, it is easy to obtain the assertion that $\inf_{x\in\mathcal{B}} \gamma(x) > 0$.

□

We recall that there is so-called real projective space $\mathbb{P}^m$, which is a compact metric space and which contains $\mathbb{R}^m$ as a **topological** subspace. If $x, y \in \mathbb{R}^m$ and $d(.,.)$ is the metric on $\mathbb{P}^m$, then

$$d(x, y) := \arccos\left(\frac{1 + \langle x, y\rangle}{\sqrt{1 + \|x\|^2}\sqrt{1 + \|y\|^2}}\right).$$

It is known that there is a constant $C > 0$ so that for $x, y \in \mathbb{R}^m$ we have $d(x, y) \leq C\|x - y\|$, see e.g. [31]. Real projective spaces were used in [32, 31] to establish good convergence guarantees for Backtracking GD and modifications.

Now we prove Theorem 1.1.

Proof of Theorem 1.1. 1) This is by construction.

2) It suffices to show that if $\{x_{n_k}\}$ is a bounded subsequence of $\{x_n\}$, then $\lim_{n\to\infty} \nabla f(x_{n_k}) = 0$. If it is not the case, we can assume, by taking a further subsequence if needed, that $\inf_k \|\nabla f(x_{n_k})\| > 0$. Then, by Lemma 4.3, upto some positive constants, $\langle \nabla f(x_{n_k}), w(x_{n_k})\rangle \sim \|\nabla f(x_{n_k})\|^2 \sim \|w(x_{n_k})\|^2$ for all k. Since $\|\nabla f(x_{n_k})\|$ is bounded, it follows that $w_{n_k} \sim \widehat{w_{n_k}}$.

It follows from the fact that $\{f(x_n)\}$ is a decreasing sequence, that $\lim_{k\to\infty}(f(x_{n_k}) - f(x_{n_k+1})) = 0$. It follows from Armijo's condition that $\lim_{k\to\infty} \gamma_{n_k}\langle \nabla f(x_{n_k}), \widehat{w(x_{n_k})}\rangle = 0$. By Lemma 4.3 we have $\inf_k \gamma_{n_k} > 0$, and hence $\lim_{k\to\infty}\langle \nabla f(x_{n_k}), \widehat{w(x_{n_k})}\rangle = 0$. From the equivalence in the previous paragraph about $w_{n_k} \sim \widehat{w_{n_k}}$ and $\langle \nabla f(x_{n_k}), \widehat{w(x_{n_k})}\rangle \sim \|w_{n_k}\|^2 \sim \|\nabla f(x_{n_k})\|^2$, it follows that $\lim_{k\to\infty} \nabla f(x_{n_k}) = 0$, a contradiction.

3) We show first that $\lim_{n\to\infty} d(x_n, x_{n+1}) = 0$, where $d(.,.)$ is the projective metric defined above. It suffices to show that each subsequence $\{x_{n_k}\}$ has another subsequence for which the needed claim holds. By taking a further subsequence if necessary, we can assume that either $\lim_{k\to\infty} \|x_{n_k}\| = \infty$ or $\lim_{k\to\infty} x_{n_k} = x_\infty$ exists.

The first case: $\lim_{k\to\infty} \|x_{n_k}\| = \infty$. In this case, since $x_{n_k+1} = x_{n_k} - \gamma_{n_k}\widehat{w_{n_k}}$, where both γ_{n_k} and $\widehat{w_{n_k}}$ are bounded, it follows easily from the definition of the projective metric that $\lim_{k\to\infty} d(x_{n_k+1}, x_{n_k}) = 0$.

The second case: $\lim_{k\to\infty} x_{n_k} = x_\infty$ exists. In this case, by part 1) of Theorem 1.1, x_∞ is a critical point of f. Since f is Morse by assumption, we have that $\nabla^2 f(x_\infty)$ is invertible. Then, by the proof of part 4) below, we have that $\lim_{k\to\infty} \widehat{w_{n_k}} = 0$. Then, since γ_{n_k} is bounded, one obtain that

$$\lim_{k\to\infty} \|x_{n_k+1} - x_{n_k}\| = \lim_{k\to\infty} \|\gamma_{n_k}\widehat{w_{n_k}}\| = 0.$$

Thus by the inequality between the projective metric and the usual Euclidean norm, one obtains $\lim_{k\to\infty} d(x_{n_k+1}, x_{n_k}) = 0$.

With this claim proven, one can proceed as in [31, part 2 in the proof of Theorem 2.1], using part 2) above and the assumption that the function has at most countably many critical points, to obtain a bifurcation: either $\lim_{n\to\infty} x_n$ converges to a critical point of f, or $\lim_{n\to\infty} \|x_n\| = \infty$.

If the initial point x_0 is randomly chosen, then we know from 4) below that the limit point x_∞ cannot be a saddle point. If the function is Morse, then a critical point is either a saddle point or a local minimum, and hence we conclude that x_∞ must be a local minimum of f.

4) Assume that x_n converges to x_∞ and x_∞ is a saddle point. Then $\nabla^2 f(x_\infty)$ is invertible while $\nabla f(x_\infty) = 0$. It follows that there is a small open neighbourhood U of x_∞ so that for all $x \in U$ we have $\delta(x) = \delta_0$. By shrinking U if necessary, we also have that $\|w(x)\| \le \epsilon$ for all $x \in U$, and hence $\widehat{w(x)} = w(x)$, for all $x \in U$. Here $\epsilon > 0$ is a small positive number, to be determined later.

By Taylor's expansion, and using $A(x) = \nabla^2 f(x) + \delta_0 \|\nabla f(x)\|^\tau Id = \nabla^2 f(x) + o(1)$, we have

$$\begin{aligned} f(x - w(x)) - f(x) &= -\langle w(x), \nabla f(x)\rangle + \frac{1}{2}\langle \nabla^2 f(x) w(x), w(x)\rangle + o(\|w(x)\|^2) \\ &= -\langle w(x), \nabla f(x)\rangle + \frac{1}{2}\langle A(x) w(x), w(x)\rangle + o(\|w(x)\|^2). \end{aligned}$$

We denote by $e_1, \ldots, e_m$ an orthonormal basis of eigenvectors of $A(x)$, and let λ_i be the corresponding eigenvalue of e_i. If we write:

$$\nabla f(x) = \sum_{i=1}^m a_i e_i = \sum_{\lambda_i > 0} a_i e_i + \sum_{\lambda_i < 0} a_i e_i,$$

then by definition

$$\begin{aligned} v(x) &= \sum_{\lambda_i>0} a_i e_i/\lambda_i + \sum_{\lambda_i<0} a_i e_i/\lambda_i = \sum_{\lambda_i>0} a_i e_i/|\lambda_i| - \sum_{\lambda_i<0} a_i e_i/|\lambda_i|, \\ w(x) &= \sum_{\lambda_i>0} a_i e_i/|\lambda_i| + \sum_{\lambda_i<0} a_i e_i/|\lambda_i| = \sum_{i=1}^{m} a_i e_i/|\lambda_i|. \end{aligned}$$

Then $A(x)w(x) = \sum_{\lambda_i>0} a_i e_i - \sum_{\lambda_i<0} a_i e_i$. Therefore,

$$\begin{aligned} \frac{1}{2}\langle A(x)w(x), w(x)\rangle &= \frac{1}{2} < \sum_{\lambda_i>0} a_i e_i - \sum_{\lambda_i<0} a_i e_i, \sum_{i=1}^{m} a_i e_i/|\lambda_i| > \\ &= \frac{1}{2}(\sum_{\lambda_i>0} a_i^2/|\lambda_i| - \sum_{\lambda_i<0} a_i^2/|\lambda_i|) \\ &\leq \frac{1}{2}\sum_{i=1}^{m} a_i^2/|\lambda_i| \\ &= \frac{1}{2}\langle w(x), \nabla f(x)\rangle. \end{aligned}$$

Combining all the above, we obtain:

$$f(x - w(x)) - f(x) \leq -\frac{1}{2}\langle w(x), \nabla f(x)\rangle + o(\|w(x)\|^2),$$

for all $x \in U$. By Lemma 4.2, for $x \in U$ we have $\langle w(x), \nabla f(x)\rangle \sim \|w(x)\|^2$. Hence, $\gamma(x) = \gamma_0$ for all $x \in U$.

This means that $x_{n+1} = x_n - \gamma_0 w_n$ for n large enough, that is, Backtracking New Q-Newton's method becomes New Q-Newton's method. We recall from [33, Proof of part 1 of Theorem 3.1 for existence of local-stable manifolds] that if $\tau > 1$, then there are local stable manifolds near saddle points of f for the dynamics of BNQN. However, that proof does not show **global** avoidance of saddle points, and this is finished in Subsection 4.2 below. The case $0 < \tau \leq 1$ is similarly proven: the same integral representation of $pr_{A(x),\pm}$ shows again that the associated dynamical system is C^1, and hence the above arguments are applicable.

5) and 6): follow easily from the corresponding parts in [33, parts 4 and 5 of Theorem 3.1] and the above parts. □

Proof of Theorem 1.2. 1) As in part 2) of Theorem 1.1, it suffices to show that for a subsequence x_{n_k} converging to a critical point x^*: $\lim_{k\to\infty} \|w_{n_k}\| = 0$. We look at a new function $F : \mathbb{R}^{2m} \to \mathbb{R}$ given by $F(x, y) = \langle \nabla f(x), y\rangle$. We have $F(x = x^*, y = 0) = 0$. Moreover, $\nabla F = (\nabla^2 f(x).y, \nabla f(x))$. Hence, by assumption, there exists $C > 0$ and $\mu < 1$, and an open neighborhood U of x^* as well as $\epsilon > 0$ so that for all $x \in U$ and all $\|y\| = 1$:

$$|\langle \nabla f(x), y\rangle|^\mu \leq C(\|\nabla^2 f(x).y\| + \|\nabla f(x)\|).$$

We let $\{e_{j,k}\}_{j=1,\dots,m}$ be an orthonormal basis of $A_k := A(x_k)$, with the corresponding eigenvalues $\{\lambda_{j,k}\}_{j=1,\dots,m}$. Then, by replacing C with a bigger constant, and substituting x by x_{n_k} and y by $e_{j,k}$ we obtain also the inequality

$$|\langle \nabla f(x_{n_k}), e_{j,k}\rangle|^\mu \leq C(\|\nabla^2 f(x_{n_k}).e_{j,k} + \delta\|\nabla f(x_{n_k})\|e_{j,k}\| + \|\nabla f(x_{n_k})\|).$$

If we choose δ appropriately from the finite set $\delta_0, \dots, \delta_m$, we obtain $A(x_{n_k})$.

Since $\tau \leq 1$ and $\|e_{j,k}\| = 1$, it follows that $\|A(x).e_{j,k}\| \geq minsp(A(x)) \geq \kappa\|\nabla f(x)\|$. Therefore, by replacing C by a bigger constant, we obtain that $|\langle \nabla f(x_{n_k}), e_{j,k}\rangle|^\mu \leq C\|A(x).e_{j,k}\|$ for every j. Therefore, we obtain

$$\|w_{n_k}\| \leq C\|\nabla f(x_{n_k})\|^{1-\mu} \to 0,$$

as wanted.

2) Similar to [33], we need to show that if x^* is a critical point of f, there exists a constant $C > 0$ and $\theta < 1$ so that if a subsequence $\{x_{n_k}\}$ converges to x^*, then:

$$\begin{aligned} \lim_{k\to\infty} \|w_{n_k}\| &= 0, \\ \langle w_{n_k}, \nabla f(x_{n_k})\rangle &\geq C\|w_{n_k}\| \times |f(x_{n_{k+1}}) - f(x_{n_k})|^\theta. \end{aligned}$$

The equality follows from part 2 of Theorem 1.1 (if $0 < \tau < 1$) and from part 1 above if $\tau = 1$.

Now we prove the inequality. We have

$$\|A_{n_k}.e_{j,n_k}\| \geq minsp(A_{n_k}) \geq \kappa\|\nabla f(x_{n_k})\|^\tau.$$

Then, computed as in [**?**], we obtain

$$\langle w_{n_k}, \nabla f(x_{n_k})\rangle \geq C\|w_{n_k}\| \times \|\nabla f(x_{n_k})\|^{1+\tau}.$$

From this, the claim follows by the definition of $\mu(x^*)$: we can choose $\theta = \mu(1+\tau) < 1$ for some choices of $\mu \geq \mu(x^*)$. □

4.2. **The proof that BNQN globally avoid saddle points (part 4 of Theorem 1.1).** We need the following technical result. We refer the reader to Algorithm 1 for the notations $pr_\pm$ and other relevant notations.

Theorem 4.4. *Let $f : \mathbb{R}^k \to \mathbb{R}$ be a C^2 function. Define $U = \mathbb{R}^k \backslash \{x \in \mathbb{R}^k : \nabla f(x) = 0\}$. For $x \in U$ and $\delta \in \mathbb{R}$ we define $A(x,\delta) = \nabla^2 f(x) + \delta||\nabla f(x)||^\alpha Id$.*

1) There exists a set $\Delta \subset \mathbb{R}$ of Lebesgue measure 0 such that: For all $\delta \in \mathbb{R}\backslash\Delta$, the set $\{x \in U : A(x,\delta)$ is not invertible $\}$ has Lebesgue measure 0.

2) Fix $\delta \in \mathbb{R}\backslash\Delta$. We define $v(x,\delta) = A(x,\delta)^{-1}.\nabla f(x)$ and $w(x,\delta) = pr_{+,A(x,\delta)}.v(x,\delta) - pr_{-,A(x,\delta)}.v(x,\delta)$. Let $U_\delta = U\backslash\{x \in U : \det(A(x,\delta)) = 0\}$. Then $w(x,\delta)$ is continuous on U_δ. If we assume moreover that f is C^3, then $w(x,\delta)$ is C^1 on U_δ.

3) Let the setting be as in part 2). Define $H(x,\delta) = x - w(x,\delta)$, as in the definition for New Q-Newton's method. Let $Jac(H(x,\delta))$ be the Jacobian matrix of the map $H(x,\delta)$. There is a set $\Delta' \subset \mathbb{R}\backslash\Delta$ of Lebesgue measure 0 so that for all $\delta \in \mathbb{R}\backslash(\Delta \cup \Delta')$, the set $\{x \in U_\delta : \ Jac((H)(x,\delta))$ is not invertible$\}$ has Lebesgue measure 0.

4) Let the setting be as in part 3). There is a set Δ" $\subset \mathbb{R}$ of Lebesgue's measure zero so that the following is satisfied. If $\delta \in \mathbb{R}\backslash\Delta$" and $\mathcal{E} \subset \mathbb{R}^k$ has Lebesgue measure 0, then $H(.,\delta)^{-1}(\mathcal{E}) \subset \mathbb{R}^k$ also has Lebesgue's measure 0.

Proof. 1) Define $p(x,\delta) = \det(A(x,\delta))$. For $x \in U$, then $p(x,\delta)$ is a polynomial in δ of degree exactly k. Note that $A(x,\delta)$ is not invertible if and only if $p(x,\delta) \neq 0$. We consider the set $\Gamma = \{(x,\delta) \in U \times \mathbb{R} : \ p(x,\delta) = 0\}$. Then Γ is a measurable subset of $U \times \mathbb{R}$, and we need to show the existence of a set $\Delta \subset \mathbb{R}$ of Lebesgue measure 0 so that for all $\delta \in \mathbb{R}\backslash\Delta$, the set $\Gamma_\delta = \{x \in U : \ p(x,\delta) = 0\}$ has Lebesgue measure 0.

First, we will show that Γ has Lebesgue measure 0. Let $1_\Gamma : \ U \times \mathbb{R} \to \{0,1\}$ be the characteristic function of Γ. For a set A, we denote by $|A|$ its Lebesgue measure. Then, by Fubini-Tonneli's theorem in integration theory for non-negative functions, one obtains:

$$|\Gamma| = \int_{U\times\mathbb{R}} 1_\Gamma(x,\delta)d(x,\delta) = \int_U (\int_{\mathbb{R}} 1_\Gamma(x,\delta)d\delta)dx = 0.$$

This is because for each $x \in U$, the set $\{\delta \in \mathbb{R} : \ p(x,\delta) = 0\} \subset \mathbb{R}$ has cardinality at most k, and hence has Lebesgue measure 0.

Applying Fubini-Tonneli's theorem again, but now doing the integration on U first, one obtains that for a.e. $\delta \in \mathbb{R}$, the set $\Gamma_\delta = \{x \in U : \ p(x,\delta) = 0\} \subset U$ has Lebesgue measure 0.

2) The claims follow from perturbation theory for linear operators [**?**], where $pr_{\pm,A(x,\delta)}$ can be represented via Cauchy's integrations on the complex plane containing the variable δ. The readers can see for example [33] on details how the arguments go.

3) For each $x \in U$, define $\Delta_x = \{\delta \in \mathbb{R} : \ p(x,\delta) = 0\}$. Then, Δ_x is a finite set (consisting of zeros of the polynomial $p(x,\delta)$).

By calculating, we see that when it is legit (c.f. part 2 above), then $Jac(H(.,\delta))$ is invertible at x iff $q(x,\delta) = det(Jac(H(.,\delta))) \neq 0$.

Now, we fix $x_0 \in U$ and $\delta_0 \in \Delta_{x_0}$. We note that there is an open interval $(a,b) \subset \mathbb{R}$ containing δ_0, on which $Jac(H(x,\delta))$ is legit and $-w(x,\delta)$ is a rational function of δ, for x close to x_0. Here is how to see this. We can write $\delta = \delta_0 + \epsilon$, where ϵ is a small real number. Then, for x close to x_0 and δ close to δ_0, the eigenvalues of $A(x,\delta)$ are close to those of $A(x_0,\delta_0)$. Therefore, we can choose two close curves γ^+ and γ^- in the complex plane containing the variable ϵ, so that for $\delta = \delta_0 + \epsilon$

close to δ_0

$$\begin{aligned} w(x,\delta_0+\epsilon) &= \frac{1}{2\pi i}\int_{\gamma^+}((A(x,\delta_0)+\epsilon||\nabla f(x)||^{\alpha})^{-1}-\zeta)^{-1}\nabla f(x)d\zeta \\ &\quad -\frac{1}{2\pi i}\int_{\gamma^-}((A(x,\delta_0)+\epsilon||\nabla f(x)||^{\alpha})^{-1}-\zeta)^{-1}\nabla f(x)d\zeta. \end{aligned}$$

The above formula for $w(x,\delta_0+\epsilon)$ clearly shows that it is a rational function on ϵ, of degree exactly k.

It follows that either $q(x,\delta_0+\epsilon)$ is zero identically, or its set of zero has Lebesgue's measure zero. We will show that the former case cannot happen. For simplicity, we will show this for $x=x_0$. Indeed, while we defined the function

$$\epsilon \mapsto \det(\nabla_x(x-w(x,\delta_0+\epsilon)))$$

only for ϵ small numbers, it is no problem to extend its domain of definition to the maximum domain $\mathbb{R}$.

Let $\zeta_1,\ldots,\zeta_l>0$ and $\zeta_{l+1},\ldots,\zeta_k<0$ be all eigenvalues of $A(x,\delta_0)$. Let also $e_1,\ldots,e_l$ and $e_{l+1},\ldots,e_k$ be the corresponding orthonormal basis. If $\nabla f(x)=a_1e_1+\ldots+a_ke_k$, then

$$w(x,\delta_0+\epsilon)=\sum_{\beta\leq l}\frac{a_\beta}{\zeta_\beta+\epsilon||\nabla f(x)||^{\alpha}}e_\beta-\sum_{\beta>l}\frac{a_\beta}{\zeta_\beta+\epsilon||\nabla f(x)||^{\alpha}}e_\beta.$$

From this, we can see that $\lim_{\epsilon\to\infty}\nabla_x w(x,\delta_0+\epsilon)=0$. Therefore, one sees that $\lim_{\epsilon\to\infty}\nabla_x(x-w(x,\delta_0+\epsilon))=Id$. Hence, in particular, we obtain that $q(x_0,\delta_0+\epsilon)$ is not zero identically.

Combining all the above, here is a summary of what we proved so far: Except for a countable set of values $\delta\in\mathbb{R}$, the map $x\mapsto x-w(x,\delta)$ is C^1 and is a local diffeomorphism. With this, now we finish the proof of this part 3). As in the proof of part 1), we define the set Γ' to consist of all pairs (x,δ) for which $A(x,\delta)$ is invertible but the gradient of $H(x,\delta)$ is not invertible. The slices $\Gamma'_x=\{\delta:\ (x,\delta)\in\Gamma'\}$ have been shown above to have Lebesgue's measure zero. Hence, by the use of Fubini-Tonelli's theorem, we get that the slices $\Gamma'_\delta=\{x:\ (x,\delta)\in\Gamma'\}$ also have Lebesgue's measure zero for $\delta\in\mathbb{R}$ outside of a set of Lebesgue's measure 0.

4) This is a consequence of part 3 and the following well-known fact:

Fact. Let $H:U\to V$ be a map, where U,V are open subsets in $\mathbb{R}^k$. Assume that there is a set $\mathcal{E}'\subset U$ of Lebesgue's measure 0 so that for all $x\in U\backslash\mathcal{E}'$ then H is C^1 near x and $Jac(H)$ is invertible there. Then, if $\mathcal{E}\subset V$ has Lebesgue's measure 0, it follows that $H^{-1}(\mathcal{E})\subset U$ has Lebesgue's measure 0.

□

Now we can continue the proof of part 4 of Theorem 1.1.

By [33], each saddle point of f has a local Stable-Center manifold for the dynamics of New Q-Newton's method. Each of these manifolds has Lebesgue measure 0. By Theorem 4.4, there is

a set $\mathcal{E}_1$ of Lebesgue measure 0 so that the dynamics H of New Q-Newton's method is C^1 and has invertible Jacobian outside of $\mathcal{E}_1$. If we take $\mathcal{E}$ be the countable union of the iterations of inverse images of $\mathcal{E}_1$ and these local Stable-Center manifold by the dynamics H (in small subsets where H is invertible), then $\mathcal{E}$ is also of Lebesgue measure 0. Therefore, if x_0 is an initial point outside of $\mathcal{E}$, then its orbit will not encounter $\mathcal{E}_1$ or local Stable-Center manifolds of saddle points. In particular, the constructed sequence will not converge to any of the saddle points.

4.3. **Solving systems of equations.** Now we discuss an application of the above results to solving systems of equations. Assume that we are given a real analytic map $F : \mathbb{R}^m \backslash \mathcal{A} \to \mathbb{R}^{m'}$, where $\mathcal{A}$ is a closed set of $\mathbb{R}^m$ of Lebesgue measure 0 for which: $\lim_{x\to\mathcal{A}} \|F(x)\| = +\infty$. We consider the question of finding solutions to $F(x) = 0$. Following the usual trick, we define $f(x) = \frac{1}{2}\|F(x)\|^2 : \mathbb{R}^m \to [0, +\infty]$. A solution to $F = 0$ is a global minimum of f, so we can consider the question of finding global minima to f (which still makes sense even if the system $F = 0$ has no solutions). Since $\mathcal{A}$ has Lebesgue measure zero and since the sequence constructed by Backtracking New Q-Newton's method has decreasing function values, if one chooses a random point x_0 then it will belong to $\mathbb{R}^m\backslash\mathcal{A}$ and the whole sequence $\{x_n\}$ will also belong to $\mathbb{R}^m\backslash\mathcal{A}$. Moreover, any cluster point of $\{x_n\}$ will be also in $\mathbb{R}^m\backslash\mathcal{A}$. One can apply Theorems 1.1 and 1.2 to show that a cluster point of $\{x_n\}$ must be a critical point of f. Moreover, the constructed sequence $\{x_n\}$ either **converges** to a unique critical point of f or the norms $\{\|x_n\|\}$ converges to $+\infty$, provided conditions in those mentioned theorems are satisfied. (For example, if f has at most countably many critical points - which is a generic condition; or if F is a polynomial map and τ is small enough - depending on the degree of F. Indeed, it is shown in [2] that the Lojasiewicz exponent of a polynomial map is globally bounded in terms of the map's degree, and hence if we choose τ small enough then conditions in part 2 of Theorem 1.2 are satsified.) Part 4 of Theorem 1.1 then allows one to conclude that the limit point of $\{x_n\}$ cannot be a saddle point, hence roughly must be a local minimum of f.

There is need for much more further research, since the above results are still not enough to solve the system $F = 0$. On the one hand, local minima of f may not be global minima. On the other hand, the requirement that τ is small enough may be difficult to check, since strictly speaking one then must know an upper bound for the global Lojasiewicz exponent, which can be too small to be useful in numerical calculations, even in the case where F is a polynomial map the bound in [2] is extremely small compared with the degree. Note, however, on the experimental side, see the next Section, Backtracking New Q-Newton's method works very well and very flexible on its parameters (for example, the algorithm run wells on various examples even if one chooses $\tau = 1$).

For the case where $g : \mathbb{C}\backslash\mathcal{A} \to \mathbb{C}$ is a holomorphic function with poles at $\mathcal{A}$, and one defines $F(x, y) = (Re(g(x + iy)), Im(g(x + iy))) : \mathbb{R}^2\backslash\mathcal{A} \to \mathbb{R}^2$ consisting of the real and imaginary parts of $g(z)$ (where $z = x + iy$), then [33] shows that the above obstacles can be resolved provided that g satisfies the following generic condition: For every $z_0 \in \mathbb{C}$, if $g(z_0)g"(z_0) = 0$ then $g'(z_0) \neq 0$.

Indeed, under this generic condition, it is proven in [33] that if one applies Backtracking New Q-Newton's method (no matter what value of τ is) for the function f, from a random initial point $z_0 \in \mathbb{R}^2 = \mathbb{C}$, then for the constructed sequence $\{z_n\}$ one has a dichotomy: either $\{z_n\}$ converges to a zero z^* of $g(z)$, or $\lim_{n\to\infty} \|z_n\| = +\infty$. The main ingredient in [33] is to make use of Cauchy-Riemann's equations to relate critical points of f which are not zeros of F to saddle points of f. The following result generalises part of the mentioned result in [33].

Theorem 4.5. *Assume that $f : \mathbb{R}^m \to \mathbb{R}$ satisfies the Lojasiewicz gradient inequality. Let $\{x_n\}$ be a sequence constructed by Backtracking New Q-Newton's method. Let $\{e_k\}$ be an orthonormal basis for $A_k := A(x_k)$. Assume also that $0 < \tau \leq 1$, and that whenever a subsequence x_{k_n} converges, we have:*

$$\liminf_{n\to\infty} \frac{\min_{i\in\Lambda_{k_n}} \|A_{k_n}.e_{i,k}\|}{\max_{i\in\Lambda_{k_n}} \|A_{k_n}.e_{i,k_n}\|} > 0,$$

where $\Lambda_{k_n} = \{i : \ \langle \nabla f(x_{k_n}), e_{i,k_n}\rangle \neq 0\}$. (Moreover, if $\tau = 1$, then assume that ∇f also satisfies the Lojasiewicz gradient inequality.)

Then either $\lim_{n\to\infty} \|x_n\| = \infty$ or $\{x_n\}$ converges to a critical point of f. .

Proof. The proof is similar to that in [33]. □

Remark 4.6. The condition in Theorem 4.5 simply means that the eigenvalues of the concerned matrices A_{k_n} are of the same magnitude.

One can also combine the ideas of Backtracking New Q-Newton's method with that of Regularized Newton's method an Levenberg-Marquardt's method to obtain methods with better theoretical guarantees and performances, specifically for systems of equations. Below we present two such algorithms, together with some theoretical results. Given $F : \mathbb{R}^m \to \mathbb{R}^{m'}$ a C^2 function, we denote by $H(x) = JF(x)$ the Jacobian of F, and $f(x) = \|F(x)\|^2$. Note that $\nabla f(x) = 2H(x)^\intercal F(x)$. Hence the last While loop in the algorithm terminates after a finite time, see the previous Subsection.

Algorithm 4: Backtracking New Q-Newton's method SE

Result: Find a zero of $F : \mathbb{R}^m \to \mathbb{R}^{m'}$

Given: $0 < \delta_0, \delta_1, \dots, \delta_m$ and $0 < \tau$;

Define: $H(x) := JF(x)$ and $f(x) = \|F(x)\|^2$;

Define: $\kappa := \inf_{i \neq j} |\delta_i - \delta_j|/2$;

Initialization: $x_0 \in \mathbb{R}^m$;

for $k = 0, 1, 2 \dots$ **do**

  $j = 0$

  if $F(x_k) \neq 0$ **then**

    if $minsp(\nabla^2 f(x_k)) > \|F(x_k)\|^\tau$ **then**

      while $minsp(\nabla^2 f(x_k) + \delta_j \|F(x_k)\|) < \kappa \|F(x_k)\|$ **do**

        j:=j+1

      end

      $A_k := \nabla^2 f(x_k) + \delta_j \|F(x_k)\| \times Id$

      else

      while $minsp(\nabla^2 f(x_k) + \delta_j \|F(x_k)\|^\tau) < \kappa \|F(x_k)\|^\tau$ **do**

        j:=j+1

      end

      $A_k := \nabla^2 f(x_k) + \delta_j \|F(x_k)\|^\tau \times Id$

    end

  end

  $v_k := A_k^{-1} H(x_k)^\intercal F(x_k) = pr_{A_k,+} v_k + pr_{A_k,-} v_k$

  $w_k := pr_{A_k,+} v_k - pr_{A_k,-} v_k$

  $\widehat{w_k} := w_k / \max\{1, \|w_k\|\}$

  $\gamma := 1$

  if $H(x_k)^\intercal F(x_k) \neq 0$ **then**

    while $f(x_k - \gamma \widehat{w_k}) - f(x_k) > -\gamma \langle \widehat{w_k}, H(x_k)^\intercal F(x_k) \rangle$ **do**

      $\gamma = \gamma/2$

    end

  end

  $x_{k+1} := x_k - \gamma \widehat{w_k}$

end

Algorithm 5: Backtracking Levenberg-Marquardt's algorithm

Result: Find a zero of $F : \mathbb{R}^m \to \mathbb{R}^{m'}$

Given: $0 < \delta_0, \delta_1$ and $0 < \tau$;

Define: $H(x) := JF(x)$ and $f(x) = \|F(x)\|^2$;

Define: $\kappa := |\delta_1 - \delta_0|/2$;

Initialization: $x_0 \in \mathbb{R}^m$;

for $k = 0, 1, 2 \dots$ **do**

  $j = 0$

  if $F(x_k) \neq 0$ **then**

    if $minsp(H(x_k)^\intercal H(x_k)) > \|F(x_k)\|^\tau$ **then**

      $A_k := H(x_k)^\intercal H(x_k) + \delta_0 \|F(x_k)\| \times Id$

      else

      $A_k := H(x_k)^\intercal H(x_k) + \delta_1 \|F(x_k)\|^\tau \times Id$

    end

  end

  $w_k := A_k^{-1} H(x_k)^\intercal F(x_k)$

  $\widehat{w_k} := w_k / \max\{1, \|w_k\|\}$

  $\gamma := 1$

  if $H(x_k)^\intercal F(x_k) \neq 0$ **then**

    while $f(x_k - \gamma \widehat{w_k}) - f(x_k) > -\gamma \langle \widehat{w_k}, H(x_k)^\intercal F(x_k) \rangle$ **do**

      $\gamma = \gamma/2$

    end

  end

  $x_{k+1} := x_k - \gamma \widehat{w_k}$

end

We have the following result. Note that by definition, for the A_k in either algorithm, we have $minsp(A_k) \geq \|F(x_k)\|^\tau$, and near a non-degenerate root of $F(x)$ we have $A_k = H(x_k)^\intercal H(x_k) + O(\|F(x_k)\|)$.

Theorem 4.7. *Let $F : \mathbb{R}^m \to \mathbb{R}^{m'}$ be a C^1 function. Define $H(x) = JF(x)$ and $f(x) = \|F(x)\|^2$. Let $x_0 \in \mathbb{R}^m$ be an initial point, and $\{x_n\}$ the corresponding constructed sequence from New Q-Newton's method Backtracking SE or Levenberg-Marquardt M.*

0) (Descent property) $f(x_{n+1}) \leq f(x_n)$ for all n.

1) If x_∞ is a **cluster point** *of $\{x_n\}$, then $H(x_\infty)^\intercal F(x_\infty) = 0$. That is, x_∞ is a* **critical point** *of f.*

2) If $0 < \tau < 1$: If f has at most countably many critical points, then either $\lim_{n\to\infty} \|x_n\| = \infty$ or $\{x_n\}$ converges to a critical point of f. Moreover, if f has compact sublevels, then only the second alternative happens.

3) If x_n converges to x_∞ which is a non-degenerate zero of F (that is, if $H(x)$ is invertible at x_∞), then the rate of convergence is quadratic.

Proof. Parts 0, 1 and 2 follow exactly as in for Backtracking New Q-Newton's method.

Part 3: this follows because near x_∞ then the update rule for either algorithm is different from the usual Newton's method only in $O(\|F(x_k)\|)$, which will still has quadratic rate of convergence. □

Next, we discuss the avoidance of saddle points. Recall that a point x^* is a generalised saddle point of a function f if $\nabla f(x^*) = 0$, and moreover $\nabla^2 f(x^*)$ has at least one negative eigenvalue. Note that if x^* is a generalised saddle point of $f(x) = \|F(x)\|^2$, then x^* cannot be a root of $F(x) = 0$. In particular, since $\nabla f(x)/2 = H(x)^\intercal F(x)$, we have that $H(x^*)$ is singular, and hence $A(x) = H(x)^\intercal H(x) + \delta_1 \|F(x)\|^\tau$ near x^*.

We consider the Backtracking Levenberg-Marquardt's method first, which is more complicated to deal with. The main reason is that the main term $H(x)^\intercal H(x)$ is not the same as $\nabla^2 f(x)$. Note that the dynamics of Backtracking Levenberg-Marquardt's method is $x \mapsto G(x) = x - \gamma(x)(H(x)^\intercal H(x) + \delta_1 \|F(x)\|^\tau)^{-1}.\nabla f(x)/2$, and it is easy to compute that $G(x) = (H(x)^\intercal H(x) + \delta_1 \|F(x)\|^\tau)^{-1}.\nabla f(x)/2$ is C^1 near x^*, and $\nabla G(x^*) = (H(x^*)^\intercal H(x^*) + \delta_1 \|F(x^*)\|^\tau)^{-1}.\nabla^2 f(x^*)/2$. Since $H(x^*)^\intercal H(x^*) + \delta_1 \|F(x^*)\|^\tau$ is strictly positive definite near x^*, it follows that the term $(H(x^*)^\intercal H(x^*) + \delta_1 \|F(x^*)\|^\tau)^{-1}.\nabla^2 f(x^*)/2$ also has at least one negative eigenvalue. Therefore, if $\gamma(x)$ varies C^1 near x^* (for example, if it is a constant), then one has a local Stable-Central manifold for x^* for the dynamics of Backtracking Levenberg-Marquardt's method, and then can use the results in the previous Subsection and in Subsection 4.2 to show (when δ_0, δ_1 are randomly chosen from beginning) that Backtracking Levenberg-Marquardt's algorithm can globally avoid x^*.

There is one way to make $\gamma(x)$ to be constant near x^*, that is to choose $\gamma(x)$ by a more sophisticated manner. The main idea (in arXiv:2008.11091), used for Backtracking line search for gradient descent, is to find a continuous quantity $R(x)$, so that if $\gamma < R(x)$ then Armijo's condition is satisfied. Then, if one chooses $\gamma(x)$ in a sequence $\{\beta^n : \ n = 0, 1, 2, \ldots\}$, where $0 < \beta < 1$ is randomly chosen, then one can make sure (in case x^* is an isolated saddle point) that $\gamma(x)$ is constant near x^*. In the case at hand, such an $R(x)$ can be bounded from the quantities involving $F(x)$ at x^*. The case of non-isolated saddle points can be treated similarly, by using Lindelof's lemma (which is first used in [**?**] for the usual Gradient descent method). Therefore, we have the following result.

Theorem 4.8. *If $F : \mathbb{R}^m \to \mathbb{R}^{m'}$ is C^2, and $f = \|F\|^2$, then by choosing δ_0, δ_1 randomly, as well as choosing the learning rate $\gamma(x)$ in a sequence $\{\beta^n : \ n = 0, 1, 2, \ldots\}$ (where $0 < \beta < 1$ is randomly chosen) by the manner in arXiv:2008.11091 mentioned above, then for a random initial point the sequence $\{x_n\}$ constructed by Backtracking Levenberg-Marquardt's algorithm cannot converge to a generalised saddle point.*

However, it is still preferable to show avoidance of saddle points for the original version of Backtracking Levenberg-Marquardt's algorithm here, since it is simpler. It is expected that settling this question - at least locally near the saddle point - is the same as settling the question of whether the original version of Armijo's Backtracking line search for Gradient descent can avoid saddle points, and the latter question is still open. (As mentioned, a more sophisticated choice of learning rate for Backtracking line search for Gradient descent can avoid saddle points.) Here, by using the ideas in the previous Subsection, we can show that Backtracking Levenberg-Marquardt's algorithm can avoid saddle points of a special type, which is described next.

Again, let $f(x) = \|F(x)\|^2$ and x^* a saddle point of f. Then $\nabla^2 f(x^*)/2 = H(x^*)^\intercal H(x^*) + \nabla^2 F(x^*).F(x^*)$ has at least 1 negative eigenvalue. Since $H(x^*)^\intercal H(x^*)$ is semi-positive definite, the matrix $\nabla^2 F(x^*).F(x^*)$ also has at least one negative eigenvalue. The special generalised saddle points we concern are:

Strong generalised saddle points: x^* is a strong generalised saddle point of $f(x) = \|F(x)\|^2$ if it is a generalised saddle point of f, and moreover $\nabla^2 F(x^*).F(x^*)$ is negative definite.

Theorem 4.9. *Let $F : \mathbb{R}^m \to \mathbb{R}^{m'}$ be C^2 and $f(x) = \|F(x)\|^2$. Assume that δ_0, δ_1 are chosen randomly. If x_0 is a random initial point, and $\{x_n\}$ is the sequence constructed from Backtracking Levenberg-Marquardt's algorithm, then $\{x_n\}$ cannot converge to a strong generalised saddle point x^* of f.*

Proof. As mentioned before the statement of Theorem 4.8, it suffices to show that for x close to x^*, then the learning rate $\gamma(x) = 1$. Denote, as usual, $w(x) = (H(x)^\intercal H(x) + \delta_1 \|F(x)\|^\tau)^{-1}.H(x)^\intercal.F(x)$. Note that $\|w(x)\| \sim \|H(x)^\intercal.F(x)\|$ Then, by Taylor's expansion we have

$$\begin{aligned} f(x - w(x)) - f(x) &= -2\langle w(x), H(x)^\intercal.F(x)\rangle \\ &\quad + \langle [H(x)^\intercal H(x) + \nabla^2 F(x).F(x)].w(x), w(x)\rangle + o(\|w(x)\|^2) \end{aligned}$$

Since $\nabla^2 F(x).F(x)$ is negative definite when x is close to x^*, and since $\delta_1 > 0$, we obtain

$$\begin{aligned} \langle [H(x)^\intercal H(x) + \nabla^2 F(x).F(x)].w(x), w(x)\rangle &\leq \langle [H(x)^\intercal H(x) + \delta_1 \|F(x)\|^\tau] w(x), w(x)\rangle \\ &= \langle H(x)^\intercal.F(x), w(x)\rangle. \end{aligned}$$

Hence, for x near x^*, we have as wanted: $f(x - w(x)) - f(x) \leq -1/3\langle w(x), H(x)^\intercal.F(x)\rangle$. □

Note that when $m = 1$, then a saddle point of f is also a strong generalised saddle point. On the other hand, for $m = 2$, the two notions are different, already for the case $f = |g(z)|^2$ where g is a univariate holomorphic function (see [33]). In particular, we do not know if there is a guarantee for avoidance of saddle points if we apply Backtracking Levenberg-Marquardt's algorithm to the system of 2 real equations associated to an analytic function in one complex variable.

Concerning avoidance of saddle points, the next result shows that currently Backtracking New Q-Newton's method SE has better theoretical guarantees than Backtracking Levenberg-Marquardt's method. The proof is similar to the proof of Theorem 4.9.

Theorem 4.10. *Let $F : \mathbb{R}^m \to \mathbb{R}^{m'}$ be C^2 and $f(x) = \|F(x)\|^2$. Assume that δ_0, δ_1 are chosen randomly. If x_0 is a random initial point, and $\{x_n\}$ is the sequence constructed from Backtracking New Q-Newton's method SE, then $\{x_n\}$ cannot converge to a generalised saddle point x^* of f.*

Like the case of Backtracking New Q-Newton's method, under the assumption on Lojasiewicz inequality (around critical points), one can prove convergence for Backtracking New Q-Newton's method SE. On the other hand, we do not know if such a convergence can be proven also for Backtracking Levenberg-Macquardt's algorithm.

Theorem 4.11. *Assume that $F : \mathbb{R}^m \to \mathbb{R}^{m'}$ is C^1 so that $f = \|F\|^2$ satisfies the Lojasiewicz gradient inequality. Let $\{x_n\}$ be a sequence constructed by Backtracking New Q-Newton's method SE. Assume also that $0 < \tau < 1$.*

1) Assume that for all critical points x^* of f, we have $\mu(x^*) \times (1 + \tau) < 1$. Then either $\lim_{n\to\infty} \|x_n\| = \infty$ or $\{x_n\}$ converges to a critical point of f.

2) If F is a polynomial map, then the condition in part 1) is satisfied, provided $\tau > 0$ is small enough.

Proof. The proof is similar to what given for Backtracking New Q-Newton's method, by using the effective upper bound [2] for the Lojasiewicz exponent of a polynomial map (in terms of degree of the map and the dimension). □

5. Experimental results on various variants of Newton's method

5.1. **Implementation details for Backtracking New Q-Newton's method.** In this Subsection, we present some practical points concerning implementation details, for the language Python. Source code is in the GitHub link [38].

Indeed, Python has already enough commands to implement New Q-Newton's method. There is a package, named numdifftools, which allows one to compute approximately the gradient and Hessian of a function. This package is also very convenient when working with a family $f(x, t)$ of functions, where t is a parameter. Another package, named scipy.linalg, allows one to find (approximately) eigenvalues and the corresponding eigenvectors of a square matrix. More precisely, given a square

matrix A, the command $eig(A)$ will give pairs (λ, v_λ) where λ is an approximate eigenvalue of A and v a corresponding eigenvector.

One point to notice is that even if A is a symmetric matrix with real coefficients, the eigenvalues computed by the command eig could be complex numbers, and not real numbers, due to the fact that these are approximately computed. This can be easily resolved by taking the real part of λ, which is given in Python codes by $\lambda.real$. We can do similarly for the eigenvectors. A very convenient feature of the command eig is that it already computes (approximate) orthonormal bases for the eigenspaces. Note that similar commands are also available on PyTorch and TensorFlow, two popular libraries for implementing Deep Neural Networks.

Now we present the coding detail of the main part of New Q-Newton's method: Given a symmetric invertible matrix A with real coefficients (in our case $A = \nabla^2 f(x_n) + \delta_j \|\nabla f(x_n)\|^{1+\alpha}$), and a vector v, compute w which is the reflection of $A^{-1}.v$ along the direct sum of eigenspace of negative eigenvectors of A. First, we use the command eig to get pairs $\{(\lambda_j, v_j)\}_{j=1,\dots,m}$. Use the command real to get real parts. If we write $v = \sum_{j=1}^m a_j v_j$, then $a_j = \langle v_j, v\rangle$ (the inner product), which is computed by the Python command $np.dot(v_j, v)$. Then $v_{inv} = A^{-1}v = \sum_{j=1}^m (a_j/\lambda_j)v_j$. Finally,

$$w = v_{inv} - 2 \sum_{j:\ \lambda_j<0} (a_j/\lambda_j)v_j.$$

We remark that in actual implementations we simplify many technical details in the algorithms, and still observe good and stable behaviours.

5.2. **Experimental results.** In this subsection we present a couple of examples which showcase the good features of BNQN in practice. To save space, we refer the readers to extensive experiments in the arXiv version of this paper, as well as in the papers [33][11][29].

Here we present a couple of illustrating experimental results. Additional experiments, which are quite extensive, can be found in [33]. We use the python package numdifftools [36] to compute gradients and Hessian, since symbolic computation is not quite efficient. The experiments here are run on a small personal laptop. The unit for running time is seconds.

We note that for all examples here, the usual assumptions (e.g. the cost function is convex or its gradient is Lipschitz continuous) in the Optimization literature are not satisfied. Also, the Riemann xi function in the last example does not satisfy the usual assumptions in the root finding literature.

5.2.1. *One polynomial system in 2 complex variables.* Here, we explore the following system (see [13]):

$$\begin{aligned} f_1 := p_{10} + p_{11}x_1 + p_{12}x_2 + p_{13}x_2^2 + p_{14}x_2^3 &= 0, \\ f_2 := p_{20} + p_{21}x_1 + p_{22}x_2 + p_{23}x_2^2 + p_{24}x_2^3 &= 0, \end{aligned}$$

where $p_{10} = -13$, $p_{11} = 1$, $p_{12} = -2$, $p_{13} = 5$, $p_{14} = -1$, $p_{20} = -29$, $p_{21} = 1$, $p_{22} = -14$, $p_{23} = 1$ and $p_{24} = 1$. Here x_1, x_2 are **complex** variables, and hence the cost function f is $f = (|f_1|^2 + |f_2|^2)/2$. It means that we are finding roots to 4 equations in 4 real variables.

In this example, we will compare the performance of BNQN to that of: the usual Newton's method (for Optimization), BFGS, Adaptive Cubic Regularization as implemented in [37], as well as Random damping Newton's method [28].

Feature reported: We report on the number of iterates needed for an algorithm to achieve a prescribed threshold (more precisely, if $\|\nabla f(z_n)\| < \epsilon$ or $\|z_{n+1} - z_n\| < \epsilon$, where we fix $\epsilon = 10^{-10}$ in the experiments) or the number of iterates reaches a threshold (10000 iterations), the time (in seconds) it needs to run, the values of the cost function and its gradient at the end point (for a system of equations, we also report the value of each function in the system), whether the Hessian of the cost function at the end point has a negative eigenvalue.

Legends: We use the following abbreviation in the reports: "ACR" for Adaptive Cubic Regularization, "BFGS" for BFGS, "Newt" for Newton's method, "Newt-SE" for Newton's method applied directly to a system of equations F instead of to the associated cost function $f = \|F\|^2/2$, "RegN" for Regularized Newton's method, "B-RegN" for Regularized Newton's method with Backtracking line search incorporated, "LM" for Levenberg-Marquardt method, "B-LM" for Levenberg-Marquardt method with Backtracking line search incorporated, "B-NewQ" for Backtracking New Q-Newton's method, and "SB-NewQ" for Simplified Backtracking New Q-Newton's method.

For the first experiment, we choose the initial point

$$z_0 = [-9.12027123 + 0.001i, -3.7284278 - 0.001i].$$

In this case, it turns out that Newt-SE still converges to the same global minimum $[5, 4]$ as in the real case, Newton's method still converges to the same point as it did in the real case but now the point is no longer a local minimum (!), ACR and Iner are stopped before close enough to a critical point, while other methods converge to a different global minimum ($[13 - 14i, -1 - 1i]$). See Table 1 for details.

For the second experiment, we choose the initial point $z_0 = [15 - 0.000001i, -2 - 0.00002i]$, which is a small perturbation of the corresponding point in the real case. We obtain similar results like in Table 1, except that the performance of ACR is improved: after 2 iterates, it stopes (because of some ill-conditioned Hessian errors) at a point whose function value is $f \sim 4230$.

For the third experiment, we choose the initial point $z_0 = [-84.439842 - 0.000001i, -1.60847421 + 0.000001i]$, which is a small perturbation of the corresponding point in the real case. We obtain similar results like in Table 1, except that the performance of ACR is improved: after 12 iterates, it stopes (because of some ill-conditioned Hessian errors) at a point close to $\sim [13 + 14i, -1 + 1i]$ which is another global minimum of f.

	$\sharp N$	$f(z_N)$	$\|\nabla f(z_N)\|$	$\|f_1(z_N)\|$	$\|f_2(z_N)\|$	eigenS	Time
ACR	3	3e+93	5e+78	5e+46	5e+46	0.00	0.02
BFGS	27	1e-26	3e-12	1e-13	4e-14	6e-1	0.30
Newt	10	24.49	2e-14	4.9	4.9	-6e-1	0.10
Newt-SE	16	2e-28	$1e-14$	1e-14	1e-14	1.4	0.76
RegN	19	24.49	1e-14	4.9	4.9	4e-1	0.19
B-RegN	20	24.49	4e-14	4.9	4.9	4e-1	0.26
LM	31	6e-30	3e-14	2e-15	2e-15	6e-1	1.50
B-LM	33	9e-30	7e-14	2e-15	3e-15	6e-1	1.59
B-NewQ	31	5e-27	1e-13	6e-14	7e-14	6e-1	1.30
SB-NewQ	25	8e-27	1e-13	8e-14	1e-12	6e-1	1.18

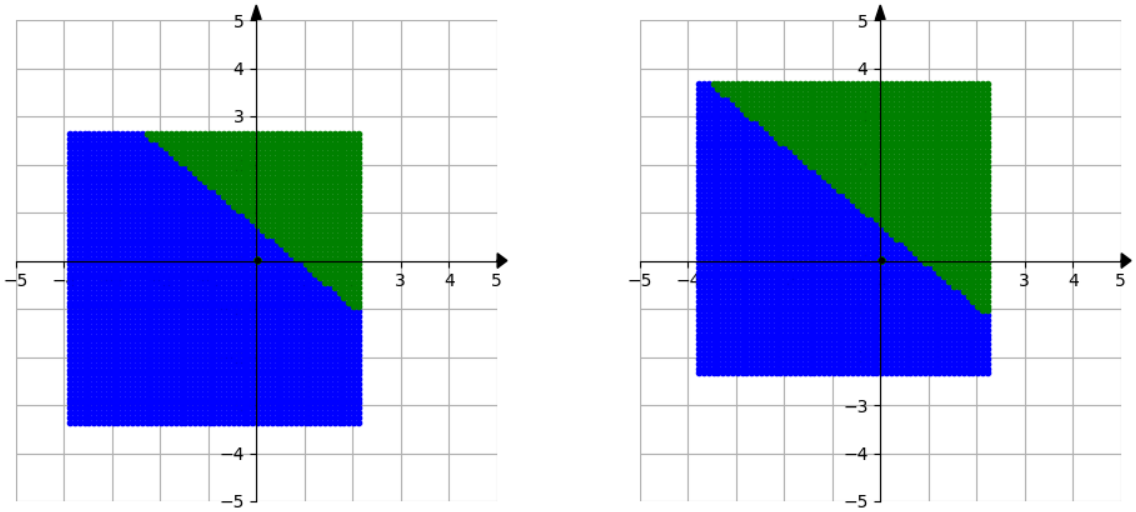
TABLE 1. Experimental results using various variants of Newton's method to solve the system of equations $f_1 := -13+1x_1+2x_2+5x_2^2-1x_2^3 = 0$ and $f_2 := -29+1x_1+14x_2+1x_2^2+1x_2^3 = 0$ in 2 complex variables x_1 and x_2. The cost function is $f = (|f_1|^2+|f_2|^2)/2$. The initial point is $z_0 = [-9.12027123+0.001i, -3.7284278-0.001i]$, at which $|f_1(z_0)| \sim 106$, $|f_2(z_0)| \sim 23.8$, and $f(z_0) \sim 5971$. **Legends**: $\sharp N$ (maximum 10000) =is the number of iterates the concerned algorithm runs before stopping when $\|\nabla f(z_N)\|$ or $\|z_N - z_{N-1}\|$ is smaller than a threshold or when some errors happen, z_N = the point constructed at the step N, "eigenS"=the smallest eigenvalue of the Hessian of f at z_N, and "Err" means that some errors (e.g. NAN, overflow, and so on) happened.

5.2.2. *Some pictures of basins of attraction.* Below, we display pictures of basins of attractions when applying Backtracking New Q-Newton's method for some polynomials of small degrees (2,3 4 and 5), and a transcendental function. We compare these with the basins obtained by the standard Newton's method and Backtracking Gradient Descent. It is observed that the basins for Backtracking New Q-Newton's method are much more regular than those for the standard Newton's method and Backtracking Gradient Descent, and hence the regularity observed does not come from properties of Backtracking line search alone. More examples like this, which explore the similarity between Backtracking New Q-Newton's method and Voronoi's diagrams, can be found in [11]. Some theoretical explanations for these experimental observations has been very recently obtained in [10].

The pictures below are created by choosing the initial point z_0 in a lattice $v + (0.1j, 0.1k)$, for $j, k \in [-30, 30]$, and v is a randomly chosen point in $[-1, 1] \times [-1.1]$.

For a polynomial of degree 2:

We consider the polynomial $P_2(z) = (z - z_1^*)(z - z_2^*)$, where $z_1^* = 0.5 - 0.2i$ and $z_2^* = 1 + 0.4i$. This polynomial satisfies the conditions for to apply the results in [33], and hence when we apply



(A) For Backtracking Gradient Descent. (B) For Backtracking New Q-Newton's method.

FIGURE 2. Basins of attraction when applying Backtracking New Q-Newton's method to the polynomial $P_2(z) = (z - z_1^*)(z - z_2^*)$. Blue: initial points z_0 for which the constructed sequence converges to z_1^*. Green: initial points z_0 for which the constructed sequence converges to z_2^*. It seems that the classical theorem of Schröder [24] (for Newton's method) is also valid for Backtracking Gradient Descent and Backtracking New Q-Newton's method. This is indeed now proven in [12].

Backtracking New Q-Newton's method for the function $f(x, y) = \frac{1}{2}|P(x + iy)|^2$, if the initial point z_0 is randomly chosen then the constructed sequence $\{z_n\}$ will converge, to either z_1^* or z_2^*. The basins of attraction (see Figure 2) seem to very much satisfy a classical result by Schröder [24] for Newton's method, see the next Conclusion subsection for details. Indeed, this is now proven rigorously in a recent paper [12].

For a polynomial of degree 3:

We consider the polynomial $P_3(z) = z^3 - 2z + 2$, which has 3 roots: $z_1^* \sim -1.76929$, $z_2^* \sim 0.884646 - 0.589743i$ and $z_3^* = 0.884646 + 0.589743i$. It is known [22], that when applying the standard Newton's method to this polynomial, then the basins of attraction are fractal. Moreover, there are sets of positive Lebesgue measure where Newton's method applied to an initial point in these sets will not converge to any of the roots.

In contrast, since this polynomial satisfies the conditions to apply the results in [33], Backtracking New Q-Newton's method applied to a random initial point z_0 will always converge to one of the 3 roots. Moreover, the basins of attraction do not seem to have a fractal structure.

Basins of attraction for Backtracking Gradient Descent seem to be more regular than that for Newton's method, but less regular than that for Backtracking New Q-Newton's method.

For a polynomial of degree 4:

We consider the polynomial $P_4(z) = (z^2 + 1)(z - 2.3)(z + 2.3)$, which has 4 roots: $z_1^* = 2.3$, $z_2^* = -2.3$, $z_3^* = i$ and $z_4 = -i$. When applying the standard Newton's method to this polynomial, then the basins of attraction are fractal. Moreover, there are sets of positive Lebesgue measure where Newton's method applied to an initial point in these sets will not converge to any of the roots.

In contrast, since this polynomial satisfies the conditions to apply the results in [33], Backtracking New Q-Newton's method applied to a random initial point z_0 will always converge to one of the 3 roots. Moreover, the basins of attraction do not seem to have a fractal structure.

Basins of attraction for Backtracking Gradient Descent seem to be more regular than that for Newton's method, but less regular than that for Backtracking New Q-Newton's method.

5.2.3. *Some experiments with the Riemann xi function.* Here we present some experiments, from [29], when applying Backtracking New Q-Newton's method to the Riemann xi function. We need to incorporate mpmath [21] for high precision computations. We refer the readers to [29] for more contexts and experiments.

It is found experimentally that by starting with initial points of the form $z_0 = (0, y_0)$, where $|y_0 - T| \leq 1$, BNQN can find all roots in the domain $T \leq \mathcal{I}(z) \leq T + 1$ of the Riemann xi function, for very large values of T such as $T = 10^9$ and $T = 10^{10}$. The motivation for looking at initial points of this special type is twofolds. First, these points lie on the boundary of the critical trip $0 \leq \mathcal{R}(z) \leq 1$, where we know that all zeros of the Riemann xi function must belong. Second, the similarity (see [11]) between BNQN and Voronoi's diagrams suggest that we indeed should be able to find all zeros of ξ by starting with these initial points, see [29] for a detailed discussion.

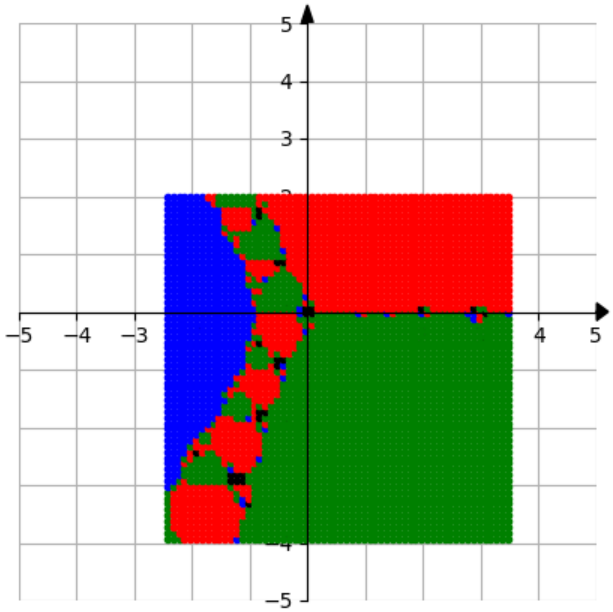
It is found experimentally that the remarkable phenomenon observed for the basins of attractions in the polynomial examples in the previous subsubsection seems also appears in very highly transcendental functions like the Riemann xi function.

We compare the basin of attractions for various methods (Newton's method, Random Relaxed Newton's method, and BNQN) in finding the first 8 roots of the Riemann xi function. Here the roots are:

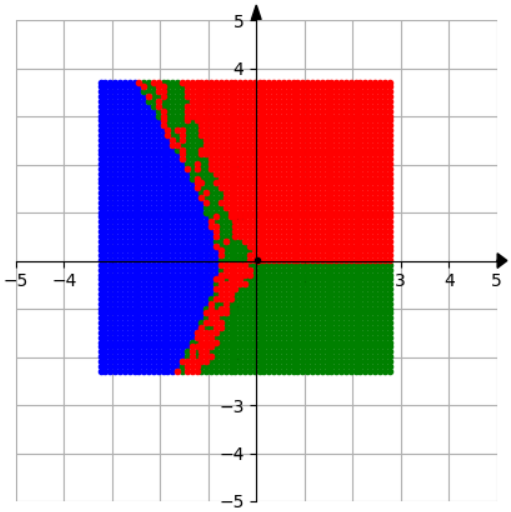
Root 1 $\sim 0.5 + 14.13472514173i$, with points in the basin of attraction having green colour.

Root 2 $\sim 0.5 - 14.13472514173i$, with points in the basin of attraction having yellow colour.

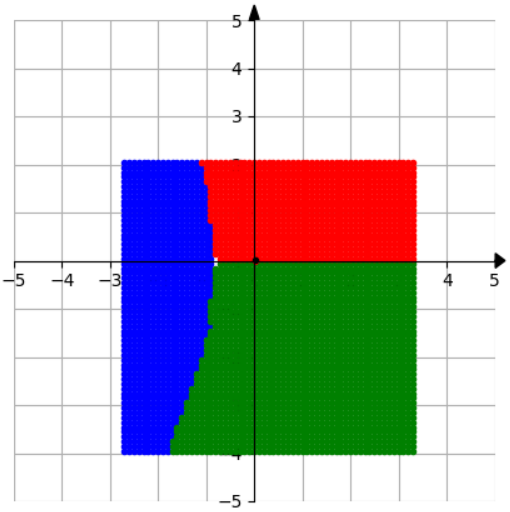
Root 3 $\sim 0.5 + 21.02203963877i$, with points in the basin of attraction having blue colour.



(A) For Newton's method.

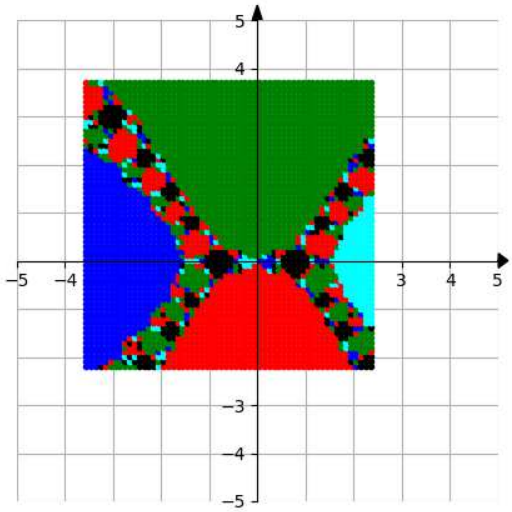


(B) For Backtracking Gradient Descent.

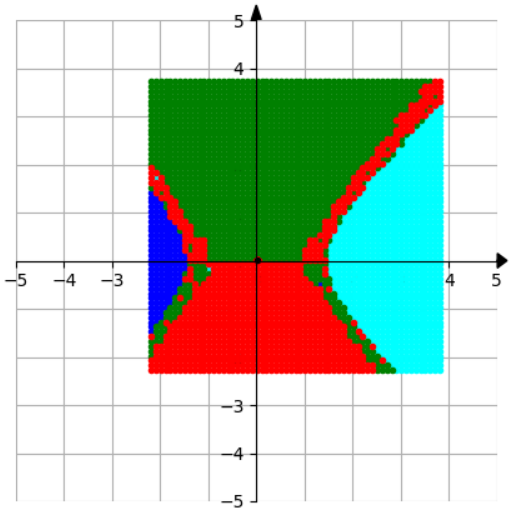


(C) For Backtracking New Q-Newton's method.

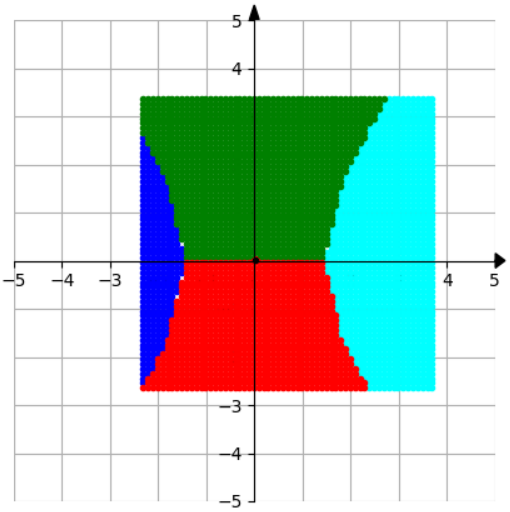
FIGURE 3. Basins of attraction for the polynomial $P_3(z) = z^3 - 2z + 2$. Blue: initial points z_0 for which the constructed sequence converges to z_1^*. Green: similar points for the root z_2^*. Red: similar points for the root z_3^*. Black: other points.



(A) For Newton's method.



(B) For Backtracking Gradient Descent.



(C) For Backtracking New Q-Newton's method.

FIGURE 4. Basins of attraction for the polynomial $P_4(z) = (z^2+1)(z-2.3)(z+2.3)$. Blue: initial points z_0 for which the constructed sequence converges to z_1^*. Cyan: similar points for the root z_2^*. Green: similar points for the root z_3^*.Red: similar points for the root z_4^*. Black: other points.

Root $4 \sim 0.5 - 21.02203963877i$, with points in the basin of attraction having red colour.

Root $5 \sim 0.5 + 25.01085758014i$, with points in the basin of attraction having pink colour.

Root $6 \sim 0.5 - 25.01085758014i$, with points in the basin of attraction having cyan colour.

Root $7 \sim 0.5 + 30.42487612585i$, with points in the basin of attraction having orange colour.

Root $8 \sim 0.5 - 30.42487612585i$, with points in the basin of attraction having purple colour.

All other points in the grid have black colour. Among other options, they may represent initial points where the sequence converge to another root or to infinity or to strange attractors. By the results in [33], BNQN has no strange attractors, but it is unknown whether Newton's method or Random Relaxed Newton's method can have.

In this experiment, we choose initial points in a grid of the size 250×250 points. We also rescale the y-axis by a factor of 0.1. For example, the point $0.1 + 1.1i$ in Figure corresponds to the point $0.1 + 11i$ in the usual coordinate system.

See Figure 5 for detail.

5.3. **An interesting example of systems of equations in complex dimension 2.** Here we first present an experiment by Takayuki Watanabe, and then provide some observations and explanations for the behaviour of BNQN found in this experiment. The experiment involves finding roots of the 2-dim polynomial map $h : \mathbf{C}^2 \to \mathbf{C}^2$ given by $h(z_1, z_2) = (z_1 + z_1 z_2, z_2^2 - z_1^2)$ which has a multiple root at $(0, 0)$. The interesting fact is that while it is known from [39] [9] that here $(0, 0)$ is not an attractor for the dynamics of Newton's method, and also seems not an attractor for Random Relaxed Newton's method, it seems that $(0, 0)$ is an attractor for the dynamics of Newton's method. We thank Takayuki Watanabe for permission to use the Figures 6 and 7.

It is apparent from Figure 7 that the basins of attraction of the 3 roots of the system $h(z_1, z_2) = 0$, under the dynamics of BNQN, has full measure. We now prove Theorem 1.3, based on theoretical results obtained so far in this paper, which indeed confirms this observation.

Proof of Theorem 1.3. We recall the function F in 4 real variables y_1, y_2, y_3, y_4 is defined by

$$F(y_1, y_2, y_3, y_4) = ||h(z_1, z_2)||^2 = |z_1(1 + z_2)|^2 + |z_2^2 - z_1^2|^2,$$

for $z_1 = y_1 + iy_2$ and $z_2 = y_3 + iy_4$.

We need to show the existence of an exception set $\mathcal{E} \subset \mathbf{R}^4$ for which if the initial point x_0 is outside $\mathcal{E}$, then the sequence constructed by BNQN applied to the function F converges to one of the 3 roots of the system $h(z_1, z_2) = 0$. To this end, we will apply Theorem 1.1, and hence will need to check the following properties:

The function F has compact sublevels: We need to show that there cannot be a sequence $(z_1^{(n)}, z_2^{(n)}) \to \infty$ for which $|z_1^{(n)}(1 + z_2^{(n)})|^2 + |(z_2^{(n)})^2 - (z_1^{(n)})^2|^2$ is bounded. We can assume that $|z_1^{(n)}|$ converges to ∞. Indeed, assume otherwise. Then the sequence $|z_1^{(n)}(1 + z_2^{(n)})|^2$ is bounded, which implies that either $z_1^{(n)}$ or $z_2^{(n)}$ must be bounded. Also, the sequence $|(z_2^{(n)})^2 - (z_1^{(n)})^2|^2$ is

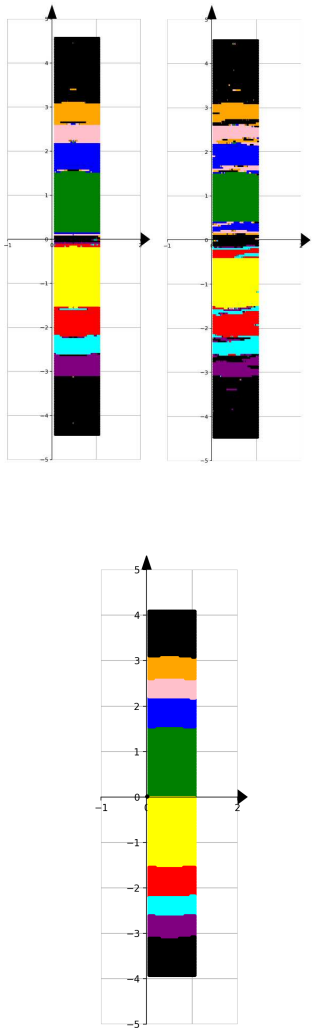


FIGURE 5. Basins of attraction when applying BNQN to find roots of $\xi(z)$, specially the first 8 roots. The y-axis is rescaled by 0.1, hence $y = 1$ in the pictures correspond to $y = 10$ in reality. Pictures are referenced to from top to bottom, from left to right. Row 1: left picture is Newton's method, right picture is Random Relaxed Newton's method. Row 2: BNQN.

One can see that pictures for Newton's method and Random Relaxed Newton's method display fractal features, while BNQN picture is very regular and similar to Voronoi's diagrams.

bounded, which implies that $|z_2^{(n)}|$ must also converge to ∞. The above facts are in contradiction. Therefore, the function F has compact sublevels.

The function F has at most countably many critical points: Indeed, we will show that F has precisely 5 critical points. We first recall the following useful fact which helps to compute the gradient of F: Let $w = u + iv$ be a complex variable, where u, v are real coordinates. If $g(w)$ is a holomorphic function, and $G(u, v) = |g(u + iv)|^2$, then

$$\begin{aligned}\frac{\partial G}{\partial u} &= \frac{\partial G}{\partial w} + \frac{\partial G}{\partial \overline{w}} = g'(u+iv)\overline{g(u+iv)} + g(u+iv)\overline{g'(u+iv)},\\ \frac{\partial G}{\partial v} &= i\frac{\partial G}{\partial w} - i\frac{\partial G}{\partial \overline{w}} = ig'(u+iv)\overline{g(u+iv)} - ig(u+iv)\overline{g'(u+iv)}.\end{aligned}$$

From this, it is ready to see (recall that $z_1 = y_1 + iy_2$ and $z_2 = y_3 + iy_4$):

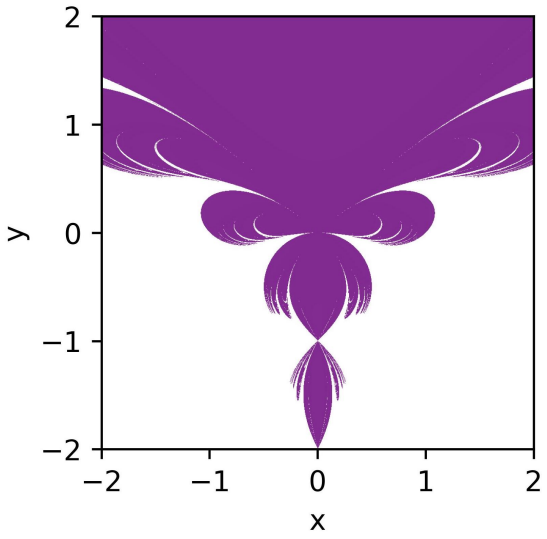
$$\begin{aligned}\frac{\partial F}{\partial y_1} &= (1+z_2)\overline{z_1(1+z_2)} + z_1(1+z_2)\overline{(1+z_2)} + 2z_1\overline{(z_1^2 - z_2^2)} + (z_1^2 - z_2^2)\overline{2z_1},\\ \frac{\partial F}{\partial y_2} &= i((1+z_2)\overline{z_1(1+z_2)} - z_1(1+z_2)\overline{(1+z_2)} + 2z_1\overline{(z_1^2 - z_2^2)} - (z_1^2 - z_2^2)\overline{2z_1}),\\ \frac{\partial F}{\partial y_3} &= z_1\overline{z_1(1+z_2)} + z_1(1+z_2)\overline{z_1} - 2z_2\overline{(z_1^2 - z_2^2)} - (z_1^2 - z_2^2)\overline{2z_2},\\ \frac{\partial F}{\partial y_4} &= i(z_1\overline{z_1(1+z_2)} - z_1(1+z_2)\overline{z_1} - 2z_2\overline{(z_1^2 - z_2^2)} + (z_1^2 - z_2^2)\overline{2z_2}),.\end{aligned}$$

We then can check (for example, with the help of BNQN) that the critical points of F are: besides the 3 roots $(0, 0)$, $(-1, -1)$, $(1, -1)$, also 2 other points $(0.4939807039..., 0, -0.5773502691896..., 0)$ and $(-0.4939807039..., 0, -0.5773502691896..., 0)$.

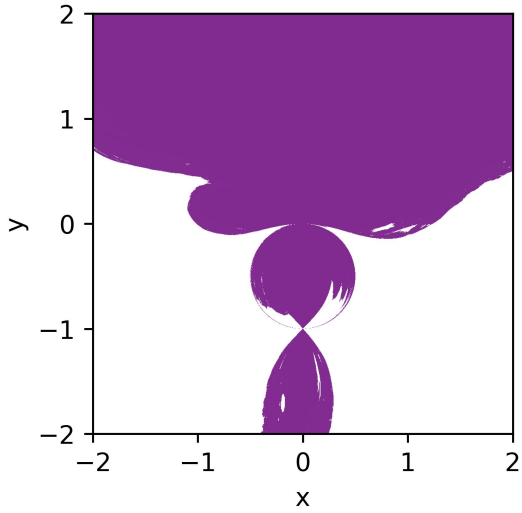
Except roots of h, the critical points of F are non-degenerate saddle points: As above, the critical points of F which are not roots of h are 2 points $(0.4939807039..., 0, -0.5773502691896..., 0)$ and $(-0.4939807039..., 0, -0.5773502691896..., 0)$. The eigenvalues of the Hessian matrix of F at $(0.4939807039..., 0, -0.5773502691896..., 0)$ are $-0.240..., 0.224..., 2.507..., 2.972...$. The eigenvalues of the Hessian matrix of F at $(-0.4939807039..., 0, -0.5773502691896..., 0)$ are also $-0.240...$, $0.224...$, $2.507...$, $2.972...$. Therefore, both these 2 points are non-degenerate saddle points of F.

Complete the proof of Theorem 1.3: Let $x_0 \in \mathbf{R}^4$ be any initial point, and let x_n be the sequence constructed by BNQN applied to the function F. Since F has compact sublevels and has only a finite number of critical points, by Theorem 1.1 the sequence x_n will converge to one of the 5 critical points $(0, 0)$, $(-1, -1)$, $(1, -1)$, $(0.4939807039..., 0, -0.5773502691896..., 0)$ and $(-0.4939807039..., 0, -0.5773502691896..., 0)$. Since the 2 points $(0.4939807039..., 0, -0.5773502691896..., 0)$ and $(-0.4939807039..., 0, -0.5773502691896..., 0)$ are non-degenerate saddle points, by Theorem 1.1 again, there is an exceptional set $\mathcal{E} \subset \mathbf{R}^4$ so that if $x_0 \notin \mathcal{E}$, then x_n converges to one of the 3 points $(0, 0)$, $(-1, -1)$, and $(1, -1)$, which are roots of the system $h(z_1, z_2) = 0$.

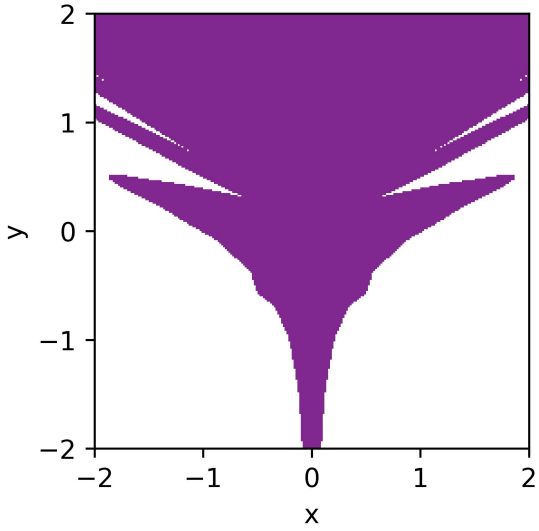
□



(A) For Newton's method.



(B) For Random Relaxed Newton's method.

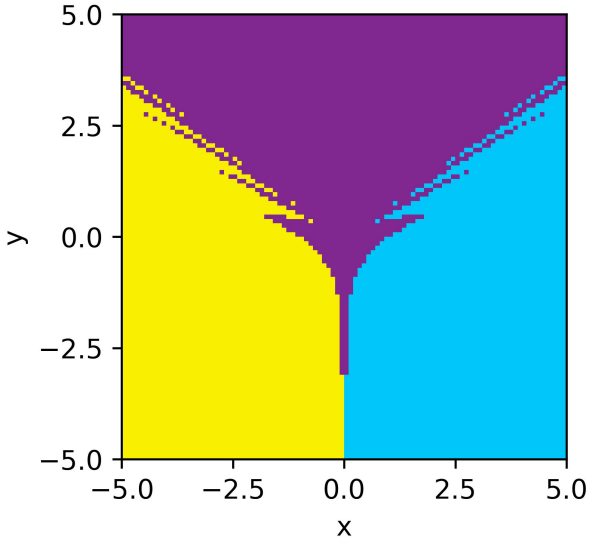


(C) For BNQN.

FIGURE 6. A slice of basins of attraction of the root $(0, 0)$ for the system of equation $z_1 + z_1 z_2 = z_2^2 - z_1^2 = 0$ in 2 complex variables, by different methods.

6. Appendix: some further developments

In this appendix, we present some further extensions. We can extend the algorithms above to the setting where at each iterate one projects to subspaces different from eigenspaces of the Hessian of f, or more generally working with a symmetric square matrix having no relations to the Hessian



(A) For BNQN.

FIGURE 7. A slice of basins of attraction of the 3 roots $(0,0)$, $(-1,-1)$ and $(-1,1)$ for the system of equation $z_1 + z_1 z_2 = z_2^2 - z_1^2 = 0$ in 2 complex variables, by BNQN. It seems that the union of these 3 basins have full measure.

of f. This way, we can unify first and second order methods, and among these algorithms some have the flavour of quasi-Newton's methods.

Algorithm 6: Backtracking New Q-Newton's method G

Result: Find a minimum of $f : \mathbb{R}^m \to \mathbb{R}$

Given: $\{\delta_0, \delta_1, \dots, \delta_m\} \subset \mathbb{R}$ (chosen **randomly**), $0 < \tau$, $q \geq 1$ and $0 < \gamma_0 < 1$;

Initialization: $x_0 \in \mathbb{R}^m$;

$\kappa := \frac{1}{2} \min_{i \neq j} |\delta_i - \delta_j|$;

for $k = 0, 1, 2 \dots$ **do**

  $j = 0$

  if $\|\nabla f(x_k)\| \neq 0$ **then**

    Choose a symmetric real matrix B_k;

    while $minsp(B_k + \delta_j \|\nabla f(x_k)\|^\tau Id) < \kappa \|\nabla f(x_k)\|^\tau$ **do**

      $j = j + 1$

    end

  end

  $A_k := B_k + \delta_j \|\nabla f(x_k)\|^\tau Id$

  $e_{1,k}, \dots, e_{m,k}$ an appropriately chosen orthonormal basis for $\mathbb{R}^m$

  $w_k := \sum_{i=1}^m \frac{\langle \nabla f(x_k), e_{i,k} \rangle}{(\sum_{j=1}^m |\langle A_k . e_{i,k}, e_{j,k} \rangle|^q)^{1/q}} e_{i,k}$

  $\widehat{w_k} := w_k / \max\{1, \|w_k\|\}$

  $\gamma := \gamma_0$

  if $\|\nabla f(x_k)\| \neq 0$ **then**

    while $f(x_k - \gamma \widehat{w_k}) - f(x_k) > -\gamma \langle \widehat{w_k}, \nabla f(x_k) \rangle / 3$ **do**

      $\gamma = \gamma / 3$

    end

  end

  $x_{k+1} := x_k - \gamma \widehat{w_k}$

end

First, we investigate some special cases of Backtracking New Q-Newton's method G where $B_k := \nabla^2 f(x_k)$, and where $q = 2$ (in which case $(\sum_{j=1}^m |\langle A_k . e_{i,k}, e_{j,k} \rangle|^q)^{1/q} = \|A_k e_{i,k}\|$ the usual Euclidean norm of the vector $A_k e_{i,k}$). We fix $e_1, \dots, e_m$ an orthonormal basis for $\mathbb{R}^m$.

Backtracking New Q-Newton's method: We choose $e_{1,k}, \dots, e_{m,k}$ eigenvectors of A_k. Hence, if λ_i is the corresponding eigenvalue of $e_{i,k}$ of A_k, then

$$w_k = \sum_{i=1}^m \frac{\langle \nabla f(x_k), e_{i,k} \rangle}{|\lambda_i|} e_{i,k}.$$

Backtracking New Q-Newton's method Backtracking G1: We choose $e_1(x), \dots, e_m(x)$'s to be eigenvectors of $\nabla^2 f(x)$ if $minsp(\nabla^2 f(x)) \geq \|\nabla f(x)\|^{1/2}$, otherwise $e_1(x)$, ..., $e_m(x)$ are e_1, ... , e_m. This is a combination between New Q-Newton's method Backtracking and New Q-Newton's method Backtracking G2 (see next). As such, it has fast rate of convergence as New

Q-Newton's method Backtracking, while also has good convergence guarantee as New Q-Newton's method Backtracking G3.

New Q-Newton's method Backtracking G2: We choose $e_{i,k} = e_i$ for all $i = 1, \dots, m$. This version is probably the less computationally expensive, and hence has some flavours of quasi-Newton's methods. On the other hand, it has quite good theoretical guarantees.

Remark 6.1. Usually, the choice of $\tau > 1$ will assure that the associated dynamical system is C^1, which helps more with avoidance of saddle points.

On the other hand, the choice of $0 < \tau < 1$ helps more with theoretical convergence guarantees.

However, in experiments, we found that there is no real difference between the choices $\tau > 1$ or $0 < \tau < 1$ or $\tau = 1$.

As in Section 2, the following result can be used to establishing that Backtracking New Q-Newton's method G can avoid saddle points. (The convergence of it can be dealt similarly to that for Backtracking New Q-Newton's method and other methods in Section 2.)

Theorem 6.2. *Let $f(x) = \|F(x)\|^2$ be as above. Let $w(x)$ be chosen as in the algorithm Backtracking New Q-Newton's method G. Assume that $q \geq 1$ is so that its Hölder's conjugate p (i.e. the number $p \geq 1$ so that $p^{-1} + q^{-1} = 1$, the value $+\infty$ is allowed) satisfies $m^{1/p} < 4/3$. Assume moreover that for x close enough to a generalised saddle point x^* of f, we have $B(x) - \nabla^2 f(x) + O(\|\nabla f(x)\|)$ is positive definite. Then, for all x close enough to x^* we have $\gamma(x) = 1$.*

Proof. Since $\|F(x)\|$ is bounded away from 0 near a generalised saddle point x^* of f, we have that $\|w(x)\| \sim \|\nabla f(x)\|$ near x^*. Also, $\langle w(x), \nabla f(x)\rangle \sim \|w(x)\|^2$. By Taylor's expansion we have

$$f(x - w(x)) - f(x) = -\langle w(x), \nabla f(x)\rangle + \frac{1}{2}\langle \nabla^2 f(x) w(x), w(x)\rangle + O(\|w(x)\|^3). \tag{2}$$

Denote $a_i = \langle \nabla f(x), e_i(x)\rangle$, $b_{i,j} = \langle A(x) e_i(x), e_j(x)\rangle$, and $B_i = (\sum_{j=1}^m |b_{i,j}|^q)^{1/q}$. We have

$$\langle w(x), \nabla f(x)\rangle = \sum_{i=1}^m a_i^2/B_i.$$

Now we bound from above the second summand in the RHS in Equation (2), note that all $\delta_j > 0$:

$$\begin{aligned}
\langle \nabla^2 f(x) w(x), w(x)\rangle &\leq \langle [B(x) + O(\|w(x)\|].w(x), w(x)\rangle \\
&\leq \langle A(x) w(x), w(x)\rangle + O(\|w(x)\|^3) \\
&= < \sum_{i=1}^m a_i A(x) e_i(x)/B_i, \sum_{j=1}^m a_j e_j(x)/B_j > \\
&= \sum_{i=1}^m \sum_{j=1}^m a_i a_j b_{i,j}/(B_i B_j)
\end{aligned}$$

Now, by Cauchy-Schwartz inequality, we have

$$|a_i a_j b_{i,j}|/(B_i B_j) \leq \frac{a_i^2|b_{i,j}|}{2B_i^2} + \frac{a_j^2|b_{i,j}|}{2B_j^2}$$

By Hölder's inequality, we obtain for all i:

$$\sum_{j=1}^{m} |b_{i,j}| \leq m^{1/p} B_i$$

This implies, from the assumption on p, that

$$f(x - w(x)) - f(x) < -\frac{1}{3}\langle w(x), \nabla f(x)\rangle$$

for all x close enough to x^*. Thus we can choose $\gamma(x) = 1$ for all such points. □

References

[1] L. Armijo, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific J. Math. 16 (1966), no. 1, 1–3.

[2] D. D'Acunto and K. Kurdyka, *Explicit bounds for the Lojasiewicz exponent in the gradient inequality for polynomials*, Ann. Pol. Math. 87 (2005), 51–61.

[3] M. Ahookhosh, F. J. Aragón, R. M. T. Fleming and P. T. Vuong, *Local convergence of the Levenberg-Marquardt method under Hölder metric subregularity*, Advances in Computational Mathematics, 1–36.

[4] D. J. Bates, J. D. Hauenstein, A. J. Sommese and C. W. Wampler, *Bertini: Software for numerical Algebraic Geometry*, Available at bertini.nd.edu with permanent doi: dx.doi.org/10.7274/R0H41PB5.

[5] C. Beltrán and L. M. Pardo, *Smale's 17th problem: average polynomial time to compute affine and projective solutions*, Journal of the American Mathematical Society 22 (2): 363–385.

[6] W. Bergweiler, Iteration of meromorphic functions, Bulletin of the American Mathematical Society, vol 29, number 2, October 1993, pp. 151–188 (1993)

[7] A. J. Bray and and D. S. Dean, *Statistics of critical points of gaussian fields on large-dimensional spaces*, Physics Review Letter, 98, 150201.

[8] P. Bürgisser and F. Cucker, *On a problem posed by Steve Smale*, Annals of Mathematics 174 (3): 1785–1836.

[9] T. C. Dinh, R. Dujardin and N. Sibony, On the dynamics near infinity of some polynomial mappings in $\mathbf{C}^2$, Mathematische Annalen vol 333, pp.703–739, 2005.

[10] J. E. Fornæss, M. Hu, and T. T. Truong, Backtracking New Q-Newton's method for finding roots of meromorphic functions in 1 complex variable: Global convergence, and local stable/unstable curves, arXiv:2412.02476.

[11] J. E. Fornæss, M. Hu, T. T. Truong and T. Watanabe, Backtracking New Q-Newton's method, Newton's flow, Voronoi's diagram and Stochastic root finding, Complex analysis and Operator theory, vol 18, article number 112, 2024.

[12] J. E. Fornæss, M. Hu, T. T. Truong and T. Watanabe, Backtracking New Q-Newton's method, Schr̈oder's theorem, and Linear conjugacy, arXiv:2312.12166.

[13] F. Freudenstein and B. Roth, *Numerical solution of systems of nonlinear equations*, Journal of the ACM, volume 10, issue 4, October 1963, pp. 550–556.

[14] A. F. Izmailov and M. V. Solodov, Newton-type methods: a broader view, Journal of Optimization Theory and Applications, vol 164, pp. 577–620, 2015.

[15] A. F. Izmailov and M. V. Solodov, Newton-type methods for Optimization and Variational problems, Springer series in Operations Research and Financial engineering, Springer International Publishing, Switzerland (2014).

[16] K. Levenberg, *A method for the solution of certain non-linear problems in least squares*, Quarterly of Applied mathematics 2 (2): 164–168.

[17] S. Lojasiewicz, *Sur les trajectoires du gradient d'une fonction analytique*, Seminari di Geometria, Bologna 1982/1983, Universita' degli studi di Bologna, Bologna (1984), pp. 115–117.

[18] S. Lojasiewicz, *Ensembles semi-analytiques*, preprint IHES, 1965.

[19] C. Macmullen, *Families of rational maps and iterative root-finding algorithms*, Ann. of Math. (2), 125 (1987), no 3, 467–493.

[20] D. Marquardt, *An algorithm for least-squares estimation of nonlinear parameters*, SIAM Journal on Applied Mathematics, 11 (2): 431–441.

[21] http://mpmath.org/

[22] Wikipedia page on Newton fractal https://en.wikipedia.org/wiki/Newton_fractal

[23] S. Paternain, A. Mokhtari and A. Ribeiro, *A Newton-based method for nonconvex optimization with fast evasion of saddle points*, SIAM Journal on Optimization, vol 29, issue 1, 10.1137/17M1150116.

[24] E. Schröder, Ueber iterirte Functionen, Math. Ann., **3**, 296-322 (1871)

[25] C. Shen, X. Chen and Y. Liang, *A regularized Newton method for degenerate unconstrained optimization problems*, Optimization Letters (2012), 6:1913–1933.

[26] M. Shub and S. Smale, *Complexity of Bezóut's theorem. I. Geometric aspects.*, Journal of the American Mathematical Society, 6(2):459–501, 1993.

[27] S. Smale, *Mathematical problems for the next century*, in Mathematics: frontiers and perspectives, pp. 271–294, AMS, Providence, RI, 2000.

[28] H. Sumi, *Negativity of Lyapunov exponents and convergence of generic random polynomial dynamical systems and random relaxed Newton's method*, Communications in Mathematical Physics 384, 1513–1583 (2021).

[29] T. Q. Tran and T. T. Truong, *The Riemann hypothesis and dynamics of Backtracking New Q-Newton's method*, arXiv:2405.05834.

[30] T. T. Truong, Some iterative algorithms on Riemannian manifolds and Banach spaces with good global convergence guarantee, arXiv:2505.22180.

[31] T. T. Truong and H.-T. Nguyen, *Backtracking gradient descent method and some applications to Large scale optimisation. Part 1: Theory*, Minimax Theory and its Applications 7 (2022), no 1, 79–108. This is the more theoretical part of arXiv: 1808.05160, with some additional experiemental results.

[32] T. T. Truong and H.-T. Nguyen, *Backtracking gradient descent method and some applications in Large scale optimisation. Part 2: Algorithms and experiments*, Applied Mathematics and Optimization 84 (2021), 2557–2586. This is the more applied part of arXiv: 1808.05160, in combination with arXiv:2001.02005 and arXiv:2007.03618.

[33] T. T. Truong, T. D. To, H.-T. Nguyen, T. H. Nguyen, H. P. Nguyen and M. Helmy, *A fast and simple modification of Newton's method avoiding saddle points saddle points*, Journal of Optimization Theory and Applications, vol 199, pp. 805–830, 2023.

[34] K. Ueda and N. Yamashita, *A regularized Newton method without line search for unconstrained optimization*, Computational Optimization and Applications (2014), 59:321–351.

[35] K. Ueda and N. Yamashita, *Convergence properties of the Regularized Newton method for the unconstrained nonconvex optimization*, Applied Mathematics and Optimization 62, 27–46 (2010).

[36] GitHub link for python's package numdifftools https://github.com/pbrod/numdifftools

[37] GitHub link for Adaptive cubic regularization for Newton's method: https://github.com/cjones6/cubic_reg . Discussion on implementation of this method: https://www.semanticscholar.org/paper/Implementing-the-cubic-and-adaptive-cubic-Jones/5bd648d2840a9753193b39e773ae4192bd43ddd7

[38] GitHub link for Python source codes for Backtracking New Q-Newton's method and variations: https://github.com/tuyenttMathOslo/New-Q-Newton-s-method-Backtracking and https://github.com/tuyenttMathOslo/NewQNewtonMethodBacktrackingForSystemEquations

[39] Y. Yamagishi, On the local convergence of Newton's method to a multiple root, J. Math. Soc. Japan 55, 897–908, 2003.

Department of Mathematics, University of Oslo, Blindern 0851 Oslo, Norway
Email address: tuyentt@math.uio.no