Backtracking New Q-Newton's method: finding roots and optimization

Tuyen Trung Truong

Matematikk Institut, Universitetet i Oslo
Supported by Research Council of Norway

Complex Analysis, Geometry and Dynamics - Portoroz conference, June 2023

Blank page

- This page is intended to be blank.

Table of content

Table of content

Table of content

Table of content (cont. 2)

Table of content (cont. 2)

Table of content (cont.3)

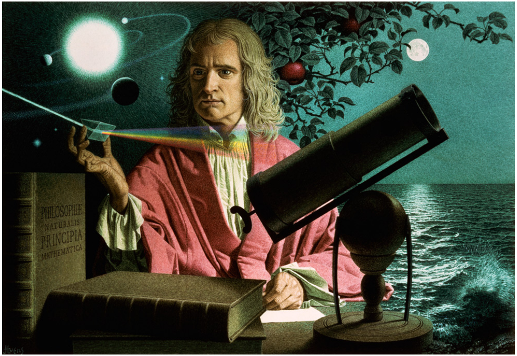
Table of content (cont.3)

Table of content (cont.3)

Blank page

- This page is intended to be blank.

§1: Some motivating questions



Section 1.1: Finding roots of a polynomial and a transcendental/meromorphic function in 1 complex variable

- Finding roots of a polynomial in 1 complex variable:

  Huge interest in mathematics in the middle age.

  Theorem (Abel-Ruffini, Galois' theory): For a general polynomial of degree ≥ 5, there is no solution in radicals. A simple example is $x^5 - x - 1$.

  Hence, in general **can only find approximate roots**. Usually use an iterative method $\Rightarrow$ Dynamics.

Section 1.1: Finding roots of a polynomial and a transcendental/meromorphic function in 1 complex variable

- Finding roots of a polynomial in 1 complex variable:

  Huge interest in mathematics in the middle age.

  Theorem (Abel-Ruffini, Galois' theory): For a general polynomial of degree ≥ 5, there is no solution in radicals. A simple example is $x^5 - x - 1$.

  Hence, in general **can only find approximate roots**. Usually use an iterative method $\Rightarrow$ Dynamics.

- Finding roots of a transcendental function/meromorphic function in 1 complex variable:

  Many interesting and useful special functions are transcendental or meromorphic.

  E.g.: Airy, elliptic, Bessel, and Riemann zeta functions.

  Less systematic work on iterative methods to find roots of these functions.

Section 1.2: Finding roots of systems of equations

- Finding roots of systems of equations is useful both in theory and applications.

- Finding roots of systems of equations is useful both in theory and applications.
- In theory:

  Find periodic points of dynamics of maps, find equilibrium measures, find if some exotic maps may exist (e.g. if there is a polynomial map of a certain degree not satisfying the Jacobian conjecture), check if there is a birational map/biregular map of a given bounded degree between 2 algebraic varieties (T., Beitrage zur Algebra und Geometrie).

Section 1.2: Finding roots of systems of equations

- Finding roots of systems of equations is useful both in theory and applications.
- In theory:

  Find periodic points of dynamics of maps, find equilibrium measures, find if some exotic maps may exist (e.g. if there is a polynomial map of a certain degree not satisfying the Jacobian conjecture), check if there is a birational map/biregular map of a given bounded degree between 2 algebraic varieties (T., Beitrage zur Algebra und Geometrie).
- In applications:

  Find critical points of a function (helpful to find minima/maxima), find orbits/solutions to physical systems (e.g. robots).

- Recasting finding roots of a system in an optimization problem.

  Consider: $F(x) = 0$, where $x \in \mathbf{R}^m$,
  $F = (f_1, \ldots, f_N) : \mathbf{R}^m \to \mathbf{R}^N$ (N equations in m variables).
  Define $g(x) = \frac{1}{2}||F(x)||^2 = \frac{1}{2}(f_1(x)^2 + \ldots + f_N(x)^2)$.
  $g(x) \geq 0$, and if $F(x) = 0$ then $g(x) = 0$ (hence, global minimum of g).

Section 1.3: A more general approach: optimization

- Recasting finding roots of a system in an optimization problem.

  Consider: $F(x) = 0$, where $x \in \mathbf{R}^m$,
  $F = (f_1, \ldots, f_N) : \mathbf{R}^m \to \mathbf{R}^N$ (N equations in m variables).
  Define $g(x) = \frac{1}{2}||F(x)||^2 = \frac{1}{2}(f_1(x)^2 + \ldots + f_N(x)^2)$.
  $g(x) \geq 0$, and if $F(x) = 0$ then $g(x) = 0$ (hence, global minimum of g).
- For complex or quaternionic variables, we can reduce to real variables (increasing dimensions).

Section 1.3: A more general approach: optimization

- Recasting finding roots of a system in an optimization problem.

  Consider: $F(x) = 0$, where $x \in \mathbf{R}^m$,
  $F = (f_1, \ldots, f_N) : \mathbf{R}^m \to \mathbf{R}^N$ (N equations in m variables).
  Define $g(x) = \frac{1}{2}||F(x)||^2 = \frac{1}{2}(f_1(x)^2 + \ldots + f_N(x)^2)$.
  $g(x) \geq 0$, and if $F(x) = 0$ then $g(x) = 0$ (hence, global minimum of g).
- For complex or quaternionic variables, we can reduce to real variables (increasing dimensions).
- Hence, methods for optimization can be used to find roots of systems of equations.

  In particular, methods for the Least Square Fit problem in classical statistics. (e.g. Linear Regression)

- **Theorem 1:** (arXiv:2006.01512, under revision in a journal, T., T.-D. To, H.-T. Nguyen, T. H. Nguyen, H. P. Nguyen, M. Helmy) New Q-Newton's method is similar to Newton's method near non-degenerate local minima, hence in particular has quadratic rate of convergence near these points . It has local Stable-central manifolds near saddle points, and (+ arXiv:2008.11091) hence can globally avoid saddle points. If a sequence constructed by this method **converges**, then the limit point is a critical point of the function.

- **Theorem 1:** (arXiv:2006.01512, under revision in a journal, T., T.-D. To, H.-T. Nguyen, T. H. Nguyen, H. P. Nguyen, M. Helmy) New Q-Newton's method is similar to Newton's method near non-degenerate local minima, hence in particular has quadratic rate of convergence near these points . It has local Stable-central manifolds near saddle points, and (+ arXiv:2008.11091) hence can globally avoid saddle points. If a sequence constructed by this method **converges**, then the limit point is a critical point of the function.
- **Theorem 2:** (arXiv:2006.01512) Backtracking New Q-Newton's method applied to a meromorphic function f in 1 complex variable, for which $\{x \in \mathbf{C} : f(x)f''(x) = f'(x) = 0\} = \emptyset$ with a random initial point, will either converge to a root or diverge to infinity.

- For entire functions, the point at infinity can be viewed as an additional root. (The case of polynomial is easy. The case of transcendental functions, we have essential singularity there.)

- For entire functions, the point at infinity can be viewed as an additional root. (The case of polynomial is easy. The case of transcendental functions, we have essential singularity there.)
- **Theorem 3:** (arXiv:2209.05378, T.) Backtracking New Q-Newton's method has, in addition to properties by New Q-Newton's method, good convergence guarantee. In particular: 1) Any **cluster point** of a sequence constructed is a critical point of the function. 2) If the function f is Morse, then for random initial points x_0, the constructed sequence $\{x_n\}$ either converges to a critical point, or diverges to infinity. 3) If the function f satisfies the gradient Lojasiewicz inequality near its critical points, with a bounded Lojasiewicz exponent, then if we choose the parameters of Backtracking New Q-Newton's method to be small enough, we have the same conclusions as in part 2).

- Gradient Lojasiewicz inequality: f satisfies the gradient Lojasiewicz inequality at a point p if there is an open neighbourhood U of p, constants $0 < \theta < 1$ and $C > 0$ such that for all $x \in U$ we have $|f(x) - f(p)|^{\theta} \leq C||\nabla f(x)||$.
- Real analytic functions satisfy the gradient Lojasiewicz inequality at every points.

- Gradient Lojasiewicz inequality: f satisfies the gradient Lojasiewicz inequality at a point p if there is an open neighbourhood U of p, constants $0 < \theta < 1$ and $C > 0$ such that for all $x \in U$ we have $|f(x) - f(p)|^{\theta} \leq C||\nabla f(x)||$.
- Real analytic functions satisfy the gradient Lojasiewicz inequality at every points.
- **Theorem:** (D. D'Acunto and K. Kurdyka) If f is a polynomial, then its Lojasiewicz exponents are explicitly bounded in terms of degree and dimension.

  Experiments show that even if the parameters in Backtracking New Q-Newton's method are not small, we still observe the conclusions of part 3 in Theorem 3.

- Gradient Lojasiewicz inequality: f satisfies the gradient Lojasiewicz inequality at a point p if there is an open neighbourhood U of p, constants $0 < \theta < 1$ and $C > 0$ such that for all $x \in U$ we have $|f(x) - f(p)|^{\theta} \leq C||\nabla f(x)||$.
- Real analytic functions satisfy the gradient Lojasiewicz inequality at every points.
- **Theorem:** (D. D'Acunto and K. Kurdyka) If f is a polynomial, then its Lojasiewicz exponents are explicitly bounded in terms of degree and dimension.

  Experiments show that even if the parameters in Backtracking New Q-Newton's method are not small, we still observe the conclusions of part 3 in Theorem 3.
- Roughly speaking, Theorem 3 says that if f is Morse or has a bounded gradient Lojasiewicz inequality, then for a random initial point x_0, Backtracking New Q-Newton's method will either converge to a local minimum or diverge to infinity.

Blank page

- This page is intended to be blank.

§ 2: Newton's method

Section 2.1: Newton's method, version 1

- This is direct application for equations.

Section 2.1: Newton's method, version 1

- This is direct application for equations.
- For 1 equation in 1 variable: $f(x) = 0$

  Choose z_0 randomly, construct

$$z_{n+1} = z_n - \frac{f(z_n)}{f'(z_n)}.$$

- This is direct application for equations.
- For 1 equation in 1 variable: $f(x) = 0$

  Choose z_0 randomly, construct

$$z_{n+1} = z_n - \frac{f(z_n)}{f'(z_n)}.$$

- For a system of N equations in m variables: $F(x) = 0$.

  Choose z_0 randomly, construct
  $z_{n+1} = z_n - (JF(z_n)^T.JF(z_n))^{-1}.JF(z_n)^T.F(z_n)$.
  Here JF is the Jacobian of F, and A^T is the transpose of a matrix A. Note: $JF(z_n)^T.F(z_n)$ is the gradient of f.

Section 2.1: Newton's method, version 1

- This is direct application for equations.
- For 1 equation in 1 variable: $f(x) = 0$

  Choose z_0 randomly, construct

$$z_{n+1} = z_n - \frac{f(z_n)}{f'(z_n)}.$$

- For a system of N equations in m variables: $F(x) = 0$.

  Choose z_0 randomly, construct
  $z_{n+1} = z_n - (JF(z_n)^T.JF(z_n))^{-1}.JF(z_n)^T.F(z_n)$.
  Here JF is the Jacobian of F, and A^T is the transpose of a matrix A. Note: $JF(z_n)^T.F(z_n)$ is the gradient of f.
- These update rules apply for both real and complex variables. For quaternionic variables, where multiplication is not commutative, need to choose either
  $z_{n+1} = z_n - f(z_n).(f'(z_n))^{-1}$ or $z_{n+1} = z_n - (f'(z_n))^{-1}.f(z_n)$.

Section 2.2: Newton's method, version 2

- This is through the associated optimization question: find global minima of $g(x)=\frac{1}{2}||F(x)||^2$.

  Choose z_0 randomly, construct

$$z_{n+1}=z_n-\frac{f(z_n)}{f'(z_n)}.$$

Section 2.2: Newton's method, version 2

- This is through the associated optimization question: find global minima of $g(x) = \frac{1}{2}||F(x)||^2$.

  Choose z_0 randomly, construct

$$z_{n+1} = z_n - \frac{f(z_n)}{f'(z_n)}.$$

- For a system of N equations in m variables: $F(x) = 0$.

  Choose z_0 randomly, construct
  $z_{n+1} = z_n - (\nabla^2 f(z_n))^{-1}.JF(z_n)^T.F(z_n)$.

  Here JF is the Jacobian of F, and A^T is the transpose of a matrix A.

Section 2.2: Newton's method, version 2

- This is through the associated optimization question: find global minima of $g(x) = \frac{1}{2}||F(x)||^2$.

  Choose z_0 randomly, construct

  $$z_{n+1} = z_n - \frac{f(z_n)}{f'(z_n)}.$$

- For a system of N equations in m variables: $F(x) = 0$.

  Choose z_0 randomly, construct
  $z_{n+1} = z_n - (\nabla^2 f(z_n))^{-1}.JF(z_n)^T.F(z_n)$.

  Here JF is the Jacobian of F, and A^T is the transpose of a matrix A.

- These update rules apply for both real and complex variables. For quaternionic variables, where multiplication is not commutative, need to choose either
  $z_{n+1} = z_n - f(z_n).(f'(z_n))^{-1}$ or $z_{n+1} = z_n - (f'(z_n))^{-1}.f(z_n)$.

Section 2.3: Results by Ernst Schröder, Arthur Cayley and Curtis McMullen

- Schröder proved his fixed point theorem for Newton's method. He connected Newton's method to iterations in dynamics.

  Theorem (Schröder): Let $f(x)$ be a polynomial of degree 2 with 2 distinct roots z_1 and z_2. Let L be the perpendicular bisector of the line segment joining z_1 and z_2. Then L divides the complex plane into 2 halves, each the basin of attraction of 1 root.

  Cayley has another proof of the above theorem, 10 years later, but more cited.

  Source: "Schröder, Cayley and Newton's method" by Daniel S. Alexander.

Section 2.3: Results by Ernst Schröder, Arthur Cayley and Curtis McMullen

- Schröder proved his fixed point theorem for Newton's method. He connected Newton's method to iterations in dynamics.

  Theorem (Schröder): Let $f(x)$ be a polynomial of degree 2 with 2 distinct roots z_1 and z_2. Let L be the perpendicular bisector of the line segment joining z_1 and z_2. Then L divides the complex plane into 2 halves, each the basin of attraction of 1 root.

  Cayley has another proof of the above theorem, 10 years later, but more cited.

  Source: "Schröder, Cayley and Newton's method" by Daniel S. Alexander.

- **Theorem** (McMullen, Annals of Mathematics): There is no algebraic dynamical systems which for a generic polynomial of degree ≥ 4, will converge to its roots when starting from a random initial point.

Section 2.3: Results by Ernst Schröder, Arthur Cayley and Curtis McMullen (cont. 2)

- The proof uses dynamical systems of the Riemann sphere. If degree of the polynomial is 3, then McMullen found an algebraic dynamical system to find roots.

  Corollary: To find roots of a general polynomial, one should look for non-algebraic iterative methods.

Section 2.4: Pros and cons

- Pros: Fast rate of convergence near non-degenerate local minima, Easy to implement, Beautiful pictures of basins of attraction (fractal structure).

- Pros: Fast rate of convergence near non-degenerate local minima, Easy to implement, Beautiful pictures of basins of attraction (fractal structure).
- Cons: Problematic with global convergence (may diverge to infinity, may have attracting cycles with more than 1 element, may converge to a local maximum or saddle points, illustrated pictures: see later), Huge calculation cost in large dimension (complexity for local convergence near non-degenerate local minima $= O(m^{\omega} \log|\log \epsilon|)$ where m is the dimension, ϵ is the error threshold, $\omega \geq 2$ the exponent for complexity of multiplying two square matrices), Fractal structure makes the algorithm sensitive to the choice of initial points.

Section 2.4: Pros and cons (con. 2)

- **Remark**: John Hubbard, Dierk Schleicher and Scott Sutherland (Inventiones Mathematicae) constructed a finite set S_d, such that if f is a polynomial in 1 complex variable of degree d and whose all roots lie in the open unit disk $\{|x| < 1\}$, then Newton's method with an initial point in this S_d will find all the roots of f.

  The cardinality of this set S_d is $O(d \log d)$. If all the roots are real, then the cardinality of S_d is $O(d)$.

  For each root of the polynomial, there is one point of this set S_d which lies in the basin of attraction for the root.
- The proof of this result heavily depends on that the function f is a polynomial.

Blank page

- This page is intended to be blank.

- Damping Newton's method (aslo known as relaxed Newton's method) for 1 variable:

$$z_{n+1} = z_n - \alpha \frac{f(z_n)}{f'(z_n)}.$$

Here α is a constant complex number.

Still has advantages/disadvantages as Newton's method.

§ 3: Random damping Newton's method

- Damping Newton's method (aslo known as relaxed Newton's method) for 1 variable:

$$z_{n+1} = z_n - \alpha \frac{f(z_n)}{f'(z_n)}.$$

  Here α is a constant complex number.

  Still has advantages/disadvantages as Newton's method.

- Random damping Newton's method: At each step n, choose randomly an $\alpha_n \in \mathbf{C}$, and use the rule

$$z_{n+1} = z_n - \alpha_n \frac{f(z_n)}{f'(z_n)},$$

- **Theorem** (Sumi, Communications in Mathematical Physics): Let f be a polynomial of degree two or more. Let $0.5 < r < 1$. Choose a sequence α_n of complex numbers randomly from $\{a : |a-1| < r\}$, with the uniform distribution. Then for each initial point (except some finite points), the random damping Newton's method:

$$z_{n+1} = z_n - \alpha_n f(z_n)/f'(z_n)$$

  will converge to a root of f almost surely.
- The paper also presents other properties regarding the dynamics of random damping Newton's method.

- Pros: as that of Newton's method, plus convergence guarantee for finding roots of polynomials in 1 complex variable.

§ 3.2: Pros and cons

- Pros: as that of Newton's method, plus convergence guarantee for finding roots of polynomials in 1 complex variable.
- Cons:

  In the statement of Sumi's theorem, if one chooses r small, e.g. $r = 0.01$, then the behaviour may like that of Newton's method. (see pictures later)

  Extensions to higer dimensions? (Remark: Recently, Sumi has some new results in dimension 2.)

  For non-polynomials? For real variables? How much random do we need? Many things are unclear.

Blank page

- This page is intended to be blank.

§ 4: Some variations

§ 4.1: The approach by Shub and Smaleb

- The approach by Shub and Smale: We want to solve a system of polynomial equations $F(x) = 0$, in several complex variables.

  Smale found an efficient method to test, using estimates of all derivatives of the polynomial map at a given point x^*, whether that point is where Banach fixed point theorem can be applied to the dynamics of Newton's method.

  Shub developed a version of Newton's method for projective spaces.

  Together, they developed in the "Bezout" series of paper, a method to find a root of a system of polynomial equations, by combining Newton's method on projective spaces and homotopy continuation.

  The key idea is to start from another (simpler) system G=0 and a point ζ, and try to use Newton's method on small intervals to reach a root of $F = 0$.

§ 4.1: The approach by Shub and Smale (cont. 2)

- Smale's 17th problem: Finding an approximate zero (i.e., one for which Newton's method applied to it will converge to a root) of a polynomial system in polynomial time, in the average case.

- Smale's 17th problem: Finding an approximate zero (i.e., one for which Newton's method applied to it will converge to a root) of a polynomial system in polynomial time, in the average case.
- This problem is solved in the affirmative by Beltrand and Pardo, Cucker and Burgisser, Lairez (Foundations of Computational Mathematics, Annals of Mathematics...).

  One key idea (Beltrand and Pardo): They show the existence of a set of polynomial systems of degree bounded by d, such that for every ϵ, if we choose randomly a (G, ζ) from the set, then for a randomly chosen polynomial system $F = 0$ of degree bounded by d, the homotopy continuation method applied to (G, ζ) will produce an approximate root of $F = 0$ with probability $\geq 1 - \epsilon$.

  Cons: Homotopy continuation may not work well in the case of real variables. Also, how about non-polynomial systems?

- The straightforward implementation of this approach may be expensive to run, and may be not working well.

  An implementation has been available in some decades, that is the software Bertini, and works quite well for polynomial systems in several complex variables.

  Performance of Bertini is better than symbolic methods (like Grobner basis,...)

§ 4.2: Levenberg-Marquardt method and Regularized Newton's method

- Levenberg-Marquardt method: one default method for Least Square Fit problem. Apply to cost functions of the form: $f = ||F||^2/2$, where $F = (f_1, \ldots, f_N)$.
  $z_{n+1} = z_n - (JF(z_n)^T.JF(z_n) + \gamma_n Id)^{-1}.JF(z_n)^T.F(z_n)$. γ_n positive number. Usually $\gamma_n = c||F(z_n)||^\tau$, constants $\tau, c > 0$.
  Pros: don't need to compute the whole Hessian of f, works well in many cases. Cons: not enough information to control near saddle points, no global convergence guaranteed.
  Remark: There is work where Armijo's Backtracking line search is incorporated. Global convergence is improved, however avoidance of saddle point is still not addressed. In arXiv:2209.05378, some improvements, with stronger results on convergence proved (with help of results in T. and H.-T. Nguyen, Applied Mathematics and Optimization & Minimax Theory and its applications).

§ 4.2: Levenberg-Marquardt method and Regularized Newton's method (cont. 2)

- Regularized Newton's method: Apply to a general cost function f.

  Add a term to the Hessian of f to make it positive definite. Update rule:

  $z_{n+1} = z_n - (\nabla^2 f(z_n) + \lambda_n Id)^{-1}.\nabla f(z_n)$. Usually, choose $\lambda_n = 2\max(0, -\lambda_1(\nabla^2 f(z_n)))$. Here $\lambda_1(\nabla^2 f(z_n))$ is the smallest eigenvalue of $\nabla^2 f(z_n)$.

  Pros: works more stably than Levenberg-Marquardt method. Cons: more expensive, no result on avoidance of saddle points is known.

§ 4.3: Cubic Regularization

- Introduced by Yurii Nesterov and B. T. Polyak in 2006, Mathematical Programming, Ser A 108, 177–205, 2006. (Earlier similar work is by Griewank.)

  At each step n, it updates $z_{n+1} = z_n + s_n$, where s_n minimises the function $s \mapsto K(s) = f(s) + < \nabla f(z_n), s > + < \nabla^2 f(z_n) s, s > + \sigma_n ||s||^3$. Here σ_n is a positive constant.

- Introduced by Yurii Nesterov and B. T. Polyak in 2006, Mathematical Programming, Ser A 108, 177–205, 2006. (Earlier similar work is by Griewank.)
  At each step n, it updates $z_{n+1} = z_n + s_n$, where s_n minimises the function $s \mapsto K(s) = f(s) + < \nabla f(z_n), s > + < \nabla^2 f(z_n)s, s > + \sigma_n ||s||^3$. Here σ_n is a positive constant.
- Remarks: In their paper, Nesterov and Polyak showed that to find the exact minimum of $K(s)$, want need to compute the eigenvector/eigenvalue pairs of $\nabla^2 f(z_n)$. So, their exact formulation has the same complexity as (Backtracking) New Q-Newton's method - see later.
  They showed that (for strongly convex functions) this method can avoid saddle points. Later work extended to adaptive Cubic Regularization, which applies to more general functions (still need some complicated constraints, like ∇f and/or $\nabla^2 f$ is Lipschitz continuous, on the function to be able to show good properties).

§ 4.3: Cubic Regularization (cont. 2)

- Global convergence for this method is proven under some similar constraints on the cost function f.

§ 4.3: Cubic Regularization (cont. 2)

- Global convergence for this method is proven under some similar constraints on the cost function f.
- However, these results are not as strong or general as those in the main theorems of this talk.

  In particular, there is no guarantee the method can be useful for finding roots of polynomials in 1 complex variable.

- Global convergence for this method is proven under some similar constraints on the cost function f.
- However, these results are not as strong or general as those in the main theorems of this talk.

  In particular, there is no guarantee the method can be useful for finding roots of polynomials in 1 complex variable.
- Since this is a trust region method, implementing it is difficult.

  There is only one public repository so far, and it is not straightforward from the theoretical algorithm.

  It contains many parameters, and the performance is very sensitive to the choice of the parameters.

§ 4.4: Summary of previous variants of Newton's method

- Prior to Backtracking New Q-Newton's method, there are thousands of variations of Newton's method. (And many new ones appear very regularly after.)

  Those which are inexpensive (like quasi-Newton's methods such as BFGS) has weak theoretical guarantee.

  Those which have stronger theoretical theoretical guarantees, usually still need quite restrictive constraints (like the gradient or the Hessian of the function is Lipschitz continuous).

  Those which have both good global convergence and avoidance of saddle points can be difficult to implement, or may not have quick rate of convergence.

Blank page

- This page is intended to be blank.

- Heuristic:

  Issue 1: A saddle point is where the Hessian has a negative eigenvalue. So if we change the signs of negative eigenvalues, maybe we can avoid saddle points.

  Issue 2: The Hessian at a point may be singular $det(\nabla^2 f(z_n)) = 0$. We need to add a term to make it invertible.

  Issue 3: To preserve the quick rate of convergence, the term we add in resolving Issue 2 should be small when the gradient is small.

- Input: A function $f : \mathbf{R}^m \rightarrow \mathbf{R}$. $m+1$ random distinct constants $\kappa_0, \ldots, \kappa_m$. A constant $\tau > 0$.

  Output: a local minimum of f.

§ 5.1: New Q-Newton's method, algorithm

- Input: A function $f : \mathbf{R}^m \rightarrow \mathbf{R}$. $m + 1$ random distinct constants $\kappa_0, \ldots, \kappa_m$. A constant $\tau > 0$.

  Output: a local minimum of f.
- Step 0: Choose an initial point z_0 randomly.

- Input: A function $f : \mathbf{R}^m \rightarrow \mathbf{R}$. $m+1$ random distinct constants $\kappa_0, \ldots, \kappa_m$. A constant $\tau > 0$.

  Output: a local minimum of f.
- Step 0: Choose an initial point z_0 randomly.
- Step n:

  If z_n is a critical point, STOP. Otherwise,

  Step n.1: Choose κ the first term in $\kappa_0, \ldots, \kappa_m$ so that $A_n := \nabla^2 f(z_n) + \kappa ||\nabla f(z_n)||^\tau$ is invertible.

  Step n.2: Let (λ_j, v_j) ($j = 1, \ldots, m$) be an orthonormal basis of pairs of eigenvalue/eigenvector of A_n.

  Step n.3: Update $z_{n+1} = z_n - \sum_{j=1}^m < v_j, \nabla f(z_n) > v_j/|\lambda_j|$.

- Proof for fast rate of convergence: Assume that z^* is a non-degenerate local minimum. We need to show that for z_0 close enough to z^*, then the sequence z_n converges to z^* with quadratic rate of convergence. This follows from that $\nabla^2 f(z^*)$ is invertible, hence A_n is close to $\nabla^2 f(z_n)$. Hence, near a non-degenerate local minimum, New Q-Newton's method behaves like Newton's method.

§ 5.2: Proofs of main properties

- Proof for fast rate of convergence: Assume that z^* is a non-degenerate local minimum. We need to show that for z_0 close enough to z^*, then the sequence z_n converges to z^* with quadratic rate of convergence. This follows from that $\nabla^2 f(z^*)$ is invertible, hence A_n is close to $\nabla^2 f(z_n)$. Hence, near a non-degenerate local minimum, New Q-Newton's method behaves like Newton's method.
- Proof for avoidance of saddle points:

  2.1: Existence of local Stable-Central manifolds near saddle points of f, under the dynamics of New Q-Newton's method

  2.2: Show that for random initial point z_0, iterates of z_0 won't land in these Stable-Central manifolds.

§ 5.2: Proofs of main properties

- Proof for fast rate of convergence: Assume that z^* is a non-degenerate local minimum. We need to show that for z_0 close enough to z^*, then the sequence z_n converges to z^* with quadratic rate of convergence. This follows from that $\nabla^2 f(z^*)$ is invertible, hence A_n is close to $\nabla^2 f(z_n)$. Hence, near a non-degenerate local minimum, New Q-Newton's method behaves like Newton's method.
- Proof for avoidance of saddle points:

  2.1: Existence of local Stable-Central manifolds near saddle points of f, under the dynamics of New Q-Newton's method

  2.2: Show that for random initial point z_0, iterates of z_0 won't land in these Stable-Central manifolds.
- Remark: Proofs of 1 and 2.1 don't need that $\kappa_0, \ldots, \kappa_m$ are random.

§ 5.2: Proofs of main properties (cont. 2)

- Proof of 2.1: Will apply Stable - Central manifold, hence need to first show that near a saddle point, the dynamics is C^1 (a priori, it is not continuous at every points).

  Since at a saddle point, the Hessian is invertible and the gradient is 0, $\kappa = \kappa_0$.

  To show that the dynamics is C^1, we need to assume that f is C^3. Also, need the following results from the theory of Linear operators, concerning eigenvalue/eigenvector of a parametrized linear operator:

§ 5.2: Proofs of main properties (cont. 2)

- Proof of 2.1: Will apply Stable - Central manifold, hence need to first show that near a saddle point, the dynamics is C^1 (a priori, it is not continuous at every points).

  Since at a saddle point, the Hessian is invertible and the gradient is 0, $\kappa = \kappa_0$.

  To show that the dynamics is C^1, we need to assume that f is C^3. Also, need the following results from the theory of Linear operators, concerning eigenvalue/eigenvector of a parametrized linear operator:

- Result: $A(x) = C^1$ family of real symmetric $m \times m$ matrices. $C =$ a circle in $\mathbf{C}$ such that for all x open set U, C contains k eigenvalues/eigenvectors $(\lambda_1(x), e_1(x))$, ..., $(\lambda_k(x), e_k(x))$ of $A(x)$. $e_1(x), \ldots, e_k(x)$ orthonormal. Moreover, ∂C contains no eigenvalue of any $A(x)$. Then the projection $pr(v) := \sum_{i=1}^{k} \lambda_i(x) < v, e_i(x) > e_i(x)$ can be represented as $\int_C (A(x) - \zeta Id)^{-1} v d\zeta$.

- Proof of 2.2:

  First, show that there is a set $\mathcal{E} \subset \mathbf{R}$ of zero Lebsegue measure such that if $\kappa \in \mathbf{R} \backslash \mathcal{E}$, then the set z where $\nabla f(z) \neq 0$ and $A(z, \delta) = \nabla^2 f(z) + \kappa ||\nabla f(z)||^{\tau}$ is not invertible belongs to $\mathcal{E}$.

  From this, it follows that for these κ, z, then the preimage by $A(z, \kappa)$ of a set of zero Lebesgue measure also has zero Lebesgue measure.

  The randomness of the constants $\kappa_0, \ldots, \kappa_{m+1}$ can be made precise: Just choose distinct m+1 numbers in the set $\mathbf{R} \backslash \mathcal{E}$.

  If z_0 is such that its orbit under $A(z, \kappa)$ converges to a saddle point, then it must meet one of the local Stable-Central manifolds where κ is one of the $\kappa_0, \ldots, \kappa_m$, then z_0 belongs to a countable union of sets of the form $A(z_1, \kappa_1')^{-1} \circ \ldots \circ A(z_k, \kappa_1')^{-1}(\mathcal{S})$, where $\mathcal{S}$ is a local Stable-Central manifold and hence is of zero Lebesgue measure.

§ 5.3: Missing point: convergence

- New Q-Newton' method is better than Newton' method: it has the same rate of convergence, it is as easy to implement (with a higher complexity), and in addition it can avoid saddle points.

  What it misses: show global convergence under some reasonable assumptions, and in case convergence not known show that every cluster point is a critical point of the function.

  Remark: if the sequence constructed by New Q-Newton's method converges, then the limit point is known to be a critical point of the function f.

  Backtracking New Q-Newton's method addresses this convergence issue.

- This page is intended to be blank.

§ 6: Backtracking New Q-Newton's method

- Heuristic idea: add a mechanism to assure global convergence. Two commonly useful techniques in the Optimization literature: trust region method and Backtracking line search.
- Trust region method: at each step n, solve a optimal subproblem on a region around the current point z_n (like in Cubic Regularization). Can prove some good theoretical results. However, there are many parameters to take care of, and no definite procedure to choose these parameters. Consequently, very difficult to implement, the implementations are very sensitive to parameters and hence unstable.
- Backtracking line search: The update rule is $z_{n+1} = z_n - \alpha_n v_v$, where v_n is a pre-determined vector (descent direction) for example $\nabla f(z_n)$, and $\alpha_n > 0$ a number to be determined by a backtracking procedure. (In this case, the truth region method would like to find minimum of the function $\alpha \mapsto f(z_n - \alpha v_n)$.) Very flexible, easy to implement, and stable.

§ 6.1: Armijo's Backtracking line search

- We will use specifically Armijo's Backtracking line search. Introduced by Armijo in the 1960's. (A related but more complicated procedure, but not giving better results or experiments, is Wolffe's condition, which needs stronger conditions than Armijo's condition.)

- We will use specifically Armijo's Backtracking line search. Introduced by Armijo in the 1960's. (A related but more complicated procedure, but not giving better results or experiments, is Wolffe's condition, which needs stronger conditions than Armijo's condition.)
- Input: z_n so that $\nabla f(z_n) \neq 0$. Also a vector w_n such that $< \nabla f(z_n), w_n >> 0$. Constants $1 \geq \alpha_0 > 0$ and $0 < \beta < 1$ (for the Backtracking procedure).
  Output: $\alpha_n > 0$ and the update $z_{n+1} = z_n - \alpha_n w_n$.

§ 6.1: Armijo's Backtracking line search

- We will use specifically Armijo's Backtracking line search. Introduced by Armijo in the 1960's. (A related but more complicated procedure, but not giving better results or experiments, is Wolffe's condition, which needs stronger conditions than Armijo's condition.)
- Input: z_n so that $\nabla f(z_n) \neq 0$. Also a vector w_n such that $< \nabla f(z_n), w_n >> 0$. Constants $1 \geq \alpha_0 > 0$ and $0 < \beta < 1$ (for the Backtracking procedure).
  Output: $\alpha_n > 0$ and the update $z_{n+1} = z_n - \alpha_n w_n$.
- Armijo's condition: $\alpha > 0$ satisfies Armijo's condition if $f(z_n - \alpha_n w_n) - f(z_n) \leq -\alpha < w_n, \nabla f(z_n) > /3$. (Remark: If f is C^1, then by Taylor's expansion, this is satisfied by $\alpha > 0$ small enough.)
  Armijo's Backtracking line search: Put $\alpha := \alpha_0$. While Armijo's condition is not satisfied, define $\alpha := \alpha\beta$.
  Remark: There are improvements of Armijos' procedure, to reduce calculations, see (T.-Nguyen, cited before).

Blank page

- This page is intended to be blank.

- The Backtracking New Q-Newton's method modifies New Q-Newton's method on the following aspects:

§ 6.2: The algorithm

- The Backtracking New Q-Newton's method modifies New Q-Newton's method on the following aspects:
- When choosing κ: Need stronger condition: Choose κ the first κ_j for which min spectral radius of $A(x,\kappa) \geq \kappa^* ||\nabla f(z_n)||^{tau}$, where $\kappa^* = \min_{i \neq j} |\kappa_i - \kappa_j|/2$.
  (Remark: in experiments, even the simpler choice in New Q-Newton's method also works well. Heuristic explanation given later.)

§ 6.2: The algorithm

- The Backtracking New Q-Newton's method modifies New Q-Newton's method on the following aspects:
- When choosing κ: Need stronger condition: Choose κ the first κ_j for which min spectral radius of $A(x,\kappa) \geq \kappa^* ||\nabla f(z_n)||^{tau}$, where $\kappa^* = \min_{i\neq j} |\kappa_i - \kappa_j|/2$.

  (Remark: in experiments, even the simpler choice in New Q-Newton's method also works well. Heuristic explanation given later.)
- Backtracking line search to update: Use Armijo's Backtracking line search for $w_n := \sum_{j=1}^{m} < v_j, \nabla f(z_n) > v_j/|\lambda_j|$. (Check that $< w_n, \nabla f(z_n) >> 0$!)

  If f does not have compact sublevels, then normalize w_n to $w_n/\max\{1, ||w_n||\}$ before running Armijo's Backtracking line search. This is to reduce the chance of divergence to infinity.

§ 6.3: The special case of function in 1 real variable

Illustration: Explicit formulas for applying Backtracking New Q-Newton's method to find roots of a function $F: \mathbf{R} \to \mathbf{R}$.

- Choose two random real numbers κ_0, κ_1. Choose two numbers $\alpha_0 > 0$ and $0 < \beta < 1$. Define $\kappa^* = |\kappa_0 - \kappa_1|/2$.

  Define $f(x) = F(x)^2/2$. Then $f'(x) = F(x)F'(x) \Rightarrow$ critical points of f are either roots of F or critical points of F.

  $f''(x) = F'(x)^2 + F(x)F''(x)$. $\Rightarrow$ If x is critical point of f but not a root of F, and $F(x)F''(x) > 0$, then x is a local minimum but not global minimum of f.

Illustration: Explicit formulas for applying Backtracking New Q-Newton's method to find roots of a function $F : \mathbf{R} \to \mathbf{R}$.

- Choose two random real numbers κ_0, κ_1. Choose two numbers $\alpha_0 > 0$ and $0 < \beta < 1$. Define $\kappa^* = |\kappa_0 - \kappa_1|/2$.

  Define $f(x) = F(x)^2/2$. Then $f'(x) = F(x)F'(x) \Rightarrow$ critical points of f are either roots of F or critical points of F.

  $f''(x) = F'(x)^2 + F(x)F''(x)$. $\Rightarrow$ If x is critical point of f but not a root of F, and $F(x)F''(x) > 0$, then x is a local minimum but not global minimum of f.

- If $|f''(z_n) + \kappa_0|f'(z_n)|^\tau| \geq \kappa^*|f'(z_n)|^\tau$, then $\kappa = \kappa_0$. Otherwise, $\kappa = \kappa_1$. $A_n = f''(z_n) + \kappa_0|f'(z_n)|^\tau$.

  Define $w_n = f'(z_n)/|A_n|$. (If f does not have compact sublevels, then define $w_n = w_n/\max\{1, ||w_n||\}$).

  Run Armijo's Backtracking line search procedure to find $\alpha_n > 0$, and update $z_{n+1} = z_n - \alpha_n w_n$.

Blank page

- This page is intended to be blank.

§ 6.4: Proof of main properties

- Avoidance of saddle points: The key point is to show that if z^* is a non-degenerate critical point of f, and z_n is close to z^*, then with w_n from Backtracking New Q-Newton's method we have for all $0 \leq \alpha \leq 1$ we have:

  $f(z_n - \alpha_n w_n) - f(z_n) \leq -\alpha < w_n, \nabla f(z_n) > /2 + o(||w_n||^2).$

  $\Rightarrow \alpha_n = \alpha_0$ near z^*.

  In particular, the dynamics of Backtracking New Q-Newton's method is C^1 near a saddle point z^*, and we can do as before for New Q-Newton's method.

  Remark: The fact $\alpha_n = \alpha_0$ used in the above proof is quite delicate, and it is true because we choose w_n as coming from some perturbations of the Hessian. If, on the other hand, one chooses $w_n = \nabla f(z_n)$ (as in Gradient Descent, introduced by Cauchy), then it is not known (even though there is strong evidence to support) if Armijo's Backtracking line search can avoid saddle point.

§ 6.4: Proof of main properties (cont. 2)

- Quick rate of convergence: If we choose $\alpha_0 = 1$, then the fact that $\alpha_n = \alpha_0 = 1$ for z_n near a non-degenerate local minimum shows that Backtracking New Q-Newton's method behaves like the usual Newton's method, and hence has quadratic convergence rate.

- Quick rate of convergence: If we choose $\alpha_0 = 1$, then the fact that $\alpha_n = \alpha_0 = 1$ for z_n near a non-degenerate local minimum shows that Backtracking New Q-Newton's method behaves like the usual Newton's method, and hence has quadratic convergence rate.
- Global convergence: This is proven in the following 2 reasonable cases:

  Case 1: f has countably many critical points. (This is the generic situation, by transversal theory.)

  Case 2: f is real analytic, or more generally satisfies gradient Lojasiewicz inequality near its critical points.

- Proof of global convergence for Case 1: We use the following result
- Result: (T.-Nguyen, cited before) 1). If one applies Armijo's Backtracking line search to descent directions w_n satisfying the condition $< w_n, \nabla f(z_n) >\geq c||\nabla f(z_n)||^{\tau}$ for all n, where $c, \tau > 0$ are constants, then any cluster point of the constructed sequence z_n is a critical point of f. 2). Let z_n be a sequence in $\mathbf{R}^m$ such that $||z_{n+1} - z_n|| \rightarrow 0$. If the set of the sequence z_n is a non-empty countable set, then the sequence z_n indeed converges.

  Remark: the proof of point 2 above uses a result by M. D. Asic and D. D. Adamovic, The American mathematical monthly, on cluster points of sequences in a compact metric space.

§ 6.4: Proof of main properties (cont. 4)

- Proof of global convergence for Case 2: We use the following result
- Theorem (P.A: Absil, R. Mahony and B. Andrews, SIAM Journal on Optimization) Let f be a function, and z_n a sequence satisfying Armijo's condition, where $< w_n, \nabla f(z_n) >\geq c||\nabla f(z_n)||^2$. Let z^* be a point at which f satisfies the gradient Lojasiewicz inequality. If z_n has a subsequence converges to z^*, then the whole sequence z_n converges to z^*.

  Remark: Lojasiewicz proved convergence of gradient flow for functions satisfying his namesake' condition.

- Finding roots of meromorphic functions

  Let F be a meromorphic function in 1 complex variable z. So singular points of F are poles. We want to solve $F(z) = 0$.

- Write $z = x + iy$, and $f(x, y) = ||F(z)||^2/2$. f is a (not defined everywhere) map from $\mathbf{R}^2$ to $\mathbf{R}$. $f(x, y) = 0$ if and only if z is a root of F. $f(x, y) = +\infty$ if and only if z is a pole of F.

- Finding roots of meromorphic functions

  Let F be a meromorphic function in 1 complex variable z. So singular points of F are poles. We want to solve $F(z) = 0$.

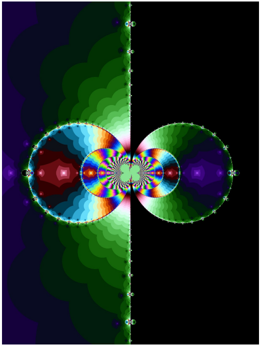
- Write $z = x + iy$, and $f(x, y) = ||F(z)||^2/2$. f is a (not defined everywhere) map from $\mathbf{R}^2$ to $\mathbf{R}$. $f(x, y) = 0$ if and only if z is a root of F. $f(x, y) = +\infty$ if and only if z is a pole of F.

- Result 1: (x, y) is a critical point of f if and only if z is a root of $F(z)F'(z)$.

  Result 2: Assume that $\{z \in \mathbf{C} : F(z)F''(z) = F'(z) = 0\} = \emptyset$. Then if (x, y) is a critical point of f but not a root of $F(z)$, then it is a saddle point of f.

Blank page

- This page is intended to be blank.



- Image created by Josep Maria Batile i Ferrer: (Rotated by -90 degree) Fractal structure of the basin of attraction of Newton's method for finding roots of the transcendental function $cosh(z) - 1$.

Blank page

- This page is intended to be blank.

- Some remarks:

  For polynomials of degree 2, Backtracking New Q-Newton's method satisfies Schr'oder/Cayley theorem. The same for Backtracking Gradient Descent.

§ 7.1: Basins of attraction

- Some remarks:

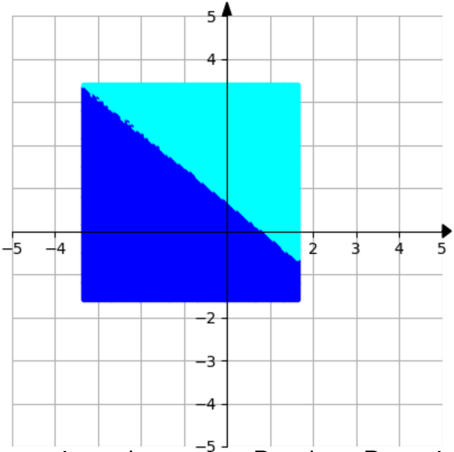
  For polynomials of degree 2, Backtracking New Q-Newton's method satisfies Schŕoder/Cayley theorem. The same for Backtracking Gradient Descent.

- For polynomials of degree 3 and 4, basins of attraction for Backtracking New Q-Newton's method seem not have fractal structure. Basins of attraction for Newton's method are well known to have fractal structures. Basins of attraction for Backtracking Gradient descent seem to be less smooth, and may have fractal structures.

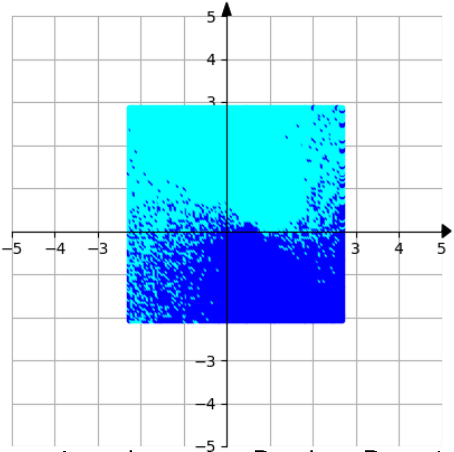
  Show examples from the paper: arXiv:2209.05378.

  More examples (including random damping Newton's method).

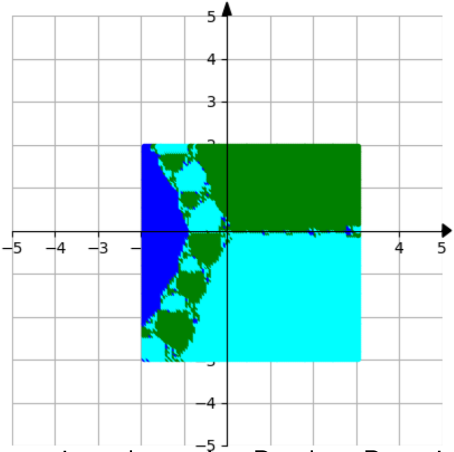
  Interestingly, Random Damping Method with α_n close to 1, applied to polynomial of degree 2 in 1 complex variable, does not satisfy Schroder/Cayley theorem!



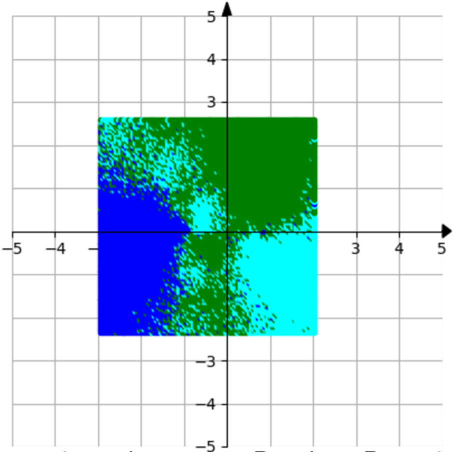
- Basin of attraction when using Random Damping Newton's method for the complex polynomial $(z - z_1^*)(z - z_2^*)$ for $z_1^* = 0.5 - 0.2i$ and $z_2^* = 1 + 0.4i$. Here $|\alpha_n - 1|$ is small.



- Basin of attraction when using Random Damping Newton's method for the complex polynomial $(z - z_1^*)(z - z_2^*)$ for $z_1^* = 0.5 - 0.2i$ and $z_2^* = 1 + 0.4i$. Here $|\alpha_n - 1|$ is large.

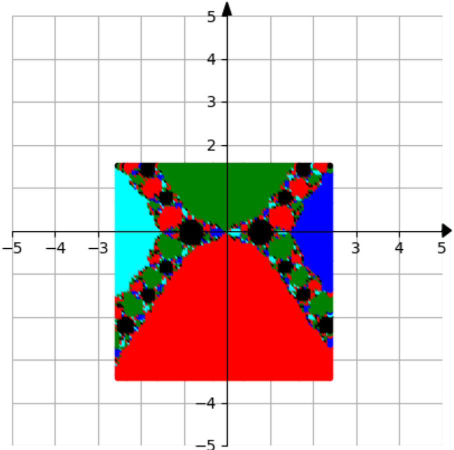


- Basin of attraction when using Random Damping Newton's method for the complex polynomial $z^3 - 2z + 2$. Here $|\alpha_n - 1|$ is small.

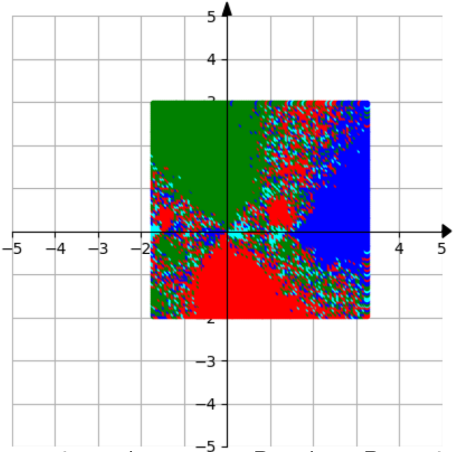


- Basin of attraction when using Random Damping Newton's method for the complex polynomial $z^3 - 2z + 2$. Here $|\alpha_n - 1|$ is large.

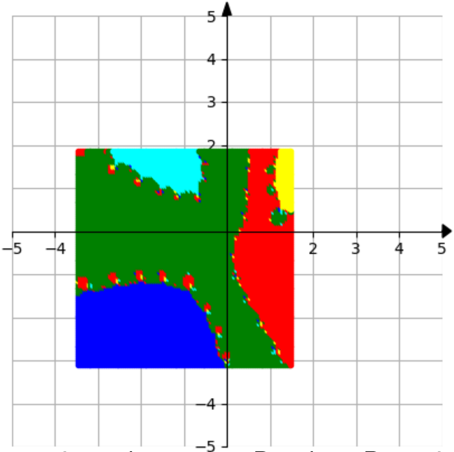
§ 7.1: Basins of attraction (cont. 6)



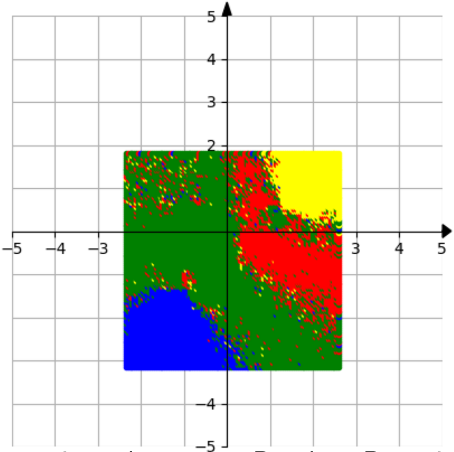
- Basin of attraction when using Random Damping Newton's method for the complex polynomial $(z^2+1)(z-2.3)(z+2.3)$. Here $|\alpha_n - 1|$ is small.



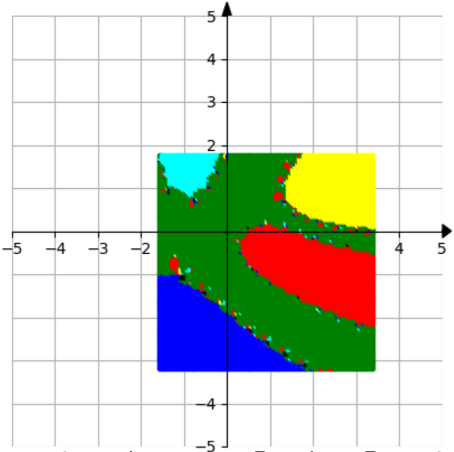
- Basin of attraction when using Random Damping Newton's method for the complex polynomial $(z^2+1)(z-2.3)(z+2.3)$. Here $|\alpha_n - 1|$ is large.



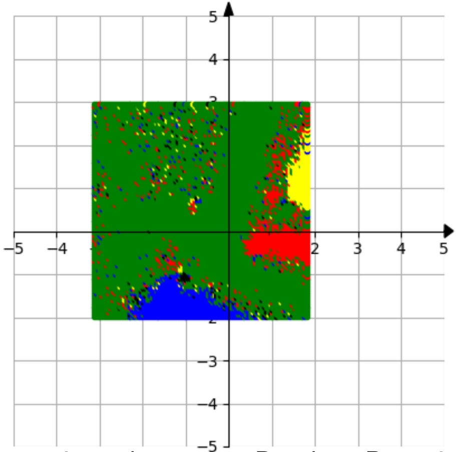
- Basin of attraction when using Random Damping Newton's method for the complex polynomial $z^5 - 3iz^3 - (5+2i)z^2 + 3z + 1$. Here $|\alpha_n - 1|$ is small.



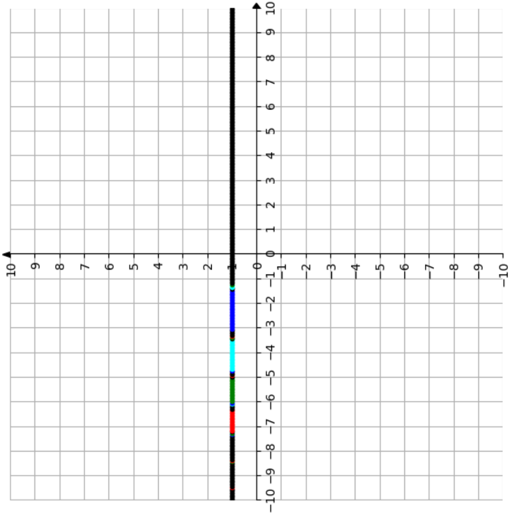
- Basin of attraction when using Random Damping Newton's method for the complex polynomial $z^5 - 3iz^3 - (5+2i)z^2 + 3z + 1$. Here $|\alpha_n - 1|$ is large.



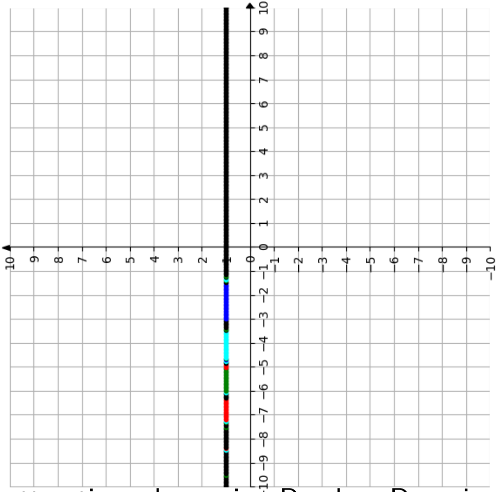
- Basin of attraction when using Random Damping Newton's method for the transcendental function in 1 complex variable $e^z(z^5 - 3iz^3 - (5+2i)z^2 + 3z + 1)$. Here $|\alpha_n - 1|$ is small.



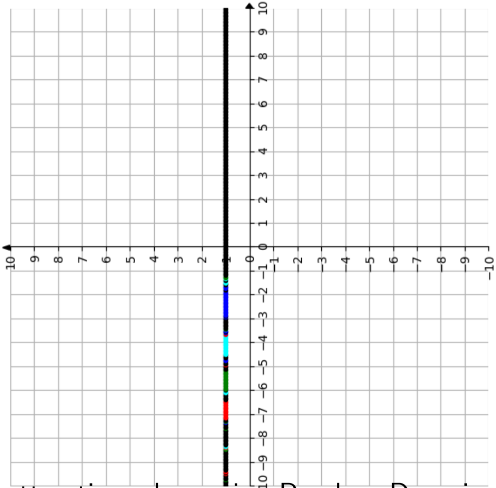
- Basin of attraction when using Random Damping Newton's method for the transcendental function in 1 complex variable $e^z(z^5 - 3iz^3 - (5+2i)z^2 + 3z + 1)$. Here $|\alpha_n - 1|$ is large.



- Basin of attraction when using Newton method for the Airy function in 1 real variable.

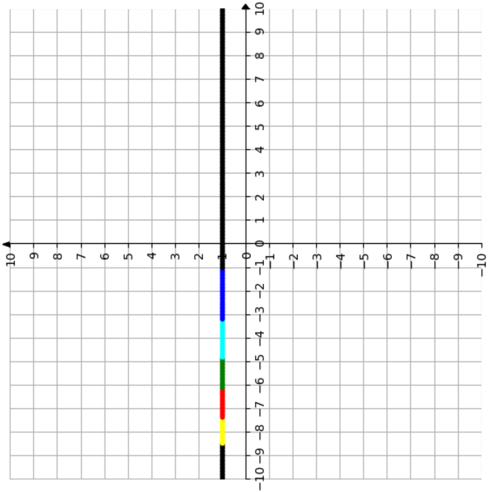


- Basin of attraction when using Random Damping Newton's method for the Airy function in 1 real variable. Here $|\alpha_n - 1|$ is small.



- Basin of attraction when using Random Damping Newton's method for the Airy function in 1 real variable. Here $|\alpha_n - 1|$ is large.

§ 7.1: Basins of attraction (cont. 15)



- Basin of attraction when using Backtracking New Q-Newton's method for the Airy function in 1 real variable.

Blank page

- This page is intended to be blank.

- Some remarks:

  To find roots of equations of holomorphic equations, we can reduce to finding roots of a system of real analytic functions (2 x number of equations, 2 x number of variables). Backtracking New Q-Newton's method may converge to local minima instead of roots (global minima).

  Some experiments show that if we regard these new real variables as **complex variables**, then reduce them to real variables again (hence, 4 x number of equations, 4 x number of equations), then convergence to global minima are observed, even if we start close to those points for the previous system the convergence is to local minima.

Blank page

- This page is intended to be blank.

§ 8: Summary and Open questions

- Summary:

  New Q-Newton's method modifies the Hessian by a term of the form $\kappa||\nabla f(z_n)||^{\tau}$. κ is one of the m+1 numbers randomly chosen from beginning. This helps resolve the case the Hessian is not invertible. Then change the sign of negative eigenvalues of the new operator. This helps to avoid saddle points. Same local rate of convergence as Newton's method.

  Backtracking New Q-Newton's method adds Armijo's Backtracking line search. This helps with global convergence.

  Roughly speaking: if f has countably many critical points or satisfies the gradient Lojasiewicz inequality, then global convergence to local minima.

  In particular, proven global convergence to roots of **ameromorphicfunction** in 1 complex variable.

  All these results are new in the literature.

- Summary (cont. 2):

  Implementation is straight forward. Performance is stable, not sensitive to the changes of parameters.

  Experiments with just a few eigenvalues/eigenvectors used show good performance.

- Summary (cont. 2):

  Implementation is straight forward. Performance is stable, not sensitive to the changes of parameters.

  Experiments with just a few eigenvalues/eigenvectors used show good performance.

- Open questions:

  Reduce complexity?

  Large scale possibility?

  More assurance of global convergence to global minima?

Thank you very much for your attention!