Strong contraction mapping and topological non-convex optimization

Siwei Luo
University of Illinois at Chicago
1200 W Harrison St, Chicago, IL 60607
sluo21@uic.edu

Abstract

The strong contraction mapping, a surjective self-mapping that its range always belongs to a subset of its domain, admits a unique fixed-point that can be pinned down by the iteration of the mapping. As an application of strong contraction mapping, we design a topological non-convex optimization method to achieve global minimum convergence. The strength of the approach is its robustness to local minima and initial point position.

1 Introduction

Calculus of variation plays an essential role in the modern calculation. Mostly, the function of interest cannot be obtained directly but makes a functional attain its extremum. This leads to many different functionals are constructed according to different problems, for instance, the least action principle, Fermat's principle, maximum likelihood estimation, finite element analysis, machine learning and so forth. These methods provide us the routine to transfer the original problem to an optimization problem, making optimization methods gain the influence and popularity.

Non-convex optimization arises naturally from many real-world problems. However, most gradient-based optimization methods being applied to a non-convex function with many local minima is unprincipled. They face great challenges of finding the global minimum of the function. Because the information from the derivative of a single point is not sufficient to recognize the global geometrical property of the function.(1; 2; 3; 4) Formally, let (X, d) be a metric space and $T : X \to X$ is a self-mapping. For the inequality that,

$$d(T(x), T(y)) \leq qd(x, y), \forall x, y \in X. \tag{1}$$

if $q \in [0, 1)$, T is called contractive; if $q \in [0, 1]$, T is called nonexpansive; if $q < \infty$, T is called Lipschitz continuous(5; 6). The gradient-based methods usually belong to nonexpansive mapping. For nonexpansive mapping or Lipschitz continuous mapping, the change of range set can involve complicated deformation(7; 8). The fixed-point may exist but is not unique for a general situation. For instance, if the objective function $f(x)$ has many local minima, for the gradient descent method mapping $T(x) = x - \gamma\nabla f(x)$, any local minimum is a fixed-point. Hence, in the optimization problems, nonexpansive mapping suffers a significant deficiency that it would likely stop at some local minimum. Because both the existence and uniqueness of the solution are important in the optimization, the contractive mapping is more favored than the nonexpansive mapping.

The well-known Banach fixed-point theorem plays an important role in solving linear or nonlinear system. But the condition of contraction mapping that $d(T(x), T(y)) \leq qd(x, y)$ is hard to be directly applied to non-convex optimization problems. In this study, we are trying to extend the Banach fixed-point theorem to an applicable method for optimization problems, which is called strong contraction mapping. We will prove that strong contraction mapping admits a unique fixed-point, demonstrate how to build an optimization method with an $O(\frac{\log(\epsilon)-\log(D(X_0))}{\log(q)})$ rate as an application of it and illustrate its fixed-point is the global minimum point of the objective function.

2 Strong contraction mapping and the fixed-point

Let (X, d) be a metric space. Recall that the metric measurement $D(X)$ refers to the maximum distance between two points in the vector space X :(10)

$$D(X) := sup\{d(x, y), \forall x, y \in X\} \tag{2}$$

Definition 1. *Let (X, d) be a complete metric space. Then a mapping $T : X \to X$ is called strong contraction mapping on X if the range of mapping T is always a subset of its domain during the iteration, namely, $\mathcal{R}(T) = X_{i+1} \subseteq X_i$ and there exists a $q \in [0, 1)$ such that $D(X_{i+1}) \leq qD(X_i)$.*

The reader should be aware of the distinction between strong contraction mapping and Banach fixed-point theorem. This contraction mapping is called strong because the requirement $D(X_{i+1}) \leq q(X_i)$ contains $d(T(x), T(y)) \leq qd(x, y)$ what in Banach fixed-point theorem as a particular case, not requiring the distance between two points to get smaller but the diameter of the range of the mapping to get smaller.

Theorem 1. *Let (X, d) be a non-empty complete metric space with strong contraction mapping $T : X \to X$. Then T admits a unique fixed-point x^* in X such that $x^* = T(x^*)$.*

The theorem is proved by following lemmas.

Lemma 1.1. *Let $x_i = T(x_{i-1})$, then $\{x_i\}$ is a Cauchy sequence and converges to a limit x^* in X.*

Proof. Let $m, n \in \mathbb{N}$ such that $m > n$.

$$d(x_m, x_n) \leq D(X_n) \leq q^n D(X_0)$$

Let $\epsilon > 0$ be arbitrary, since $q \in [0, 1)$, we can find a large $N \in \mathbb{N}$ such that

$$q^N \leq \frac{\epsilon}{D(X_0)}.$$

Hence, by choosing m, n large enough:

$$d(x_m, x_n) \leq q^n D(X_0) \leq \frac{\epsilon}{D(X_0)} D(X_0) = \epsilon.$$

Because the x_0 doesn't show up in the inequality, as long as $D(X_0)$ is bounded, the convergence is guaranteed and independent of the choice of x_0. This inequality indicates the Cauchy sequence of strong contraction mapping converge to an ϵ-approximate point in $O(\frac{\log(\epsilon)-\log(D(X_0))}{\log(q)})$ iterates.

Since $\{x_i\}$ is a Cauchy sequence, it converges to a limit point:

$$x^* = \lim_{i\to\infty} x_i$$

Lemma 1.2. *x^* is a fixed-point of T in X.*

Proof.

$$x^* = \lim_{i\to\infty} x_i = \lim_{n\to\infty} T(x_{i-1}) = T(\lim_{n\to\infty} x_{i-1}) = T(x^*)$$

Thus,$x^* = T(x^*)$.

Lemma 1.3. *x^* is the unique fixed-point of T in X.*

Proof. Suppose there exists another fixed-point y that $T(y) = y$, then

$$y = \lim_{i\to\infty} T^i(y)$$

And the same rule applies to the fixed-point x^*. Thus, both the y and x^* should be elements of X_i, as i goes to infinity.

Consequently, the following inequality holds

$$0 \leq d(x^*, y) \leq D(X_i) \leq q^i D(X_0)$$

By the sandwich theorem,

$$0 \leq d(x^*, y) \leq \lim_{i\to\infty} D(X_i) \leq \lim_{i\to\infty} q^i D(X_0) = 0$$

$$d(x^*, y) = 0$$

Thus, x^* is the unique fixed-point of T in (X, d) as $x^* = y$.

3 Topological Non-convex Optimization Algorithm Implementation

Suppose the objective function is continuous, occupied with a unique global minimum, can be decomposed into locally convex components and is defined on non-empty complete domain X, $f : X \to R$. To overcome the dilemma that the optimization method may be saturated by local minima, intuitively, one can utilize a hyperplane(9) parallels to the domain to cut the objective function f such that the intersection between the hyperplane and the function forms contours. Imagine that as the hyperplane moving downwards the contours will eventually shrink to a point. While, the difficulty is how to iterate the hyperplane to move downwards and decide the position of the global minimum point.

As a numerical method, instead of getting all points of contours or symbolic expression of contours, we only want to get n number of roots on the contours using root-finding algorithm, that is

$$r_j^i = f^{-1}(L_i) = f^{-1}(f(x_i)), \; \forall j \in \{1, 2, ..., n\} \tag{3}$$

where, i indicates the ith iterate and j indicate the jth root.

Now, our task is to map to a point lower than the height of contours.

First, provide one arbitrary initial point x_0 to the function and calculate the height of contours $f(x_0) = L_0$ at the height of the initial point; If the objective function f is non-convex, as a consequence, to map to a point lower than the height of contours cannot be achieved by simply averaging the roots on contours. However, we can map to a point lower than the height of contours by averaging the roots belong to the same locally convex component and iterate hyperplane moving downwards.

Recall the property of convexity. Let a function $g : \mathcal{R}^n \to \mathcal{R}$ is convex, for every $r_1, r_2 \in \mathcal{R}^n$ and $\lambda \in [0, 1]$, we have the Inequality.4, (19)

$$g(\lambda r_1 + (1 - \lambda) r_2) \leq \lambda g(r_1) + (1 - \lambda) g(r_2) \tag{4}$$

Furthermore, the inequality $d(x_m, x_n) \leq q^n D(X_0)$ indicates the rate of convergence. The smaller q is, the higher rate of convergence is achieved. To achieve high rate of convergence, it is important to extend the Inequality.4 to the Jensen's Inequality.5, namely (20)

$$g(\lambda_1 r_1 + \lambda_2 r_2 + ... + \lambda_n r_n) \leq \lambda_1 g(r_1) + (1 - \lambda_1) g(\frac{\lambda_2 r_2 + ... + \lambda_n r_n}{1 - \lambda_1})$$

By induction,

$$g(\sum_{j=1}^{n} \lambda_j r_j) \leq \sum_{j=1}^{n} \lambda_j g(r_j) \tag{5}$$

where,

$$\sum_{j=1}^{n} \lambda_j = 1, \lambda_j \in [0, 1]$$

When we apply the Jensen's Inequality.5 to a locally convex component of the objective function f, let $\lambda_j = \frac{1}{n}$,for $j = 1, 2, ..., n$, then the Jensen's Inequality.5 turns to be a strict inequality,

$$f(\lambda_1 r_1 + \lambda_2 r_2 + ... + \lambda_n r_n) < \lambda_1 f(r_1) + \lambda_2 f(r_2) + ... + \lambda_n f(r_n) = f(x_i) = L_i \tag{6}$$

Let $x_{i+1} = \frac{1}{n}\sum_{j=1}^{n} r_j^i$,

$$f(x_{i+1}) = L_{i+1} < f(x_i) = L_i, \quad \forall i = 1, 2, ..., n \tag{7}$$

Therefore, it is important to get roots belong to the same locally convex subset. Based on the Inequality.4, one practical way to achieve that is to pair each two roots on contours and scan function's value along the segment between the two roots and check whether there exists a point higher than contours' height. N number of random points along the segment between two roots are checked whether higher than the contour's level(15; 16), if the Inequality.4 always holds for N times of test, then we believe the two roots locate at the same locally convex subset. Traverse all the roots and apply this examination on them, then we can decompose the roots with regards to different locally convex subsets. The average of roots in the same locally convex subset is always lower than the contours' height due to Jensen's Inequality.5. Hence, this method can be named as Jensen's Inequality descent.

After the set of roots are decomposed into several locally convex subsets, the averages of roots with regards to each subset are calculated and the lowest one is returned as an updated point per iterate. Thereafter, the remaining calculation is to repeat the iteration until convergence and return the converged point as the global minimum.

Rewrite the mapping of each iterate that $x_{i+1} = T(x_i)$, averaging the roots $r_1^i, r_2^i, ...r_n^i$ which locate at the same locally convex subset in the ith iterate,

$$x_{i+1} = T(x_i) = \frac{1}{n}\sum_{j=1}^{n} r_j^i = \frac{1}{n}\sum_{j=1}^{n} f^{-1}(f(x_i)) \tag{8}$$

Subsequently, like a Russian doll, the topological non-convex optimization algorithm can be explicitly written as an expansion of iterates as

$$\begin{aligned}
x_{i+1} &= T(x_i) = T^{i+1}(x_0) \\
&= \frac{1}{n}\sum_{j=1}^{n} r_j = \frac{1}{n}\sum_{j=1}^{n} f^{-1}(f(x_i)) \\
&= \frac{1}{n}\sum_{j=1}^{n} f^{-1}(f(\frac{1}{q}\sum_{j=1}^{q} f^{-1}(f(x_{i-1})))) \\
&\vdots \\
&= \frac{1}{n}\sum_{j=1}^{n} f^{-1}(f(\frac{1}{q}\sum_{j=1}^{q} f^{-1}(f(\ldots\frac{1}{k}\sum_{j=1}^{k} f^{-1}(f(x_0))\ldots))))
\end{aligned} \tag{9}$$

Now, we are going to prove that the iterating point calculated in this way converges to the unique global minimum of the objective function.

Proposition 1. *Provided x_{min} is the unique global minimum of the objective function f, then the fixed-point x^* of the strong contraction mapping T as discussed above coincides with the global minimum of the function.*

Proof. Since the iterating point x_{i+1} is always lower than x_i ,

$$0 \leq f(x_{i+1}) - f(x_{min}) < f(x_i) - f(x_{min})$$

Hence, there exists a $\xi \in (0, 1)$ such that,

$$0 \leq f(x_{i+1}) - f(x_{min}) < \xi^i(f(x_0) - f(x_{min}))$$

As i goes to infinity, then

$$
\begin{aligned}
0 \leq \lim_{i\to\infty} f(x_{i+1}) - f(x_{min}) &< \lim_{i\to\infty} \xi^i (f(x_0) - f(x_{min})) \\
\lim_{i\to\infty} f(x_{i+1}) - f(x_{min}) &= 0 \\
\lim_{i\to\infty} f(x_{i+1}) &= f(x_{min}) \\
f(\lim_{i\to\infty} x_{i+1}) &= f(x_{min}) \\
f(x^*) &= f(x_{min})
\end{aligned}
$$

Because the fixed-point x^* is at the same height as the global minimum x_{min} and the global minimum point is unique. Thus, the fixed-point x^* coincides with the global minimum point.(14)

In summary, the main procedure of topological non-convex optimization method is following steps: I. Given the initial guess point x_0 for the objective function, calculate the contour's level L_0; II. Solve the equation $f(x_i) = L_i$ and get n number of roots. Decompose the set of roots into several locally convex subsets, return the lowest average of roots as an update point from each iterate; III. Repeat the above iteration until convergence.

4 Experiments on Sphere, McCormick and Ackley functions

We conduct toy experiments on Sphere, McCormick and Ackley functions to test the non-convex optimization algorithm. The Sphere function is a good and simple example to start with because it is a convex function. We test the non-convex optimization algorithm with the initial point (1,1,1). The contour for the Sphere function in each iterate is a spherical surface as shown in the FIG.1. The red points are roots solved by the root-finding algorithm and the black spherical surface is the contour at their height. From left to right in the FIG.1, it shows the procedure that the contour moving downwards and shrinking to the global minimum. We present the average of roots and the height per iterate in TABLE.1.

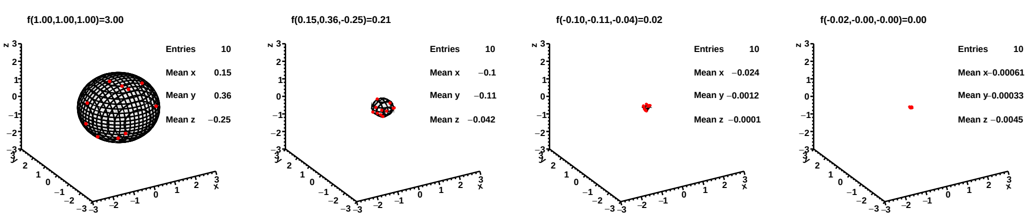

Figure 1: The red point markers are the roots and spherical surface is the contour is 3D space for each iteration. The intermediate steps illustrate how the contour shrink to a point during the procedure.

Table 1: It present the average of roots and the level of contour from each iterate when the optimization method is tested on Sphere function.

iterate	the average of roots	height of contour
0	(1.00,1.00,1.00)	3.0000
1	(0.15,0.36,-0.25)	0.2146
2	(-0.1,-0.11,-0.042)	0.0237
3	(-0.024,-0.0012,-0.0001)	0.0004
4	(-0.0061,-0.00033,-0.0045)	

Subsequently, we test the non-convex optimization algorithm on the McCormick function. And the first 4 iterates of roots and contours are shown in FIG.2 and the detailed iteration of the searching point from the numerical calculation is shown in TABLE.2.

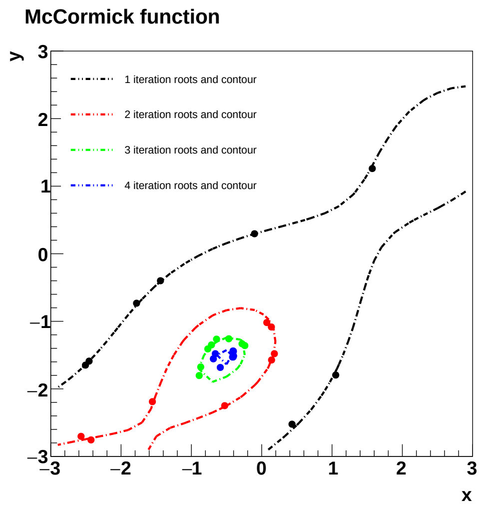

Figure 2: The point markers are the roots and the dash lines are the contours during iterates. The first 4 iteration results are drawn to illustrate the procedure of optimization test on the McCormick function.

Table 2: It present the average of roots and the level of contour from each iterate when the optimization method is tested on McCormick function.

iterate	the average of roots	height of contour
0	(2.00000,2.0000)	2.2431975047
1	(-0.6409,-0.8826)	-1.1857067055
2	(-0.8073,-1.8803)	-1.7770492114
3	(-0.5962,-1.4248)	-1.8814760998
4	(-0.4785,-1.5162)	-1.9074191216
5	(-0.5640,-1.5686)	-1.9125755974
6	(-0.5561,-1.5467)	-1.9131043354
7	(-0.5474,-1.5465)	-1.9132219834
8	(-0.5473,-1.5472)	

Ackley function is notorious for its many local minima and would be a nightmare for most gradient-based methods. We test the non-convex optimization algorithm on Ackley function starting from the point (2,2). As shown in the FIG.3, contours at each iterate are complicated and the average of roots is not guaranteed to be lower than contours' height. The method of decomposing them into different locally convex subsets discussed above is used to separate the set of roots and the lowest average is return as the updated searching point. We present the lowest average per iterate in TABLE.3.

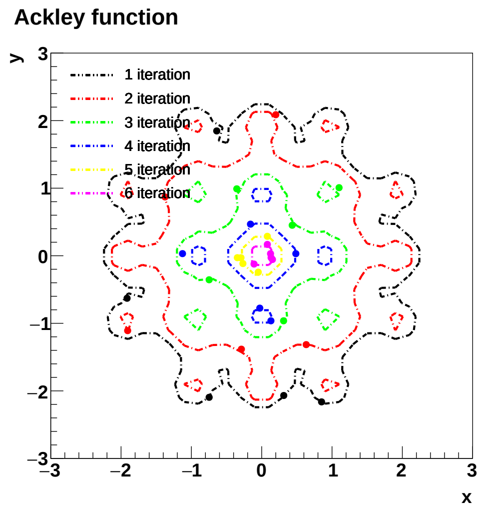

Figure 3: The point markers are the roots and the dash lines are contours during iterates. The first 6 iteration results are drawn to illustrate the procedure of optimization test on the Ackley function. The intermediate steps show how contours ignore the existence of local minima and safely approach to the global minimum point.

Table 3: It present the average of roots and the level of contour from each iterate when the optimization method is tested on Ackley function.

iterate	the average of roots	height of contour
0	(2.00000000,2.00000000)	6.59359908
1	(-0.78076083,-1.34128187)	5.82036224
2	(-0.35105371,-0.62030933)	4.11933422
3	(-0.20087095,0.38105138)	3.09359564
4	(0.06032320,-0.88101860)	2.17077104
⋮	⋮	⋮
15	(0.00000404,-0.00000130)	0.00001199
16	(-0.00000194,-0.00000079)	0.00000591
17	(-0.00000034,0.00000003)	

From the topological viewpoint, these experiments explain how the non-convex optimization algorithm works and observe contours' deformation during the hyperplane moving downwards that eventually converging to the global minimum point. This global minimum convergence empirically present that the proposed non-convex optimization method will outperform many gradient-based methods in global minimization tasks.

5 Conclusion

This paper conceptually introduces the strong contraction mapping, proves the existence and uniqueness of its fixed-point and illustrates the implementation of a non-convex optimization method. The decomposition of roots set provided a divide-and-conquer method to transfer the original problem to a number of subproblems and solve them recursively. The optimization method has been tested on Sphere, McCormick and Ackley functions and successfully converged to the global minimum. These experiments demonstrate the contours' deformation and the Cauchy sequence of iterating points approaching the global minimum. The global minimum convergence regardless of local minima and initial point position is a very significant strength of an optimization algorithm. We hope that this study can improve our understanding of common property of fixed-point theorems and provide us valuable insight and vista of the development of efficient optimization methods.

References

[1] Allen-Zhu, Zeyuan, and Elad Hazan. "Variance reduction for faster non-convex optimization." In International conference on machine learning, pp. 699-707. 2016.

[2] Arora, Sanjeev, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. "A practical algorithm for topic modeling with provable guarantees." In International Conference on Machine Learning, pp. 280-288. 2013.

[3] Arora, Sanjeev, Rong Ge, and Ankur Moitra. "New algorithms for learning incoherent and overcomplete dictionaries." In Conference on Learning Theory, pp. 779-806. 2014.

[4] Hsu, Daniel, Sham M. Kakade, and Tong Zhang. "A spectral algorithm for learning hidden Markov models." Journal of Computer and System Sciences 78, no. 5 (2012): 1460-1480.

[5] Husain, T., and Abdul Latif. "Fixed points of multivalued nonexpansive maps." International Journal of Mathematics and Mathematical Sciences 14.3 (1991): 421-430.

[6] Latif, Abdul, and Ian Tweddle. "On multivalued f-nonexpansive maps." Demonstratio Mathematica 32.3 (1999): 565-574.

[7] Ahues, Mario, Alain Largillier, and Balmohan Limaye. Spectral computations for bounded operators. Chapman and Hall/CRC, 2001.

[8] Rudin, Walter. "Functional analysis. 1991." Internat. Ser. Pure Appl. Math (1991).

[9] Brezis, Haim. Functional analysis, Sobolev spaces and partial differential equations. Springer Science Business Media, 2010.

[10] Croom, Fred H. Principles of topology. Courier Dover Publications, 2016.

[11] Kiwiel, K. C. (2001). Convergence and efficiency of subgradient methods for quasiconvex minimization. Mathematical programming, 90(1), 1-25.

[12] Banach, S. (1922). Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. Fund math, 3(1), 133-181.

[13] Branciari, A. (2002). A fixed point theorem for mappings satisfying a general contractive condition of integral type. International Journal of Mathematics and Mathematical Sciences, 29(9), .

[14] Taylor, A. E., & Lay, D. C. (1958). (Vol. 2). New York: Wiley.

[15] Schachter, Bruce. "Decomposition of polygons into convex sets." IEEE Transactions on Computers 11 (1978): 1078-1082.

[16] Tseng, Paul. "Applications of a splitting algorithm to decomposition in convex programming and variational inequalities." SIAM Journal on Control and Optimization 29.1 (1991): 119-138.

[17] Arveson, W. (2006). A short course on spectral theory (Vol. 209). Springer Science & Business Media.

[18] Blanchard, P.; Devaney, R. L.; Hall, G. R. (2006). Differential Equations. London: Thompson. pp. 96–111. ISBN 0-495-01265-3.

[19] Bertsekas, Dimitri (2003). Convex Analysis and Optimization. Athena Scientific.

[20] Chandler, David. "Introduction to modern statistical mechanics." Introduction to Modern Statistical Mechanics, by David Chandler, pp. 288. Foreword by David Chandler. Oxford University Press, Sep 1987. ISBN-10: 0195042778. ISBN-13: 9780195042771 (1987): 288.

[21] Cormen, Thomas H., Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. Introduction to algorithms. MIT press, 2009.