Convergence of Line Search A-Function methods

Dimitri Kanevsky, David Nahamoo, Tara Sainath, Bhuvana Ramabhadran

IBM T.J. Watson Research Center, Yorktown, NY 10598, USA
{kanevsky, nahamoo, tsainath, bhuvana}@us.ibm.com

Abstract

Recently, the Line Search A-Function (LSAF) was introduced as a technique that generalizes Extended Baum-Welch (EBW) algorithm for functions of continuous probability densities. It was shown that LSAF provides a unified scheme for a large class of optimization problems that involve discriminant objective functions of different probability densities or sparse representation objective functions such as Approximate Bayesian Compressive Sensing. In this paper, we show that a discrete EBW recursion (that was initially developed to optimize functions of discrete distributions) also fits the scope of LSAF technique. We demonstrate the utility and robustness of the technique for discrete distributions thru the experimental set up of a TIMIT phone classification task using a Convex Hull Sparse Representation approach with different Lq regularization (q being any positive number).

Index Terms: EBW, convergence, $\mathcal{A}$-function, LSAF

1. Introduction

The Extended Baum-Welch (EBW) technique was initially introduced for estimating the discrete probability parameters of multinomial distribution functions of HMM speech recognition problems under the Maximum Mutual Information discriminative objective function [1]. Later, in [9], EBW technique was extended to estimating the parameter of Gaussian Mixture Models (GMMs) of HMMs under the MMI discriminative function for speech recognition problems. In ([5]) the EBW technique was generalized to the novel Line Search A-functions (LSAF) optimization technique. In ([5]), a simple geometric proof was provided to show that LSAF recursions result in a growth transformation (i.e. the value of the original function increases for the new parameters values). In the paper we show that a discrete EBW that was invented more than 23 years ago can be also represented using $\mathcal{A}$-functions (like its continuous EBW version). We also for the first time give a convergence proof for a discrete EBW. This technique of this proof can be easily applied also for proving convergence of continuous EBW (that optimize functions of GMMs) and LSAF process. Finally, we provide experimental results for TIMIT phonetic classification task using fractional norm generalization of a recently introduced Convex-Hull sparse representation technique that deploys discrete EBW for optimization ([10]). The rest of the paper is structured as follows. In Section 2, we provide a short review of the LSAF method which was introduced in ([5]). In Section 3 we show that a discrete EBW can be represented in the LSAF framework, In Section 4 we outline a proof of convergence of a discrete EBW. In Section 5 we describe a discrete EBW recursion for optimization problems with fractional norms constraints. In Section 6 we demonstrate this techniques for phonetic tasks ([8]). In Section 7 we give conclusions and in Appendix we give a detailed proof of convergence for EBW.

2. LSAF

2.1. Definition

Let $f(x) : \mathcal{U} \subset \mathbb{R}^n \to \mathbb{R}$ be a real valued differentiable function in an open subset $\mathcal{U}$. Let $\mathbf{A}_f = \mathbf{A}_f(x,y) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be twice differentiable in $x \in \mathcal{U}$ for each $y \in \mathcal{U}$. We define $\mathbf{A}_f$ as an $\mathcal{A}$-function for f if the following properties hold.
1. $\mathbf{A}_f(x,y)$ is a strictly convex or strictly concave function of x for any $y \in \mathcal{U}$ (recall that twice differentiable function is strictly concave or convex over some domain if its Hessian function is positive or negative definite in the domain, respectively).
2. Hyperplanes tangent to manifolds defined by $z = g_y(x) = \mathbf{A}_f(x,y)$ and $z = f(x)$ at any $x = y \in \mathcal{U}$ are parallel to each other, i.e.

$$\nabla_x \mathbf{A}_f(x,y)|_{x=y} = \nabla_x f(x) \tag{1}$$

It was shown in ([5]) that a general optimization technique can be constructed based on $\mathcal{A}$-function. We formulated a growth transformation such that the next step in the parameter update that increases $f(x)$ is obtained as a linear combination of the current parameter values and the value that optimizes the $\mathcal{A}$-function, i.e. $\nabla_x \mathbb{A}_f(x,y)|_{x=\tilde{x}} = 0$. More precisely, we stated that $\mathcal{A}$-function gives a set of iterative update rules with a ”growth” property as the following. Let x_0 be some point in $\mathcal{U}$ and $\mathcal{U} \ni \tilde{x}_0 \neq x_0$ be a solution of $\nabla_x A(x,x_0)|_{x=\tilde{x}_0} = 0$ (it is the minimum of $\mathbf{A}_f(x,x_0)$ if $\mathbf{A}_f$ is concave and the maximum if $\mathbf{A}_f$ is convex). Let

$$x_1 = x(\alpha) = \alpha \tilde{x}_0 + (1-\alpha) x_0. \tag{2}$$

Then for sufficiently small $|\alpha| \neq 0$, $f(x(\alpha)) > f(x_0)$ where $\alpha > 0$ if $A(x,x_0)$ concave and $\alpha < 0$ if $A(x,x_0)$ convex.
We called this technique Line Search A-Function (LSAF).

2.2. Gaussian example

Let

$$\xi_t(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x_t-\mu)^2\}}{2\sigma^2} \right\}$$

be densities over a space $x_t \in \mathcal{X}$ with model parameters $\theta = \{\mu, \sigma\}$ and let us consider a function $f(\{\xi_t(\theta)\})$. Let $c_t(\theta) = \xi_t \frac{\partial f(\{\xi_t\})}{\partial \xi_t}$. Then

$$Assoc_f(\{\xi_t(\theta_0), \xi_t(\theta)\}) = \sum c_t(\theta_0) \log f(\{\xi_t(\theta)\})$$

is $\mathcal{A}$-function. This gives us the following updates for model parameters:

$$\hat{\mu} = \hat{\mu}(\alpha) = \alpha\tilde{\mu}_0 + (1-\alpha)\mu_0 \ , \ \hat{\sigma}^2 = \hat{\sigma}^2(\alpha) = \alpha\tilde{\sigma}_0^2 + (1-\alpha)\sigma_0^2 \tag{3}$$

where $\tilde{\mu}_0 = \frac{c_t(\theta_0)x_t}{\sum c_t(\theta_0)}$ and $\tilde{\sigma}_0^2 = \frac{c_t(x_t - \tilde{\mu}_0)^2}{\sum c_t}$. Since $Assoc_f(\{\xi_t(\theta_0), \xi_t(\theta)\})$ is an $\mathcal{A}$-function, updates (3) are growth for sufficiently small $|\alpha|$. It was shown in ([5]) that EBW recursions for functions of Gaussian densities belong to this family of transformations (3).

Copyright © 2011 ISCA

3. Discrete EBW

In this section we show that discrete EBW can be described using the LSAF framework. In what follow we describe the case of a single distribution but this technique is easily generalizable to several distributions.

Let $\mathcal{S} = \{\beta : \beta \in \mathbb{R}^n, \beta_i \geq 0, i = 1, ...n, \sum \beta_i = 1\}$ and $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function on some subset $X \subset \mathcal{S}$. We want to solve the following maximization problem for a function $f(\beta)$

$$\text{Maximize } \ f(\beta) \ \text{ such that } \ \beta \in \mathcal{S} \tag{4}$$

Let for $\beta \in X$ $a_i^k = \frac{\partial f(\beta^k)}{\partial \beta_i^k}, i = 1, ...n$. For any $D \in \mathbb{R}$ and $\beta^k \in \mathbb{R}^n$ such that $\sum_{j=1}^n a_j^k \beta_j^k + D \neq 0$ we define a recursion $T_D : \mathbb{R}^n \to \mathbb{R}^n$ as

$$\beta_i^{k+1} = T_D(\beta^k) = \frac{a_i^k \beta_i^k + D\beta_i^k}{\sum_{j=1}^n a_j^k \beta_j^k + D} \tag{5}$$

It was shown in [2] that for sufficiently large D if $\beta^{k+1} \neq \beta^k$ then $f(\beta^{k+1}) > f(\beta^k)$.

An $\mathcal{A}$-function $\mathbb{A}_f$ for the function f in (4) differentiable in some compact neighborhood $\mathcal{U} \subset X$ of a point $\beta_0 \in \mathcal{S}$ is defined as following:

$$\mathbb{A}_f(\beta_0, \beta) = \sum (c_i + \beta_{0i} D) \log \beta_i \tag{6}$$

where $c_i = c_i(\beta_0) = \beta_{0i} \frac{\partial f(\beta)}{\partial \beta_i}|_{\beta=\beta_0} = \beta_{0i} a_i(\beta_0)$ and D is any number such that $a_i(\beta) + D > 0$ for all i and any $\beta \in \mathcal{U}$ (such D can be found since f is differentiable in $\mathcal{U}$ and $\mathcal{U}$ is compact). To show that the function $\mathbb{A}_f(\beta_0, \beta)$ in (6) is $\mathcal{A}$-function one need to check (1) as following. Replace $\beta_n = 1 - \sum \beta_i$ in (4, 6), i.e. consider functions $g(\beta') = f(\beta_1, ..., \beta_{n-1}, 1 - \sum_1^{n-1} \beta_i)$, $\mathbb{A}_g(\beta_0; \beta') = \mathbb{A}_f(\beta_0, \{\beta_1, ..., \beta_{n-1}, 1 - \sum_1^{n-1} \beta_j)$ where $\beta' = \{\beta_1, ...\beta_{n-1}\}$. We have:

$$\frac{\partial \mathbb{A}_g(\beta_0, \beta')}{\partial \beta_i}|_{\beta_i = \beta_{0i}} = a_i(\beta_0) \frac{\partial f(\beta)}{\partial \beta_i}|_{\beta_i=\beta_{0i}} +$$

$$D\beta_{0i} \frac{\partial \log \beta_i}{\partial \beta_i}|_{\beta_i=\beta_{0i}} +$$

$$D(1 - \sum_1^{n-1} \beta_{0i}) \frac{\partial \log(1 - \sum_1^{n-1} \beta_i)}{\partial \beta_i}|_{\beta=\beta_0} = \frac{\partial g(\beta')}{\partial \beta'}|_{\beta_i' = \beta_{0i}'}$$

4. Outline of convergence proof

We consider the optimization problem (4) and its EBW recursion (5). We first define the convergence statement for this problem.

For this we note that, necessary and/or sufficient conditions for some point to be a local maximum of a general constrained problem require this point to be a critical point. The notion of a critical point usually requires the derivative of the Lagrangian (that is a weighted sum of the gradient of an objective function and gradients of constraints equations) to be equal to zero as well as a system of equalities and inequalities associated with constraint equations. It can be easily seen that if a critical point of the A-function results into a fixed point in the recursion formula, this critical point will also become the critical point of the objective function f. Since each $\mathcal{A}$-function at each iteration is convex with a single critical point, at each iteration n we create a new β_n which is a critical point of the $\mathcal{A}$-function at that iteration. The sequence of $\beta_1, \beta_2, ..., \beta_n$ has a set of limit points since for discrete EBW the sequence of β belongs to a bounded set (a simplex). Based on Lemma 2 in Appendix, we know that each limit point is a fixed point with respect to the recursion formula. Depending then on regularity conditions at a critical point we get either necessary or/and sufficient conditions for this point to be a local extremum.

In our special case of a discrete EBW we use the following approach. First, we notice that if there exists some entry $\beta_i^k = 0$ then for all consequent updates we have $\beta_i^r = 0, r > k$. Therefore we consider a starting point β^0 for the problem (4) such that all $\beta_i > 0$. And for each k-th iteration (5) we consider D so large to ensure that $\beta_i^{k+1} > 0$ for all i. This means that if the problem (4) has a local maximum with some entries equal zero then these entries should be obtained as entries of a limit of an infinite sequence of updates $\beta^k, k = 1, 2....$ Second, the function f in (4) may not be defined everywhere for $\beta \in \mathcal{S}$, for example if it has poles in $\mathcal{S}$. In this case our "convergence" strategy would be to start with a point $\beta^0 \in \mathcal{S}$ over which f is defined and choose D at each iteration in such a way that an updated parameters belong to an area where f is defined. This process may lead to an infinite sequence of updated parameters that have limit points over which f is not defined (for example, this updated sequence converge to a pole, i.e. to a point where f has the "infinite" value).

Next, to define Lagrangian for the problem (4) we consider only an equality constraint $\sum \beta_i = 1$ to define critical points (of a Lagrangian). We then show that the update sequence to this problem has limits points that are critical points of this constrained problem. We also show that each update in this sequence satisfies inequalities $\beta \in \mathcal{S}_0 = \{\beta \in \mathcal{S}, \beta_i > 0, i = 1, ...n\}$. Therefore limits points of this sequence still belong to the domain $\mathcal{S}$. In Appendix we give a detailed proof following this strategy.

5. EBW method for fractional norms

We now show that discrete EBW methods can be applied to optimization of objective functions with fractional norm constraints as was suggested in ([7]):

$$\max f(\{\beta_i\}) \ \text{ s.t. } \ ||\beta||_q = 1 \ \text{ and } \ \beta_i \geq 0, i = 1....n \tag{7}$$

where $||\beta|_q = (\sum \beta_i^q)^{1/q}$. Changing variables and function as in

$$\gamma_i = \beta_i^{1/q} \quad g(\{\gamma_i\}) = f(\{\beta_i\}) \tag{8}$$

transforms the problem (7) into a discrete EBW problem for which recursion (5) could be applied. We will demonstrate this method for phonetic classification tasks using the objective function from ([10])

$$F(\beta) = -1/2(y - H\beta)^T R^{-1}(y - H\beta) +$$

$$\sum_{t=1}^{N_{classes}} \sum_{i=1}^{C_t} w_{it} \exp\left(-1/2 \times \frac{(\mu_{it} - H\beta)^T(\mu_{it} - H\beta)}{\sigma_{it}^2}\right) \tag{9}$$

Here the unique GMM classes are denoted by t, C_t is the number of Gaussians belonging to class t, and w_{it}, μ_{it} and σ_{it} are the weight, mean and variance parameters for Gaussian i in GMM t.

6. Experiments

Classification experiments are conducted on TIMIT acoustic phonetic corpus using feature representation that is similar to the classification work in ([8], [10]). We performed the following three experiments on functions (9) using the TIMIT dataset with constraints $\sum \beta_i^q = 1$ for three values $q = 0.5, 0.75, 1$. We run 60 iterations of a (5) for (7) (with transformation of parameters as described in (8)). We chose $D = 100$ at the first iteration and changed it at each iteration by multiplying by 1.02. Dimension of β in our experiments was 200. For each q we chose the uniform initial β with entries $1/200^{1/q}$. Our plots for one phone (Figure 1) demonstrate that we have the robust method to update model parameters and obtain a sparse representation. We also observe that the sparsity degree (i.e. number of near zero entries) of β decreases with the growth of q as expected.

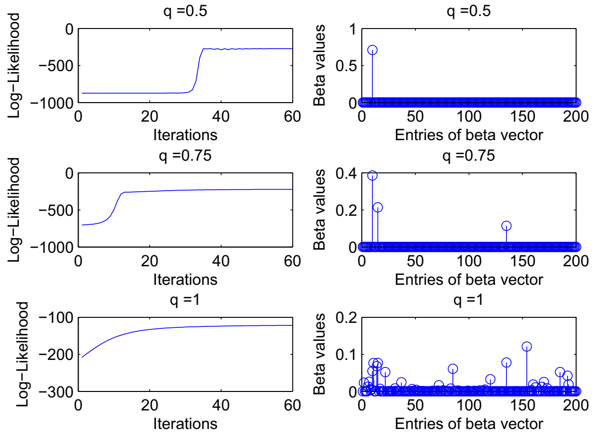

Fig. 1. Likelihood functions and sparsity representation for β for different values of q for 60 iterations

We also described in Table 1 the accuracy results on the test Timit set. The classification rule in this experiment is a combination of residual error between $y - H\beta$ and a GMM score. Number of iterations that is given in each column as iter # and D were tuned on a development set to optimize accuracy. The accuracy result for $q = 1$ coincides with the result that was reported in ([10]) and is the best reported number to date.

Table 1: *Accuracy of SR method. TIMIT test set*

Lq	q=0.5/iter 6	q=.075/iter 4	q=1.0/iter 15
Accuracy	81.93	82.95	85.11

7. Conclusions

In this paper, we showed that the discrete EBW method can be formulated in the LSAF framework. We provided a convergence proof for the discrete EBW recursion scheme. We showed how discrete EBW technique can be used to solve sparse representation problems with fractional norm constraints. By applying the technique to TIMIT phonetic classification task, we demonstrated that our method gives robust results, i.e. objective function values have a steady growth and we get a sparse representation. Finally, we believe that we can generalize our convergence proof of the discrete EBW to the broad LSAF case with the goal of presenting the generalized proof in a later paper.

8. References

[1] P.S. Gopalakrishnan, D. Kanevsky, D. Nahamoo and A. Nadas, "An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems", IEEE Trans. Information Theory, 1991, vol 37(1)

[2] D. Kanevsky,"Extended Baum Transformations for General Functions", in Proc. ICASSP, 2004

[3] D. Kanevsky, "Extended Baum Transformations For General Functions, II", IBM Research Report IBM, 2005. RC23645(W0506-120)

[4] D. Kanevsky, T. Sainath, B. Ramabhadran and D. Nahamoo, "Generalization of Extended Baum-Welch Parameter Estimation for Discriminative Training and Decoding", in Proc. Interspeech, 2008

[5] D. Kanevsky, D. Nahamoo, T. Sainath and B. Ramabhadran and ,"A-Functions: A Generalization of Extended Baum-Welch Transformations to Convex Optimization", in Proc. ICASSP, 2011

[7] A. Carmi, P. Gurfil, D. Kanevsky and B. Ramabhadran, "Extended Compressed Sensing: Filtering Inspired Methods for Sparse Signal Recovery and Their Nonlinear Variants", IBm Research Report, RC 24785 (W0903.015) March 4, 2009

[8] Sainath, T. N., Carmi, A., Kanevsky, D., and Ramabhadran, "Bayesian Compressive Sensing for Phonetic Classification". In Proc. ICASSP, 2010 small vocabulary, continuous speech recognition, in Proc. ICASSP, 1991.

[9] D. Povey, "Discriminative Training for Large Vocabulary Speech Recognition", Ph.D. thesis, University of Cambridge, 2003

[10] T. N. Sainath, D. Nahamoo, D. Kanevsky and B. Ramabhadran, "Convex Hull Sparse Representation Modeling Technique for Exemplar-Based Speech Recognition", IBM Research Report, 2011, March

9. Appendix: Convergence proof

For any map T we say that β is fixed if $T(\beta) = \beta$.

Lemma 1 *If β is fixed for T_D then β is a critical point of the Lagrange function*

$$\Lambda(\beta, \lambda) = f(\beta) - \lambda(\sum \beta_i - 1) \qquad (10)$$

for the constrained problem

$$Maximize\ \ f(\beta)\ \ such\ that\ \sum \beta_i = 1 \qquad (11)$$

In addition, if the Hessian of f at $\beta \in \mathcal{S}_0 = \{\beta \in \mathcal{S}, \beta_i > 0$ for all $i\}$ is negative definite then β is a local maximum of the problem (11)

Proof
If

$$\beta_i = T_D(\beta) = \frac{a_i\beta_i + D\beta_i}{\sum_{j=1}^{n} a_j\beta_j + D} \qquad (12)$$

then $\beta_i \sum_{j=1}^{n} a_j\beta_j = a_i\beta_i$, i.e.

$$\sum_{j=1}^{n} a_j\beta_j = a_i \qquad (13)$$

Next, $\nabla_\beta \Lambda(\beta,\lambda) = 0$ implies that $\lambda = a_i$ for all i or $\lambda = \sum a_i\beta_i$ and this coincides with (13), i.e. with (12). The condition $\nabla_\lambda \Lambda = 0$ is exactly the constraint $\sum \beta_i = 1$. Next, the Hessian for the Lagrangian (10) coincides with the Hessian of the function f at a point β that belongs to an interior of $\mathcal{S}$, i.e. β is the local maximum of f. This completes the proof of the lemma.

Lemma 2 *Let $T : X \to X$ be a continuous map of a compact set X such that $f(T(\beta)) > f(\beta)$ if $T(\beta) \neq \beta$. Then all limits points of a sequence $T^i(\beta)$ are fixed points of T.*

Proof Let a be a limit point of $T^{n_i}(\beta), i = 1, 2, ...$ Then $b = T(a)$ is a limit point for the sequence $T^{n_i+1}(\beta)$. If $b \neq a$ then we have

$$f(T^{n_i}(\beta)) \leq f(T^{n_i+1}(\beta)) \leq f(T^{n_{i+1}}(\beta)) \qquad (14)$$

I.e. $F(a) \leq F(b) \leq F(a)$. This implies $a = b$. Q.E.D.

Lemma 3 *Let $f(\beta)$ be continuously differentiable on a compact subset $X \in \mathbb{R}^n$. Then for any $q < 1$ there exist a constant D_0 such that for any $\beta \in X$ and $D \geq D_0$*

$$|T_D(\beta) - \beta| < q|\beta| \qquad (15)$$

Proof
We have

$$T_D(\beta) = \frac{a_i\beta_i + D\beta_i}{\sum_{j=1}^{n} a_j\beta_j + D} = \frac{1/Da_i\beta_i + \beta_i}{1/D\sum_{j=1}^{n} a_j\beta_j + 1} =$$

$$\beta_i(1 + a_i/D)(1 - 1/D\sum_{j=1}^{n} a_j\beta_j) + O(1/D^2) =$$

$$\beta_i + \beta_i \frac{a_i - \sum_{j=1}^{n} a_j\beta_j}{D} + O(1/D^2) \qquad (16)$$

Therefore, there exists a sufficiently large D_β such that

$$|T_D(\beta) - \beta| < q_\beta|\beta| \qquad (17)$$

for any $D \geq D_\beta$ where $q_\beta < 1$. Since f is continuously differentiable, there exist a small ball $U_\beta \subset \mathbb{R}^n$ with the center at β such that for any $\beta' \in U_\beta$

$$|T_D(\beta') - \beta'| < q_\beta|\beta'| \qquad (18)$$

Since X is compact it can be covered by a finite number of such balls U_β and therefore one can choose D_0 and q satisfying the lemma conditions.

Lemma 4 *Let $f(\beta)$ be continuously differentiable on a compact subset $X \in \mathbb{R}^n$. There exist a constant D_0 such that for any $\beta \in X$ and $D \geq D_0$*

$$f(T_D(\beta)) > f(\beta) \;\; \textit{if} \;\; T_D(\beta) \neq \beta \qquad (19)$$

Proof If $T_D(\beta) = \beta$ for some D then β is a critical point of Lagrangian $\mathcal{L}(\beta, \lambda) = F - \lambda(\sum_i \beta_i - 1)$ and for any D for which $T_D(\beta)$ is defined $T_D(\beta) = \beta$. Otherwise it was shown in [2] that there exists such large D that $f(T_D(\beta)) > f(\beta)$. Since f is continuous differentiable one can find a small open ball U_β of β such that $F(T_{D_\beta}(\beta')) > f(\beta')$ for any $\beta' \in U_\beta$. Since X is compact it can be covered by a finite set of open balls U_i such that for each ball U_i there is a D_i for which the lemma holds. Taking an upper limit of finite number of D_i proves the lemma.

Theorem 1 *Let $f(\beta)$ be continuously differentiable on $X \subset \mathcal{S}$. Let $X = \bigcup C_i, i = 1, 2, ...,$where C_i are compact and $C_i \subset C_{i+1}$. Assume also that for each C_i and any point $\beta \in C_i$ there is an open ball $U_i \in \mathbb{R}^n$ at some center u_i such that $\beta \in U_i \bigcap \mathcal{S} \subset C_i$. Assume that for each $\beta \in X$ there exists D_β such that for all $D \geq D_\beta$ $T_D(\beta) \in X$. Then there exists a transformation $T : X \to X$ such that the following holds:*

1. *$T(X) \subset X$*
2. *For any $\beta \in X$ such that $T(\beta) \neq \beta$ we have: $f(T(\beta)) > f(\beta)$*
3. *All limit points of a sequence $\{T^n(\beta)\}$ are fixed points.*

In addition, if a fixed point belongs to an interior of $\mathcal{S}$ and the Hessian of f at this fixed point is negative definite then this fixed point is a local maximum for f.

Proof
Let define T inductively as follows. Let D_1 be such that for any $D \geq D_1$ and $\beta \in C_1$ the following holds: $T_D(C_1) \subset X$, $f(T_D(\beta))) > f(\beta)$. Such D_1 exist by lemmas 3 and 4. Then for any $\beta \in C_1$ we define $T = T_1(\beta) = T_{D_1}(\beta)$. Assume that we defined T_k for C_k. And let us define T on C_{k+1} as following. Let D_{k+1} be such that for any $D \geq D_{k+1}$ and $\beta \in C_{k+1}$ we have $T_D(C_{k+1}) \subset X$, $f(T_D(\beta))) > f(\beta)$. Then we define $T_{k+1}(\beta) = T_{D_{k+1}}(\beta)$ for $\beta \in C_{k+1} \setminus C_k$ and for $\beta \in C_k$ we define $T_{k+1}(\beta) = T_k\beta$. Since for any $i < k$ $T_k(\beta) = T_i(\beta)$ if $\beta \in C_i$ all T_k give rise to the uniquely defined map $T(\beta) = T_k(\beta)$ for $\beta \in C_k, k = 1, 2, ...$ The 3d statement of the theorem follows from Lemma 2. And the last statement of the theorem about local maximum follows from Lemma 1.

Corollary 1 *Let f be continuously differentiable everywhere in $\mathcal{S}$, except of finite number of points $\mathcal{P}$. Then the following holds for any $\beta \in \mathcal{S}_0 \setminus \mathcal{P}$*

1. *$f(T^n(\beta)) > f(T^{n-1}(\beta))$ if $T^n(\beta) \neq T^{n-1}(\beta)$*
2. *All limit points of $T^n(\beta), n = 1, 2, ...$ are either critical points of the function f or belong to $\mathcal{P}$.*

In addition, let a limit point β^* of $T^n(\beta), n = 1, 2, ...$ belongs to $\mathcal{S}_0$ and does not belong to $\mathcal{P}$. Let the Hessian of f at this point β^* is negative definite then this point β^* is the local maximum of f .

Proof
The corollary follows from the fact that the set $X = \mathcal{S} \setminus \mathcal{P}$ can be represented as a union of compact sets C_i as described in Theorem 1. Namely, one can inductively build this set C_i as follows. For any point $\beta_1 \in X$ one can find an open ball U_1 with a center at β_1 such that this ball does not intersect $\mathcal{P}$. Let C_1 will be the closure of this ball. Next, we take any point $\beta_2 \in X \setminus \mathcal{P}$ that lies outside of C_1 and find another open ball U_2 at the center of β_2 that does not intersect C_1 and $\mathcal{P}$. Then we define C_2 as union of C_1 and the closure of U_2. Assume that we already constructed C_k. Then we can for any point β_{k+1} find a ball U_{k+1} at the center of β_{k+1} such that U_{k+1} does not intersect C_k and $\mathcal{P}$. Then C_{k+1} is defined as a union of the closure of U_{k+1} and C_k. Continue this process we build a set of C_k satisfying the conditions of Theorem 1 thereby proving the corollary.