Homotopy Techniques in Solving Systems of Nonlinear Equations: A Theoretical Justification of Convex Combinations

Nicholas Sun
Department of Mathematics
Rutgers University
110 Frelinghuysen Road
Piscataway, NJ 08854-8019, USA
nicholas.sun@rutgers.edu

Abstract

One of the techniques for solving systems of non-linear equations $F_1(x_1, \ldots, x_n) = 0$, $\ldots$, $F_n(x_1, \ldots, x_n) = 0$ ($\vec{F}(\vec{x}) = \vec{0}$) is a *homotopy method*, when we start with a solution of a simplified (and thus easier-to-solve) approximate system $G_i(x_1, \ldots, x_n) = 0$, and then gradually adjust this solution by solving intermediate systems of equation $H_i(x_1, \ldots, x_n) = 0$ for an appropriate "transition" function $\vec{H}(\vec{x}) = \vec{f}(\lambda, \vec{F}(\vec{x}), \vec{G}(\vec{x}))$. The success of this method depends on the selection of the appropriate combination function $\vec{f}(\lambda, \vec{u}_1, \vec{u}_2)$. The most commonly used combination function is the *convex homotopy* function $\vec{f}(\lambda, \vec{u}_1, \vec{u}_2) = \lambda \cdot \vec{u}_1 + (1-\lambda) \cdot \vec{u}_2$. In this paper, we provide a theoretical justification for this combination function.

1 Formulation of the Problem

Need to solve systems of equations. In many practical situations, it is difficult (or even impossible) to directly measure the values of the desired physical quantities $x_1, \ldots, x_n$. For example, it is difficult to directly measure the distance to a faraway star or the temperature inside this star. To measure these quantities, we measure easier-to-measure quantities $y_1, \ldots, y_m$ which are related to the desired quantities $x_1, \ldots, x_n$ by known dependencies $y_i = d_i(x_1, \ldots, x_n)$. For example, to measure the distance to a star, we measure the angles in the direction to this star at two different moments of

time, when the Earth is at different parts of its solar orbit. To find the desired values x_i, we then need to solve the corresponding system of equations $F_i(x_1, \ldots, x_n) = 0$, where we denoted $F_i(x_1, \ldots, x_n) \stackrel{\text{def}}{=} d_i(x_1, \ldots, x_n) - y_i$.

To find the values of n unknowns $x_1, \ldots, x_n$, in general, we need n equations. The corresponding system of n equations $F_i(x_1, \ldots, x_n) = 0$, $1 \le i \le n$, can be also described in the vector form, as $\vec{F}(\vec{x}) = \vec{0}$.

Homotopy method for solving systems of equations: a brief reminder. The dependencies $d_i(x_1, \ldots, x_n)$ are often non-linear; as a result, the corresponding system of nonlinear equations is difficult to solve.

One of the methods of solving systems of nonlinear equations is the *homotopy* method; see, e.g., [1, 2, 3, 5]. The main idea behind this method is that it is usually easier to solve a system of equations if we know a good approximation to the solution; in this case, we can, e.g., linearlize the system, and solve the resulting linear system. One way to get a good approximation is to take a solution to the approximate system.

As a result, we start with a *simplified* easier-to-solve system $\vec{G}(\vec{x}) = \vec{0}$, and then form a continuous family of functions $\vec{H}(\lambda, \vec{x}) = \vec{f}(\lambda, \vec{A}(\vec{x}), \vec{B}(\vec{x}))$ with a parameter $\lambda \in [0, 1]$ that starts, for $\lambda = 0$, at the simplified system $\vec{H}(0, \vec{x}) = \vec{G}(\vec{x}) = \vec{0}$ and ends up, for $\lambda = 1$, at the desired system $\vec{H}(1, \vec{x}) = \vec{F}(\vec{x}) = \vec{0}$.

Let us select a sequence of real numbers $\lambda_0 = 0 < \lambda_1 < \ldots < \lambda_k = 1$ for which all the differences $\lambda_i - \lambda_{i-1}$ are small. Since the differences $\lambda_i - \lambda_{i-1}$ are small, we have $\lambda_i \approx \lambda_{i-1}$ for all i. Thus, the system corresponding to λ_{i-1} is a close approximation to the system corresponding to λ_i – and therefore, the solution of the system corresponding to λ_{i-1} is a good approximation to the solution of the system corresponding to λ_i.

This idea leads to the following algorithm.

- We start by finding a solution to the easy-to-solve simplified system corresponding to $\lambda_0 = 0$.
- Then, for each i, we use the solution corresponding to λ_{i-1} (which, as we have mentioned, is a good approximation to the system corresponding to λ_i) to solve the system corresponding to λ_i.

Once we get to the value $i = k$, we thus have a solution $\vec{x}$ to the desired system $\vec{H}(\lambda_k, \vec{x}) = \vec{H}(1, \vec{x}) = \vec{F}(\vec{x}) = \vec{0}$

Convex homotopy. The success of the homotopy method in solving systems of equations depends on the proper selection of a combination function

$\vec{f}(\lambda, \vec{u}_1, \vec{u}_2)$. The most widely used combination function is the function

$$\vec{f}(\lambda, \vec{u}_1, \vec{u}_2) = \lambda \cdot \vec{u}_1 + (1 - \lambda) \cdot \vec{u}_2 \tag{1}$$

known as the *convex homotopy* function.

Problem. In many cases, the convex homotopy function leads to a successful solution of the corresponding system of nonlinear equations. However, there seems to be no convincing theoretical explanation for this empirical success – and thus, it is not clear whether the convex homotopy function is indeed the best combination function or there may be other combination functions which are even better.

What we do in this paper. In this paper, we provide a possible theoretical justification for the empirical success of the convex homotopy function.

Comment. A similar justification for a homotopy method for solving optimization problems is described in [4].

2 Analysis of the Problem

We need a family of homotopy functions. As we can see from the description of the homotopy method, the exact parametrization of different functions $\vec{f}(\lambda, \vec{u}_1, \vec{u}_2)$ is not that important; what is important is that we have a family of functions $\vec{h}(\vec{u}_1, \vec{u}_2) \stackrel{\text{def}}{=} \vec{f}(\lambda, \vec{u}_1, \vec{u}_2)$ corresponding to different values of the parameter λ.

Reasonable properties of functions from the homotopy family must satisfy. What properties should these functions $\vec{h}(\vec{u}_1, \vec{u}_2)$ satisfy?

First, in the case when the mapping $\vec{F}(\vec{x})$ corresponding to the original system of equations is already simple, i.e., if $\vec{G}(\vec{x}) = \vec{F}(\vec{x})$ for all $\vec{x}$, then it is reasonable to require that the above procedure do not force us to perform any unnecessary job of solving any other system of equations. In other words, for each of the homotopy functions $\vec{h}(\vec{u}_1, \vec{u}_2)$, the resulting objective function $\vec{H}(\vec{x}) = \vec{h}(\vec{F}(\vec{x}), \vec{G}(\vec{x})) = \vec{h}(\vec{F}(\vec{x}), \vec{F}(\vec{x}))$ should coincide with $\vec{F}(\vec{x})$: $\vec{h}(\vec{F}(\vec{x}), \vec{F}(\vec{x})) = \vec{F}(\vec{x}))$.

This property should be satisfied for all possible values of $\vec{F}(\vec{x})$. Thus, we must have $\vec{h}(\vec{u}, \vec{u}) = \vec{u}$ for all vectors $\vec{u}$.

Another reasonable property is *continuity*: if we change the original system $\vec{F}(\vec{x}) = \vec{0}$ and/or the simplified system $\vec{G}(\vec{x}) = \vec{0}$ a little bit, this should lead to a small change in the new system $\vec{H}(\vec{x}) = \vec{0}$. In other words, the combination function $\vec{h}(\vec{u}_1, \vec{u}_2)$ should be continuous.

Finally, in many physical situations, while the vector $\vec{F}$ has a direct physical sense, the numerical values $F_1, \ldots$ depend on what coordinate system we choose. The simplest are linear coordinate transformations $\vec{F} \to \vec{T}(\vec{F})$. It is reasonable to require that the homotopy function does not change under such transformation.

In other words, if we apply the homotopy function $\vec{h}(\vec{u}_1, \vec{u}_2)$ to the two systems of equations described in the new coordinates, i.e., to the systems $\vec{T}(\vec{F}(\vec{x})) = \vec{0}$ and $\vec{T}(\vec{G}(\vec{x})) = \vec{0}$, then we should get the same system as when we apply this same homotopy function to the original systems $\vec{F}(\vec{x}) = \vec{0}$ and $\vec{G}(\vec{x}) = \vec{0}$ – except that it is now described in new coordinates as well. In precise terms, if $\vec{H}(\vec{x}) = \vec{h}(\vec{F}(\vec{x}), \vec{G}(\vec{x}))$, then we should have $\vec{T}(\vec{H}(\vec{x})) = \vec{h}(\vec{T}(\vec{F}(\vec{x})), \vec{T}(\vec{G}(\vec{x})))$.

This property must be true for all possible values of $\vec{u}_1 = \vec{F}(\vec{x})$ and $\vec{u}_2 = \vec{G}(\vec{x})$, and for all possible linear transformations. Thus, we conclude that for all $\vec{u}_1$, $\vec{u}_2$, and $\vec{T}$, if $\vec{u} = \vec{h}(\vec{u}_1, \vec{u}_2)$, then $\vec{T}(\vec{u}) = \vec{h}(\vec{T}(\vec{u}_1), \vec{T}(\vec{u}_2))$.

Now, we are ready to formulate our main result.

3 Definitions and the Main Result

Definition. *Let us call a continuous function $\vec{h}(\vec{u}_1, \vec{u}_2)$ of two vector variables a* reasonable homotopy function *if it is satisfies the following two properties:*

- $\vec{h}(\vec{u}, \vec{u}) = \vec{u}$ *for all vectors* $\vec{u}$*;*
- *for each pair of vectors $\vec{u}_1$, $\vec{u}_2$, and for each linear transformation $\vec{T}$, if $\vec{u} = \vec{g}(\vec{u}_1, \vec{u}_2)$, then $\vec{T}(\vec{u}) = \vec{g}(\vec{T}(\vec{u}_1), \vec{T}(\vec{u}_2))$.*

Comment. One can easily check that for every real number λ, the function $\vec{h}(\vec{u}_1, \vec{u}_2) = \lambda \cdot \vec{u}_1 + (1 - \lambda) \cdot \vec{u}_2$ is a reasonable homotopy function (in the sense of the above Definition). It turns out that these are the only reasonable homotopy functions.

Proposition. *Every reasonable homotopy function has the form*

$$\vec{h}(\vec{u}_1, \vec{u}_2) = \lambda \cdot \vec{u}_1 + (1 - \lambda) \cdot \vec{u}_2 \tag{2}$$

for some real number λ.

Discussion. This result provides the desired theoretical justification for the convex homotopy function.

Comment. We are analyzing *generic* homotopy functions, i.e., homotopy functions which can be applied to all possible systems of non-linear equations. For *specific* systems of equations, different homotopy functions – which take the specific character of the corresponding systems into account – are sometimes better than the convex one; see, e.g., [2, 3].

4 Proof

1°. Let us take two unit vectors $\vec{e}_1 \stackrel{\text{def}}{=} (1, 0, \ldots, 0)$ and $\vec{e}_2 \stackrel{\text{def}}{=} (0, 1, 0, \ldots, 0)$, and let us denote $\vec{w} \stackrel{\text{def}}{=} \vec{g}(\vec{e}_1, \vec{e}_2)$. Let us prove that for the vector $\vec{w} = (w_1, w_2, w_3, \ldots, w_n)$, only the first two components w_1 and w_2 may be different from 0, the rest are zeros.

Indeed, the invariance property implies that for any linear transformation $\vec{T}$, we have

$$\vec{T}(\vec{w}) = \vec{g}(\vec{T}(\vec{e}_1), \vec{T}(\vec{e}_2)). \tag{3}$$

In particular, this is true for the following linear transformation

$$\vec{T}(x_1, x_2, x_3, \ldots, x_n) \stackrel{\text{def}}{=} (x_1, x_2, -x_3, \ldots, -x_n).$$

By definition of this transformation $\vec{T}$, we have $\vec{T}(\vec{e}_1) = \vec{e}_1$ and $\vec{T}(\vec{e}_2) = \vec{e}_2$. Thus, formula (3) implies that $\vec{T}(\vec{w}) = \vec{g}(\vec{e}_1, \vec{e}_2)$. By definition of $\vec{w}$, this means that $\vec{T}(\vec{w}) = \vec{w}$, i.e., that

$$(w_1, w_2, w_3, \ldots, w_n) = (w_1, w_2, -w_3, \ldots, -w_n).$$

So, for every $i \geq 3$, we have $w_i = -w_i$ and therefore, $w_i = 0$.

2°. We have just proved that $\vec{h}(\vec{e}_1, \vec{e}_2) = (w_1, w_2, 0, \ldots, 0)$. In vector form, this formula can be represented as

$$\vec{h}(\vec{e}_1, \vec{e}_2) = w_1 \cdot \vec{e}_1 + w_2 \cdot \vec{e}_2. \tag{4}$$

3°. Let us now prove that if $\vec{u}_1 \not\parallel \vec{u}_2$, then

$$\vec{h}(\vec{u}_1, \vec{u}_2) = w_1 \cdot \vec{u}_1 + w_2 \cdot \vec{u}_2. \tag{5}$$

Indeed, from the formula (4), we conclude that for any linear transformation $\vec{T}$, we get $\vec{h}(\vec{T}(\vec{e}_1), \vec{T}(\vec{e}_2)) = \vec{T}(w_1 \cdot \vec{e}_1 + w_2 \cdot \vec{e}_2)$. Since the transformation $\vec{T}$ is linear, this implies that

$$\vec{h}(\vec{T}(\vec{e}_1), \vec{T}(\vec{e}_2)) = w_1 \cdot \vec{T}(\vec{e}_1) + w_2 \cdot \vec{T}(\vec{e}_2). \tag{6}$$

Since $\vec{u}_1 \not\parallel \vec{u}_2$, we can extend $\vec{u}_1, \vec{u}_2$ to a basis $\vec{u}_1$, $\vec{u}_2$, $\vec{u}_3$, $\ldots$, $\vec{u}_n$. We can then form the following linear transformation:

$$T(x_1, x_2, \ldots, x_n) = x_1 \cdot \vec{u}_1 + x_2 \cdot \vec{u}_2 + \ldots + x_n \cdot \vec{u}_n.$$

For this transformation $\vec{T}$, we have $\vec{T}(\vec{e}_1) = \vec{u}_1$ and $\vec{T}(\vec{e}_2) = \vec{u}_2$. Thus, the formula (6) takes the desired form (5).

4°. Let us prove that the same formula (4) also holds when $\vec{u}_1 \parallel \vec{u}_2$.

Indeed, let us take any vector $\vec{e} \not\parallel \vec{u}_1$. Then for every $\varepsilon > 0$, we have $\vec{u}_2' \stackrel{\text{def}}{=} (\vec{u}_2 + \varepsilon \cdot \vec{e}) \not\parallel \vec{u}_1$. So, due to Part 3 of our proof, we have $\vec{h}(\vec{u}_1, \vec{u}_2 + \varepsilon \cdot \vec{e}) = w_1 \vec{u}_1 + w_2 \cdot (\vec{u}_2 + \varepsilon \cdot \vec{e})$. Due to continuity, when $\varepsilon \to 0$, we get the desired formula $\vec{h}(\vec{u}_1, \vec{u}_2) = w_1 \cdot \vec{u}_1 + w_2 \cdot \vec{u}_2$.

5°. From Parts 3 and 4, we conclude that the formula (4) holds for all pairs of vectors $\vec{u}_1$ and $\vec{u}_2$.

In particular, when $\vec{u}_1 = \vec{u}_2 = \vec{u}$, we get $\vec{h}(\vec{u}, \vec{u}) = w_1 \cdot \vec{u} + w_2 \cdot \vec{u} = (w_1 + w_2) \cdot \vec{u}$. By definition of a reasonable homotopy function, we must have $\vec{h}(\vec{u}, \vec{u}) = \vec{u}$. Thus, $(w_1 + w_2) \cdot \vec{u} = \vec{u}$ for all $\vec{u}$, i.e., $w_1 + w_2 = 1$. Thus, $w_2 = 1 - w_1$.

So, if we denote $\lambda \stackrel{\text{def}}{=} w_1$, we get $w_1 = 1 - \lambda$. For these values w_i, the formula (4) becomes the desired formula (2). The proposition is proven.

Acknowledgments. This work was performed when the author was working at the the University of Texas at El Paso's National Center of Border Security and Immigration, a Center of Academic Excellence under the Department of Homeland Security.

The author would also like to thank Shahriar Hossain, Vladik Kreinovich, and Luc Longpré for their invaluable guidance.

References

[1] S. N. Chow, J. Mallet-Paret, and J. A. Yorke, "Finding zeros of maps: homotopy methods that are constructive with probability one", *Mathematics of Computation*, 1978, Vol. 32, pp. 887–899.

[2] D. R. Easterling, M. S. Hossain, L. T. Watson, and N. Ramakrishnan, "Probability-one Homotopy Maps for Tracking Constrained Clustering Solutions", *Proceedings of the Internatonal Socety of Modeling and Simulation 21st High Performance Computing Symposium HPC'13*, San Diego, California, April 7–10, 2013, Article No. 17.

[3] S. Ji, L. T. Watson, and L. Carin, "Semisupervised learning of hidden Markov models via a homotopy method", *IEEE Transactions on Pattern Analysis and Machine Intellidence*, 2009, Vol. 31, pp. 275–287.

[4] N. Sun, "Why Convex Homotopy Is Very Useful in Optimization: A Possible Theoretical Explanation", *Journal of Uncertain Systems*, 2015, Vol. 9, to appear.

[5] L. T. Watson, M. Sosonkina, R. C. Melville, A. P. Morgan, and H. F. Walker, "Algorithm 777: HOMPACK90: a suite of Fortran 90 codes for globally convergent homotopy algorithms", *ACM Transactions on Math. Software*, 1997, Vol. 23, pp. 514–549.