A Proof of Local Convergence for the Adam Optimizer

Sebastian Bock
Faculty of Computer Science and Mathematics
OTH Regensburg
Regensburg, Germany
sebastian2.bock@oth-regensburg.de

Martin Weiß
Faculty of Computer Science and Mathematics
OTH Regensburg
Regensburg, Germany
martin.weiss@oth-regensburg.de

***Abstract*—Adaptive Moment Estimation (Adam) is a very popular training algorithm for deep neural networks, implemented in many machine learning frameworks. To the best of the authors knowledge no complete convergence analysis exists for Adam. The contribution of this paper is a method for the local convergence analysis in batch mode for a deterministic fixed training set, which gives necessary conditions for the hyperparameters of the Adam algorithm. Due to the local nature of the arguments the objective function can be non-convex but must be at least twice continuously differentiable.**

***Index Terms*—Non-convex optimization, Adam optimizer, convergence, momentum method, dynamical system, fixed point**

I. Introduction

Many problems in machine learning lead to a minimization problem in the weights of a neural network: Consider e.g. training data $(x_1, y_1), \ldots, (x_N, y_N)$ consisting of inputs x_i and outputs y_i, and the task to determine a neural network that has learned the relationship between inputs and outputs. This corresponds to a function $y = F(w, x)$, parametrized by the weights w, which minimizes the average loss function

$$f(w) = \frac{1}{N} \sum_{i=1}^{N} l(x_i, y_i, w) =: \frac{1}{N} \sum_{i=1}^{N} f_i(w)$$

over the training data. Typically the loss is built using some norm for regression problems, e.g. $l(x, y, w) = \frac{1}{2} \|y - F(w, x)\|_2^2$, or using cross entropy for classification. Optimization algorithms construct a sequence $\{w_t\}_{t \in \mathbb{N}_0}$ starting from an initial value w_0, which under appropriate assumptions converges to some local minimum $w_\star$ for general non-convex f. The most simple optimization algorithm for differentiable f is gradient descent with the update $w_{t+1} = w_t - \alpha \nabla f(w_t)$ and a learning rate $\alpha > 0$. For convex f conditions on the Lipschitz constant L of ∇f guarantee convergence and give estimates for the rate of convergence, see [1]. However L is hard to get in practice, and choosing α too big leads to oscillatory behaviour. Besides it is well known from optimization that the gradient is not the only optimum descent direction for f, see [2], but computation of the Hessian is usually prohibitive. This has led to the development of a family of algorithms which compute moments of first order, that is approximate descent directions based on previous iterates of the gradient like the initial momentum method [1], as well as second order moments to control the componentwise scaling and / or to adapt the learning rate in AdaGrad [3] and Adam [4]. More algorithms exist with variants like: batch mode vs. online or incremental mode – using $\nabla f(w_t)$ in iteration t vs. $\nabla f_k(w_t)$ where k iterates in a cyclic fashion over $1, \ldots, N$, or deterministic vs. stochastic choice of the index k for $\nabla f_k(w_t)$, stochastic assumptions for the observation of $\nabla f(w_t)$ or $\nabla f_k(w_t)$, and so on.

However for most of these algorithms only partial convergence results are known. The original proof of [4] is wrong as has been noted by several authors, see [5], [6]. Modifying the algorithm to AMSGrad, [7] establishes bounds on $\|\nabla f(w_t)\|$, similar to the results in [8] for a class called Incremental Generalized Adam. Though none of the results shows convergence of the sequence $\{w_t\}_{t \in \mathbb{N}_0}$. Also the proofs are lengthy and hardly reuse results from each other, giving not much insight. General results from optimization cannot be used for several reasons: First, the moments usually cannot be proven to be a descent direction. Second, the learning rate cannot be shown to be a step size valid for the Wolfe conditions for a line search, see [2]. The algorithm for the step taken in iteration t may explicitly contain the variable t in much more complicated ways than $\frac{1}{t}$ in the Robbins-Monro approach [9].

The contribution of this paper is a generally applicable method, based on the theory of discrete time dynamical systems, which proves local convergence of Adam. The results are purely qualitative because the results hold for learning rates sufficiently small, where "sufficiently small" is defined in terms of the eigenvalues of the Hessian in the unknown minimum $w_\star$.

II. Fixed Point Analysis under Perturbation

A. Notation

With Mat_n we denote the set of all real n-by-n matrices. The symbol $\perp$ denotes the transpose of a vector or matrix. With $\otimes, \oplus$ and $\oslash$ we denote the component-wise multiplication and division of vectors, as well as component-wise addition of vectors and scalar. For $f : \mathbb{R}^n \to \mathbb{R}$ the gradient

This paper presents results of the project "LeaP – Learning Poses" supported by the Bavarian Ministry of Science and Art under Kap. 15 49 TG 78.

and Hessian are written as ∇f and $\nabla^2 f$, provided they exist. Throughout this paper we assume $f : \mathbb{R}^n \to \mathbb{R}^n$ at least continuously differentiable, twice continuously differentiable for some results. The open ball with radius r around $x \in \mathbb{R}^n$ is denoted by $B_r(x) = \{y \in \mathbb{R}^n : \|y - x\| < r\}$ and $\|x\|$ is any norm. We denote $\rho(A) = \max\{|\lambda| | \lambda \text{ eigenvalue of } A\}$ the spectral radius of a matrix A and $\text{diag}(v) \in Mat_n$ describes a matrix with components of $v \in \mathbb{R}^n$ on the diagonal.

B. Related Work

Stochastic gradient descent (SGD) becomes an effective method for optimizing noisy tasks. Especially in the area of neural networks SGD variants are partly responsible for big successes in the last years, see e.g. [10] or [11].
Popular first-order SGD methods are AdaGrad [3] and RMSProp [12]. Kingma and Ba combine the advantages of these two methods and introduce the Adaptive Moment Estimation (Adam) in [4] (see Algorithm 1). Unfortunately,

Algorithm 1 Adam Optimizer

Require: $\alpha \in \mathbb{R}^+$, $\epsilon \in \mathbb{R}$ $\beta_1, \beta_2 \in (0,1)$, $w_0 \in \mathbb{R}^n$ and the function $f(w) \in C^2(\mathbb{R}^n, \mathbb{R})$

$m_0 = 0$, $v_0 = 0$, $t = 0$
while w not converged **do**
  $m_{t+1} = \beta_1 m_t + (1 - \beta_1)\nabla_w f(w_t)$
  $v_{t+1} = \beta_2 v_t + (1 - \beta_2)\nabla_w f(w_t) \otimes \nabla_w f(w_t)$
  $w_{t+1} = w_t - \alpha \frac{\sqrt{1-\beta_2^{t+1}}}{(1-\beta_1^{t+1})} m_{t+1} \oslash \sqrt{v_{t+1} \oplus \epsilon}$
  $t = t + 1$
end while

the Adam optimizer is not always defined the same way. Kingma and Ba [4] use $\sqrt{v} \oplus \epsilon$ and the bias correction in $\hat{m}$ and $\hat{v}$. The authors in [7] and [8] do not use an ϵ as well as [13], but the latter initialize $v_0 = \epsilon$. All three apply the bias correction in the learning rate α_t. We use the bias correction as described in [4, Section 2] and $\sqrt{v \oplus \epsilon}$ as used in [14]. The differences between the two possible usages of ϵ are minimal (see figure 1[1]) especially in the area around $v \approx 0$ by choosing ϵ as the square root of ϵ from [4]. Some differences in figure 1 are due to the dropout layer, which randomly eliminates neurons in the training to avoid overfitting. The main aim by the introduction of ϵ – avoiding division by 0 – holds in both variants, but $\sqrt{v \oplus \epsilon}$ gives the additional advantage of making the right hand side continuously differentiable for $v \in [0, \infty)$ whereas $\sqrt{v} \oplus \epsilon$ is not differentiable at $v = 0$. Differentiability will be essential in our proof.

During the last years the Adam Optimizer has become one of the most used optimization methods for training neural networks. Even if it is apparently working, there is, to the best of our knowledge, still no convergence proof for Adam. The proof in the original paper [4] was shown wrong, see

[1]The experiment is programmed with Keras 2.2.4, Tensorflow 1.11.0 and Python 3.6.

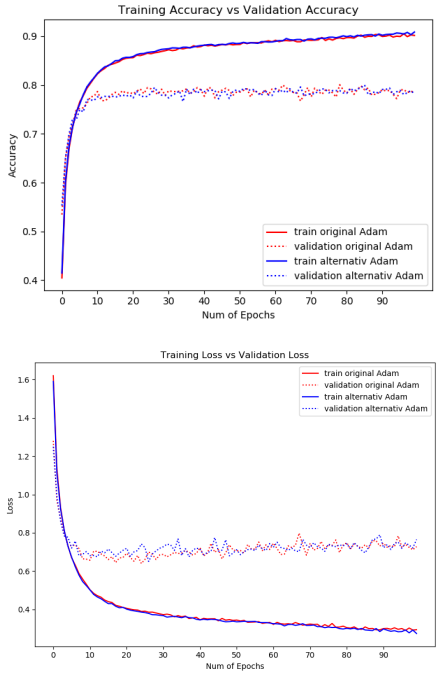

Fig. 1. Training the cifar-10 dataset ([15]) with the two different Adam methods

[5], [16] or [6]. Reddi et. al. in [7] present even a counter example and also introduce an improved method called AMSGrad. However in experiments AMSGrad does not show this improvements. On the contrary, in some cases it ends up with worse accuracy than Adam. Chen et al. [8] are showing for non-convex f $\min_{t=1}^{T} E[\|\nabla f(x_t)\|^2] = O\left(\frac{s_1(T)}{s_2(T)}\right)$. With the assumption of $s_1(T)$ growing slower than $s_2(T)$, one reaches a minimum of $E[\|\nabla f(x_t)\|^2]$ but without a guarantee of staying there. [17] use a similar interpretation of the Adam optimizer with a dynamical system viewpoint. But Barakat and Bianchi uses a non-autonomous ordinary differential equation without Lipschitz condition because of the term $\sqrt{v} \oplus \epsilon$ instead of our non-autonomous system of difference equations. In our opinion, this choice makes the proof longer and more complicated.

In the current work we present a convergence proof of the Adam optimizer [4] in a complete batch mode. This means the batch size is equal to the amount of training data. With this assumption we can guarantee, that we are searching the same minimum $w_\star$ in each time step t. Due to the local nature our proof does not assume the convexity of $f(w)$, thus we can guarantee local convergence even for non-convex settings. The hyperparameter setting is only bounded by

$$\frac{\alpha \max_{i=1}^{n}(\mu_i)}{\sqrt{\epsilon}}(1 - \beta_1) < 2\beta_1 + 2 \tag{1}$$

with μ_i the i-th eigenvalue of the Hessian $\nabla^2 f(w_\star)$. The counter example in [7] does not affect our convergence proof,

because we consider batch mode only; the incremental function in [7] becomes a linear function in batch mode.

C. Idea

We consider the learning algorithm from the standpoint of dynamical systems and define a common state vector x consisting of the moments – like m and v for Adam – and the weights, so we have $x = (m, v, w)$. Then the optimization can be written as an iteration $x_{t+1} = T(t, x_t)$ for some function $T : \mathbb{N}_0 \times X \to X$ with $X \subset \mathbb{R}^p$, which defines a non-autonomous dynamical system. The function f to be minimized in the learning process, or rather its gradient, becomes a part of T. If f is at least continuously differentiable a local minimum gives the necessary condition $\nabla f(w_\star) = 0$. We show that this condition leads to a fixed point $x_\star$ of T, where the moments are all zero. We analyse the stability properties of this fixed point and prove local asymptotic stability. This is done by considering a time-variant iteration T as the perturbation of a time-invariant iteration $\bar{T}$ where Banach-like fixed point arguments can be applied. We use the second method of Lyapunov for stability analysis where the vanishing moments simplify the computation and estimates for the eigenvalues. Asymptotic stability however is equivalent to convergence of the iteration defined T to $x_\star$ for all x_0 sufficiently close to $x_\star$. The conditions needed for the fixed point analysis and stability results require the learning rate to be sufficiently small. Note that these results cannot be obtained directly from standard fixed point theorems for autonomous systems, because the iteration index t enters the dynamics. Therefore also estimates of the eigenvalues depend on the iteration t, and even a bound on the spectral radius uniform in t does not give the convergence results presented here: It is well known that $\rho(A) < 1$ implies the existence of a vector norm with induced matrix norm such that $\|A\| < 1$, but this norm depends on A. So $\rho(A_t) \leq c < 1$ for some c for all $t \in \mathbb{N}_0$ does not imply the existence of a *single* norm such that $\|A_t\| < 1$ for all t. We emphasize that the result is purely qualitative, giving no explicit guidance to the choice of the learning rates. The main advantage of our approach is the clearness of the proof, only computation of eigenvalues is needed once the iteration has been written in terms of T and $\bar{T}$. These calculations are much more simple than the lengthy estimates in [4], [7] and [8].

D. Preliminaries

We recall some standard definitions and facts from the theory of difference equations and discrete time dynamical systems, see e.g. [18, Definition 5.4.1] or [19]. Consider $T : \mathbb{N}_0 \times M \to M$ with $M \subset \mathbb{R}^n$ which defines a non-autonomous dynamical system by the iteration

$$x_{t+1} = T(t, x_t), \quad t \in \mathbb{N}_0, x_0 \in M \tag{2}$$

with solutions $x : \mathbb{N}_0 \to M$, $t \mapsto x_t$ depending on the initial value x_0. We use the notations $x_t = x(t; x_0)$ and $x = x(\cdot; x_0)$ to emphasize the dependence of solutions on the initial value if necessary. We always use the initial time $t_0 = 0$.

Autonomous systems constitute the special case where T does not depend on t, so we can abbreviate to $\bar{T} : M \to M$ and write

$$x_{t+1} = \bar{T}(x_t), \quad t \in \mathbb{N}_0, x_0 \in M \tag{3}$$

A point $x_\star \in M$ is called *equilibrium* or *fixed point* if $T(t, x_\star) = x_\star$ for all $t \in \mathbb{N}_0$, so the constant function $x_t = x_\star$ for all $t \in \mathbb{N}_0$ is a solution of (2). In the following the asterisk will always denote equilibria or their components. Consider a solution $x = x(\cdot; x_0)$ of (2). x is called *stable*, if for each $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon)$ such that any solution $\tilde{x} = \tilde{x}(\cdot; \tilde{x}_0)$ of (2) with $\|\tilde{x}_0 - x_0\| < \delta$ fulfills $\|\tilde{x}_t - x_t\| < \varepsilon$ for all $t \in \mathbb{N}_0$.

x is called *attractive* if there exists $\delta > 0$ such that any solution $\tilde{x}$ with $\|\tilde{x}_0 - x_0\| < \delta$ fulfills $\lim_{t\to\infty} \|\tilde{x}_t - x_t\| = 0$. x is called *asymptotically stable* if it is stable and attractive.

Recall that a *contraction* is a self-mapping on some set with Lipschitz constant $L < 1$, i.e. a mapping $\bar{T} : M \to M$, $M \subset \mathbb{R}^n$ with $\left\|\bar{T}(x) - \bar{T}(y)\right\| \leq L \|x - y\|$ for all $x, y \in M$. If M is complete, i.e. all Cauchy sequences converge, then a unique fixed point $x_\star \in M$ of $\bar{T}$ exists by the Banach fixed point theorem.

Theorem II.1. ***Linearized asymptotic stability implies local nonlinear stability*** *Consider $\bar{T} : M \to M$ with a fixed point $x_\star$ and $\bar{T}$ continuously differentiable in an open neighbourhood $B_r(x_\star) \subset M$ of $x_\star$. Denote the Jacobian by $D\bar{T}_{x_\star}$, and assume $\left\|D\bar{T}_{x_\star}\right\| < 1$ for some norm on Mat_n. Then there exists $0 < \varepsilon \leq r$ and $0 \leq c < 1$ such that for all x_0 with $\|x_0 - x_\star\| < \varepsilon$*

$$\|x(t; x_0) - x_\star\| \leq c^t \|x_0 - x_\star\| \qquad \forall t \in \mathbb{N}_0.$$

i.e. $x_\star$ is locally exponentially and asymptotically stable.

The theorem is the core of the first method of Lyapunov for discrete time systems. Unfortunately, we could not find a proof in any english textbook like [20], [18]. For a proof see the preprint [21, Theorem 3.3] or the German textbook [22, Theorem 5.4]

III. Convergence Proof

Let $w \in \mathbb{R}^n$ be the weights of the function $f(w) \in C^2(\mathbb{R}^n, \mathbb{R})$, which has to be minimized. We also define $g(w) := \nabla f(w) \in \mathbb{R}^n$ as the gradient of f and the state variable of our dynamical system $x = (m, v, w)$. With these definitions we can rewrite the Adam-Optimizer as a system of the form (2).

$$\begin{aligned} m_{t+1} &:= \beta_1 m_t + (1 - \beta_1)\, g(w_t) \in \mathbb{R}^n \\ v_{t+1} &:= \beta_2 v_t + (1 - \beta_2)\, g(w_t) \otimes g(w_t) \in \mathbb{R}^n \\ w_{t+1} &:= w_t - \alpha \frac{\sqrt{1 - \beta_2^{t+1}}}{\left(1 - \beta_1^{t+1}\right)} \left(m_{t+1} \oslash \sqrt{v_{t+1} \oplus \epsilon}\right) \in \mathbb{R}^n \end{aligned} \tag{4}$$

So the Adam optimizer can be written as the iteration of a time-variant dynamical system $x_{t+1} = [m_{t+1}, v_{t+1}, w_{t+1}]^\perp =$

$T(t,x) = T\left(t, [m_t, v_t, w_t]^\perp\right) \in \mathbb{R}^{3n}$. We split the system in an autonomous and a non-autonomous part

$$x_{t+1} = T(t, x_t) = \bar{T}(x_t) + \Theta(t, x_t) \tag{5}$$

with

$$\bar{T}(x_t) = \begin{bmatrix} \beta_1 m_t + (1-\beta_1) g(w_t) \\ \beta_2 v_t + (1-\beta_2) g(w_t) \otimes g(w_t) \\ w_t - \alpha \left(m_{t+1} \oslash \sqrt{v_{t+1} \oplus \epsilon}\right) \end{bmatrix} \tag{6}$$

and

$$\Theta(t, x_t) = \begin{bmatrix} 0 \\ 0 \\ -\alpha\left(\frac{\sqrt{1-\beta_2^{t+1}}}{1-\beta_1^{t+1}} - 1\right)\left(m_{t+1} \oslash \sqrt{v_{t+1} \oplus \epsilon}\right) \end{bmatrix} \tag{7}$$

To avoid lengthy expressions we use m_{t+1} and v_{t+1} as an abbreviation for the updated terms instead of the filters depending on m_t, $g(w_t)$ and v_t. The autonomous system is Adam without bias correction, the disturbance term Θ adds bias correction which leads to a non-autonomous system. The Jacobian matrix of the autonomous system (6) is

$$J_{\bar{T}}(m_t, v_t, w_t) = \begin{pmatrix} \beta_1 I & 0 & (1-\beta_1)\nabla_w g(w_t) \\ 0 & \beta_2 I & \frac{\partial v}{\partial w} \\ \frac{\partial w}{\partial m} & \frac{\partial w}{\partial v} & \frac{\partial w}{\partial w} \end{pmatrix}$$

with

$$\begin{aligned} \frac{\partial v}{\partial w} =& 2(1-\beta_2)\operatorname{diag}(g(w_t))\nabla_w g(w_t) \\ \frac{\partial w}{\partial m} =& -\alpha \operatorname{diag}\left(\beta_1 \oslash \sqrt{v_{t+1} \oplus \epsilon}\right) \\ \frac{\partial w}{\partial v} =& \frac{\alpha}{2\beta_2} \operatorname{diag}\left(m_{t+1} \oslash (v_{t+1} \oplus \epsilon)^{\frac{3}{2}}\right) \\ \frac{\partial w}{\partial w} =& I - \alpha\Big((1-\beta_1)\operatorname{diag}(v_{t+1} \oplus \epsilon)^{-\frac{1}{2}} \\ & - \operatorname{diag}(m_{t+1} \otimes (v_{t+1} \oplus \epsilon)^{-\frac{3}{2}} \otimes g(w_t)\Big)\nabla_w g(w_t) \end{aligned}$$

We have the following simple observation:

Lemma III.1. *Consider a critical point $w_\star$ for f, $\nabla f(w_\star) = 0$. Then $x_\star = (0,0,w_\star)^\perp$ is a fixed point for (6) and (4).*

Proof. We start the iteration with $w_0 = w_\star$, $v_0 = 0$ and $m_0 = 0$, i.e. $x_0 = (0,0,w_\star)$. Then (6) gives $x_1 = T(x_0) = x_0$, and inductively $x_t = x_0$ for all t. The same holds for (4). □

Now we investigate the stability of this equilibrium with the goal of asymptotic stability for local minima $w_\star$. The analysis is simplified because the m and v components of $x_\star$ are 0. So we reach the following Jacobian:

$$J_{\bar{T}}(0,0,w_\star) = \begin{pmatrix} \beta_1 I & 0 & (1-\beta_1)\nabla_w g(w_\star) \\ 0 & \beta_2 I & 0 \\ \frac{-\alpha\beta_1}{\sqrt{\epsilon}} I & 0 & I - \frac{\alpha(1-\beta_1)}{\sqrt{\epsilon}}\nabla_w g(w_\star) \end{pmatrix}$$

Theorem III.2. *Let $J_{\bar{T}}(m,v,w) \in \mathrm{Mat}_{3n}$ be the Jacobian of system (6) and $w_\star \in \mathbb{R}^n$ a minimum of f with positive definite Hessian $\nabla_w^2 f(w_\star) = \nabla_w g(w_\star)$. Denote $\mu_i \in \mathbb{R}$ with $i \in \{1, \cdots, n\}$ the i-th eigenvalue of $\nabla_w g(w_\star)$, $\varphi_i := \frac{\alpha}{\epsilon}(1-\beta_1)$ and all other parameters are defined as in Algorithm 1. Then $J_{\bar{T}}(0,0,w_\star)$ has the eigenvalues, for $i = 1, \ldots, n$:*

$$\begin{aligned} \lambda_{1,i} &= \beta_2 \\ \lambda_{2,i} &= \frac{(\beta_1+1) + \sqrt{(\beta_1+1)^2 - 4\left(\beta_1 - \frac{\alpha\mu_i(\beta_1-1)}{\sqrt{\epsilon}}\right)}}{2} \\ \lambda_{3,i} &= \frac{(\beta_1+1) - \sqrt{(\beta_1+1)^2 - 4\left(\beta_1 - \frac{\alpha\mu_i(\beta_1-1)}{\sqrt{\epsilon}}\right)}}{2} \end{aligned}$$

In the combination of Theorem III.2 and II.1 we still have to show, that $|\lambda_{j,i}| < 1$ holds, then the spectral radius for the Jacobian is smaller than 1 and we prove the local convergence.

Theorem III.3. *Let the parameters be defined as in Theorem III.2 and inequality (1) holds, then $\rho(J_{\bar{T}}(0,0,w_\star)) < 1$.*

Corollary III.4. *Let the parameters be defined as in Theorem III.2 and such that $\frac{\alpha \max_{i=1}^n(\mu_i)}{\sqrt{\epsilon}}(1-\beta_1) < 2\beta_1 + 2$ holds for $i \in \{1, \ldots, n\}$, then Algorithm 1 converges locally with exponential rate of convergence.*

Proof. Consider the non-autonomous system (5) with $\bar{T}(x_t)$ and $\Theta(t, x_t)$ as defined in equations (6) and (7). The Hessian of f is continuous, so the gradient of f is locally Lipschitz with some constant $L > 0$, $\|g(w_1) - g(w_2)\| \leq L\|w_1 - w_2\|$ for all w_1, w_2 in some neighbourhood of $w_\star$. Let all other parameters be defined as in Theorem III.2, especially $\frac{\alpha \max_{i=1}^n(\mu_i)}{\sqrt{\epsilon}}(1-\beta_1) < 2\beta_1 + 2$. Using $m_\star = 0$ and $g(w_\star) = 0$ we estimate

$$\begin{aligned} \|\Theta(t,x)\| =& \alpha \left|\frac{\sqrt{1-\beta_2^{t+1}}}{1-\beta_1^{t+1}} - 1\right| \\ & \cdot \frac{\|\beta_1 m + (1-\beta_1)g(w)\|}{\sqrt{\beta_2 v + (1-\beta_2)g(w) \otimes g(w) \oplus \epsilon}} \\ \leq& \frac{\alpha}{\sqrt{\epsilon}} \left|\frac{\sqrt{1-\beta_2^{t+1}} - \left(1-\beta_1^{t+1}\right)}{\left(1-\beta_1^{t+1}\right)}\right| \\ & \cdot \|\beta_1 m + (1-\beta_1)g(w)\| \\ \leq& \frac{\alpha}{\sqrt{\epsilon}(1-\beta_1)} \left|\frac{\left(1-\beta_2^{t+1}\right) - \left(1-\beta_1^{t+1}\right)^2}{\sqrt{1-\beta_2^{t+1}} + \left(1-\beta_1^{t+1}\right)}\right| \\ & \cdot (\beta_1 \|m\| + (1-\beta_1)\|g(w)\|) \\ \leq& \frac{C}{4}\left|\left(1-\beta_2^{t+1}\right) - \left(1-\beta_1^{t+1}\right)^2\right| \\ & \cdot (\beta_1 \|m - m_\star\| + (1-\beta_1)\|g(w) - g(w_\star)\|) \\ \leq& \frac{C}{4}\left|-\beta_2^{t+1} - 2\beta_1^{t+1} + \beta_1^{2(t+1)}\right| \\ & \cdot (\beta_1 \|m - m_\star\| + (1-\beta_1)L\|w - w_\star\|) \\ \leq& C\beta^{t+1}(\beta_1 \|m - m_\star\| + (1-\beta_1)L\|w - w_\star\|) \end{aligned}$$

where we have used the Lipschitz continuity of g, and set $\beta = \max\{\beta_1, \beta_2, \beta_1^2\}$, $C := \frac{4\alpha}{\sqrt{\epsilon}(1-\beta_1)\left(\sqrt{1-\beta_2}+(1-\beta_1)\right)}$. The

term $\beta_1 \|m - m_\star\| + (1-\beta_1)L\|w - w_\star\|$ corresponds to a norm

$$\|(\tilde{m}, \tilde{w})\|_* := \beta_1 \|\tilde{m}\| + (1-\beta_1)L\|\tilde{w}\|, \qquad \tilde{m}, \tilde{w} \in \mathbb{R}^n$$

on $\mathbb{R}^{2n}$ (which does not depend on $w_\star$). By the equivalence of norms in finite dimensional spaces we can estimate $\|(\tilde{m}, \tilde{w})\|_* \leq \tilde{C}\|(\tilde{m}, \tilde{w})\|$ for some $\tilde{C} > 0$. We continue the estimate:

$$\begin{aligned} &\leq C\beta^{t+1}\tilde{C}\|(m - m_\star, w - w_\star)\| \\ &\leq (C\beta\tilde{C})\beta^t \|x - x_\star\| =: \bar{C}\beta^t\|x - x_\star\| \end{aligned}$$

for some $\bar{C} > 0$. With this estimate and Theorem V.1, it is sufficient to prove exponential stability of a fixed point of $\bar{T}$. By Theorem III.3 we get $\rho(J_{\bar{T}}(0, 0, w_\star)) < 1$. Thus with Theorem II.1 the fixed point $(0, 0, w_\star)$ corresponding to the minimum $w_\star$ is locally exponentially stable, and Theorem V.1 gives local exponential convergence of the non-autonomous system $T(t, x)$, i.e. the Adam algorithm. □

The following corollary is a combination of our results with the results in [23] to show global convergence in the strictly convex case. The idea is: The iteration reaches an ϵ-bounded gradient $\|\nabla f(w_{\tilde{t}})\| < \epsilon$ in some iteration $\tilde{t}$ for suitable choice of hyperparameters according to [23]. The arguments of [23] do not imply that $\|\nabla f(w_t)\|$ remains bounded, nor $\lim_{t\to\infty} \nabla f(w_t) = 0$, nor that $\lim_{t\to\infty} w_t$ exists.

At this point we use our results to show that for ϵ small enough the condition $\|\nabla f(w_t)\| < \epsilon$ implies that w_t is in the domain of local convergence.

Corollary III.5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ strictly convex with minimum $w_\star \in \mathbb{R}^n$. Assume $f \in C^2$ and $\nabla^2 f(w_\star)$ positive definite. Assume that the conditions of Theorem 2.2 in [23] hold (boundedness of $\|\nabla f\|$, conditions on hyperparameters of Adam). Then Adam converges globally for the minimum $w_\star$.*

Proof. Denote $x_\star = (m_\star, v_\star, w_\star) = (0, 0, w_\star)$ as in Theorem III.2. Fix $\alpha > 0$ such that the assumptions of Corollary III.4 hold. Choose $\tilde{\varepsilon} > 0$ small enough such that for all $x_0 = (m_0, v_0, w_0) \in B_{\tilde{\varepsilon}}(x_\star)$ the Adam algorithm converges according to Corollary III.4.

Theorem 2.2 in [23] shows that for suitable choice of parameters, for any $\varepsilon > 0$ we have $\|\nabla f(w_t)\| < \varepsilon$ for some $t \in \{0, \ldots, T\}$ independent of $w_0 \in \mathbb{R}^n$ where T depends on ε. Fix this index t. By Lemma V.3 there exist $C > 0$, $\bar{\epsilon} > 0$ such that $\|\nabla f(x)\| \geq C\|x - x_\star\|$ for all $x \in B_{\bar{\epsilon}}(x_\star)$.

Choosing x_0 with $\|x_0 - x_\star\|$ small enough leads to convergence to the minimum according to Corollary III.4. □

Note that the strict convexity was only used to guarantee uniqueness of a minimum, and that the only critical point is this minimum. Otherwise $\|\nabla f(w_t)\| < \epsilon$ might hold near a maximum or saddle point where Corollary III.4 does not apply.

Of course other results which guarantee ϵ-boundedness of the gradient at some iteration t for Adam in batch mode can also be combined with our approach.

TABLE I
COLOUR DESCRIPTION FOR THE CONVERGENCE INVESTIGATIONS

	Inequality (9) satisfied	Inequality (1) satisfied	Adam finds solution
green	yes	yes	yes
blue	no	yes	yes
yellow	yes	no	yes
white	no	no	yes
black	yes	yes	no
cyan	no	yes	no
magenta	yes	no	no
red	no	no	no

IV. EXPERIMENTS

To compare our requirements for convergence to the requirements taken by [7] or [4], we make some empirical experiments. First, we look at the different requirements to the hyperparameter.

$$\beta_1 < \sqrt{\beta_2} \tag{8}$$

$$\beta_1^2 < \sqrt{\beta_2} \tag{9}$$

Inequality (1) describes the needed requirement presented in this paper. Problematically in this estimation is, that we need the maximum eigenvalue of $(1-\beta_1)\nabla_w g(w_\star)$ and consequently $w_\star$. Therefore our estimation is an a posteriori estimation. But with (1) we learn something about the relationship between the hyperparameters. $\frac{\alpha}{\sqrt{\epsilon}}$ has to be very small to fullfill inequality 1. With a small α or a big ϵ we always make the weight change smaller and so we do not jump over $w_\star$. Inequality (8) was presented in [7] and inequality (9) was originally presented in [4]. Both are a priori estimations for the hyperparameters.

To show the behaviour of all estimations we set up the following experiments. In Experiment 1 and 2 we want to minimize $f(w) := w^4 + w^3$ with the minimum $w_\star = -\frac{3}{4}$. In Experiment 3 we minimize the multidimensional function $f(w_1, w_2) := (w_1 + 2)^2 (w_2 + 1)^2 + (w_1 + 2)^2 + 0.1(w_2 + 1)^2$ with the minimum $w_\star = (-2, -1)$. We run the Adam optimizer 10000 times in every hyperparameter setting and if the last five iterations $w_{end} \in \mathbb{R}^5$ are near enough to the known solution $w_\star$ the attempt is declared as convergent. Near enough in this setting means that all components of w_{end} are contained in the interval $[w_\star - 10^{-2}, w_\star + 10^{-2}]$. The colour coding of our experiments can be found in Table I. To keep the clarity of our results we only compare the original Adam inequality with our inequality. With inequality (8) we obtain similar figures.

Experiment 1

First, we iterate over $\epsilon \in \{10^{-8}, \ldots, 10^{-6}\}$ and $\beta_1 \in \{0.01, \ldots, 0.99\}$. The other hyperparamteres are fixed $\alpha = 0.001$, $\beta_2 = 0.1$. This setting leads us to figure 2. The only area, where the Adam optimizer is not finding a solution (red dots), is inside the white area. So both inequalities are not satisfied and the convergence is not given. The white area – Adam converge but no inequality is satisfied – is

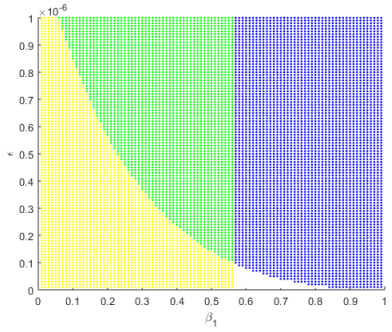

Fig. 2. Iterating over ϵ and β_1

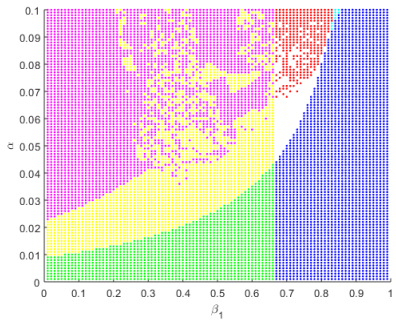

Fig. 4. Iterating over α and β_1 with $x_0 = -0.750000001$

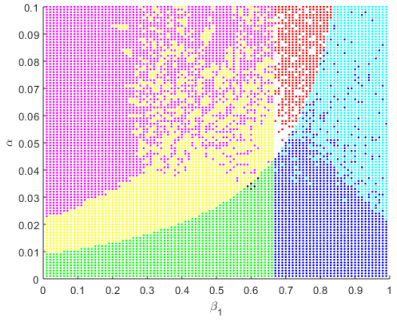

Fig. 3. Iterating over α and β_1

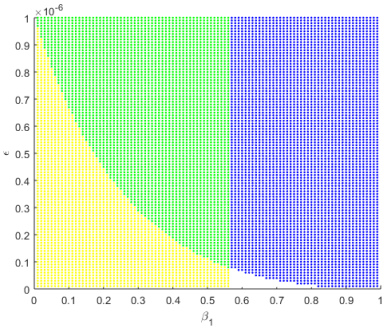

Fig. 5. Experiment 3 :Iterating over ϵ and β_1.

formed because we only talk about estimation and not clear boundaries. The blue and yellow area can be made larger or smaller by changing β_2 or α.

Experiment 2

In the second experiment we iterate over $\alpha \in \{0.001, \ldots, 0.1\}$ and $\beta_1 \in \{0.01, \ldots, 0.99\}$. $\beta_2 = 0.2$ and $\epsilon = 10^{-4}$ are fixed. With the starting point $x_0 = -2$ we reach figure 3. In the magenta and the cyan area the Adam method is not reaching the solution, although inequality (9) or (1) is satisfied. The Adam is oscillating around the solution but do not reach them. The big difference is that the non-convergence in the cyan area is attributable to the fact that our proof only shows local convergence. By starting in $x_0 = -0.750000001$ the cyan area is almost complete blue (see figure 4). In contrary the magenta area does not change that much.

Experiment 3

In the last experiment we use the same hyperparameters as in experiment 1. Therefore we reach a similar looking figure 5 by iterating over the parameters. The reason for the enlargement of the blue and green area is the different function $f(x)$, thus different eigenvalues in inequality (1). By observing the convergence behaviour from each of the four differently coloured areas in figure 5, we can not spot big differences.

V. Conclusion and Discussion

In this paper we introduce a local convergence proof of the Adam method and to the best of our knowledge it is the first at all. We also give an a posteriori boundary for the hyperparameters and show, that the choice of β_2 does not matter for the convergence near a minimum.

However the proof is based on the vanishing gradient condition $\nabla f(w_\star) = 0$ and cannot be used for an incremental algorithm for $f(w) = \frac{1}{N}\sum_{i=1}^{N} f_i(w)$ where different component gradients $g_t = \nabla f_{i_t}(w_t)$ are used in the iterations for the moments. Clearly $\nabla f(w_\star) = 0$ does not imply $\nabla f_i(w_\star) = 0$ for all components. We are investigating how the incremental dynamical system can be related to the batch system.

The analysis applies to any local minimum with positive definite Hessian and therefore does not require overall convexity. In order to show global convergence of Adam-like algorithms other methods have to be applied.

References

[1] J. E. Nesterov, *Introductory lectures on convex optimization: A basic course*, ser. Applied optimization. Boston, Mass.: Kluwer Acad. Publ, 2004, vol. APOP 87.

[2] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York: Springer, 2006.

[3] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, Jul. 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=1953048.2021068.

[4] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," *CoRR*, vol. abs/1412.6980, 2014.

[5] S. Bock, "Rotationsermittlung von Bauteilen basierend auf neuronalen Netzen (unpublished)," Ostbayerische Technische Hochschule Regensburg, M.Sc. thesis, 2017.

[6] D. M. Rubio, "Convergence Analysis of an Adaptive Method of Gradient Descent," University of Oxford, Oxford, M.Sc. thesis, 2017.

[7] J. S. Reddi, S. Kale, and S. Kumar, "On the Convergence of Adam and Beyond," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=ryQu7f-RZ.

[8] X. Chen, S. Liu, R. Sun, and M. Hong, *On the Convergence of A Class of Adam-Type Algorithms for Non-Convex Optimization*, 2018. [Online]. Available: https://arxiv.org/abs/1808.02941.

[9] H. Robbins and S. Monro, "A Stochastic Approximation Method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., Curran Associates, Inc, 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[11] G. E. Hinton, N. Srivastava, and K. Swersky, *Overview of mini-batch gradient descent (unpublished)*, Toronto, 2012. [Online]. Available: http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.

[12] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[13] F. Zou, L. Shen, Z. Jie, W. Zhang, and W. Liu, *A Sufficient Condition for Convergences of Adam and RMSProp*, 2018. [Online]. Available: https://arxiv.org/abs/1811.09358.

[14] A. B. da Silva and M. Gazeau, *A general system of differential equations to model first order adaptive algorithms*, 2018. [Online]. Available: https://arxiv.org/pdf/1810.13108.

[15] A. Krizhevsky, *Learning Multiple Layers of Features from Tiny Images*, Toronto, 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[16] S. Bock, M. Weiß, and J. Goppold, *An improvement of the convergence proof of the ADAM-Optimizer*. Clusterkonferenz 2018, 2018. [Online]. Available: http://arxiv.org/abs/1804.10587.

[17] A. Barakat and P. Bianchi, *Convergence of the ADAM algorithm from a Dynamical System Viewpoint*, Paris, 2018. [Online]. Available: https://arxiv.org/abs/1810.02263.

[18] R. P. Agarwal, Z. Nashed, and E. Taft, *Difference Equations and Inequalities: Theory, Methods, and Applications*, 2nd ed. Boca Raton: Chapman and Hall/CRC, 2000.

[19] W. G. Kelley and A. C. Peterson, *Difference equations: An introduction with applications*, 2. ed. San Diego, Calif.: Harcourt/Academic Press, 2001.

[20] H. Freeman, *Discrete-time systems: An introduction to the theory*. New York: J. Wiley, 1965.

[21] N. Bof, R. Carli, and L. Schenato, *Lyapunov Theory for Discrete Time Systems*. [Online]. Available: https://arxiv.org/abs/1809.05289.

[22] U. Krause and T. Nesemann, *Differenzengleichungen und diskrete dynamische Systeme: Eine Einführung in Theorie und Anwendungen*, 1. Aufl., ser. De Gruyter Studium. s.l.: Walter de Gruyter GmbH Co.KG, 2012.

[23] S. De, A. Mukherjee, and E. Ullah, *Convergence guarantees for RMSProp and ADAM in non-convex optimization and an empirical comparison to Nesterov acceleration*, 2018. [Online]. Available: http://arxiv.org/pdf/1807.06766v3.

[24] J. R. Silvester, *Determinants of Block Matrices*, London, 1999. [Online]. Available: http://www.ee.iisc.ac.in/new/people/faculty/prasantg/downloads/blocks.pdf.

APPENDIX

Proof. **(Theorem III.2)**
We see that $J_{\bar{T}}(0,0,w_\star)$ has the n-fold eigenvalue β_2. So we can drop second block row and column of $J_{\bar{T}}$ and investigate the eigenvalues of

$$\begin{pmatrix} \beta_1 I & (1-\beta_1)\nabla_w g(w_\star) \\ -s\beta_1 I & I - s(1-\beta_1)\nabla_w g(w_\star) \end{pmatrix} =: \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

where we use the abbreviation $s := \frac{\alpha}{\sqrt{\epsilon}}$. B and D are symmetric since $\nabla_w g(w_\star)$ is the Hessian of f. By the spectral theorem we can diagonalize B as $B = Q\Lambda Q^\perp$ with an orthogonal matrix Q and a diagonal matrix of eigenvalues Λ. Analogously holds $D = I - Q\Lambda Q^\perp = Q(I-\Lambda)Q^\perp$. We make a similarity transformation with $\tilde{Q} := \begin{pmatrix} Q & 0 \\ 0 & Q \end{pmatrix} \in \mathrm{Mat}_{2n}$. This leaves the eigenvalues unchanged and gives

$$\tilde{Q}^\perp \begin{pmatrix} A & B \\ C & D \end{pmatrix} \tilde{Q} = \begin{pmatrix} \beta_1 I & (1-\beta_1)\mu_i I \\ -s\beta_1 I & I - s(1-\beta_1)\mu_i \end{pmatrix}$$

with μ_i the i-th eigenvalue of the Hessian. Eigenvalues does not change in similarity transformations, so we can also calculate the eigenvalues of our new block matrix with four diagonal sub matrices.

$$\begin{aligned} &\det \begin{pmatrix} (\beta_1 - \lambda) I & (1-\beta_1)\mu_i I \\ -s\beta_1 I & (1 - s(1-\beta_1)\mu_i - \lambda) I \end{pmatrix} \\ &= \det\left((\beta_1 - \lambda)(1 - s(1-\beta_1)\mu_i - \lambda) I + (1-\beta_1) s\beta_1 \mu_i I\right) \\ &\overset{!}{=} 0 \end{aligned}$$

Therefore the matrix is a diagonal matrix, we can conclude:

$$\det\left(J_{\bar{T}}\left(0,0,w_\star\right)-\lambda I\right)=\prod_{i=1}^{n}\left(\beta_1-\lambda\right)\left(1-s\left(1-\beta_1\right)\mu_i-\lambda\right)+\left(1-\beta_1\right)s\beta_1\mu_i$$

Each factor can be written as

$$\lambda^2-\left(1-s\left(1-\beta_1\right)\mu_i+\beta_1\right)\lambda+\beta_1\overset{!}{=}0$$

and following the statement

$$\lambda_{23,i}=0.5\left(1-s\left(1-\beta_1\right)\mu_i+\beta_1\pm\sqrt{\left(1-s\left(1-\beta_1\right)\mu_i+\beta_1\right)^2-4\beta_1}\right)$$

is true. □

Proof. **(Theorem III.3)**
We already have calculate the eigenvalues of the Jacobian in Theorem III.2. With these we can easily see, that $|\lambda_1|=|\beta_2|<1$ is satisfied per the requirements of algorithm 1. Therefore we only have to look at:

$$\lambda_{23,i}=\frac{1+\beta_1-\varphi_i\pm\sqrt{\left(1+\beta_1-\varphi_i\right)^2-4\beta_1}}{2}$$

We define $\varphi_i:=\frac{\alpha\mu_i}{\sqrt{\epsilon}}\left(1-\beta_1\right)$ and simplify the eigenvalue.

$$|\lambda_{23,i}|=\frac{1}{2}\left|\left(1+\beta_1-\varphi_i\right)\pm\underbrace{\sqrt{\left(1+\beta_1-\varphi_i\right)^2-4\beta_1}}_{①}\right|$$

First we look at upper bound of the eigenvalues. For this we take term ① combined with the regrets for φ_i:

$$\sqrt{\left(1+\beta_1-\varphi_i\right)^2-4\beta_1}<\sqrt{\left(1+\beta_1\right)^2-4\beta_1}=\pm\left(1-\beta_1\right)$$

So if we put this in $\lambda_{23,i}$ we have the inequality $|\lambda_{23,i}|<\frac{1}{2}\left|1+\beta_1-\varphi_i\pm\left(1-\beta_1\right)\right|$. Easy to see are the two cases:

$$\begin{aligned}&|\lambda_{23,i}|<1 && \text{with +}\\&|\lambda_{23,i}|<\beta_1<1 && \text{with -}\end{aligned}$$

In both cases we see that the eigenvalues are smaller than 1 in absolute value. To show the lower bound $\lambda_{23,i}>-1$, we look again at term ①.

$$\underbrace{\sqrt{\left(1+\beta_1-\varphi_i\right)^2-4\beta_1}}_{\in\mathbb{C}\setminus\mathbb{R}}=i\sqrt{4\beta_1-\left(1+\beta_1-\varphi_i\right)^2}$$

Then we can write:

$$\begin{aligned}|\lambda_{23,i}|=&\frac{1}{2}\sqrt{\left(1+\beta_1-\varphi_i\right)^2+4\beta_1-\left(1+\beta_1-\varphi_i\right)^2}\\=&\sqrt{\beta_1}<1\end{aligned}$$

The last inequality is given by the requirements of Theorem III.3 and so we proved the whole Theorem. □

Theorem V.1. ***Convergence to fixed point with perturbation***
Let $M\subset\mathbb{R}^n$ be a complete set, $\bar{T}:M\to M$ Lipschitz continuous with $L<1$, $x_\star\in M$ the unique fixed point of $\bar{T}$. Assume $B_r(x_\star)\subset M$ for some $r>0$. Recall that the non-autonomous system (5) is defined by

$$\tilde{x}_{t+1}=T(\tilde{x}_t):=\bar{T}(\tilde{x}_t)+\Theta(t,\tilde{x}_t)$$

for $\Theta:\mathbb{N}_0\times M\to\mathbb{R}^n$ with the bound $\|\Theta(t,\tilde{x})\|\leq C\beta^t\|\tilde{x}-x_\star\|$ for all $\tilde{x}_t\in M$, $t\in\mathbb{N}_0$ for some $C\geq 0$ and $0<\beta<1$. Then there exists $\varepsilon>0$ such that for all $\tilde{x}_0\in M$ with $\|\tilde{x}_0-x_\star\|<\varepsilon$ the iteration defined by (5) is well-defined, i.e. stays in M, and converges to $x_\star$.

Proof. Let $x=x(\cdot,\tilde{x}_0)$ be the solution of the undisturbed iteration $x_{t+1}=\bar{T}(x_t)$ with initial condition $\tilde{x}_0$, $\tilde{x}=\tilde{x}(\cdot,\tilde{x}_0)$ the corresponding solution of (5). We define $e_t:=\|\tilde{x}_t-x_\star\|$, and estimate using the assumptions

$$\begin{aligned}e_{t+1}&=\left\|\bar{T}(\tilde{x}_t)+\Theta(t,\tilde{x}_t)-x_\star\right\|\\&=\left\|\bar{T}(\tilde{x}_t)-\bar{T}(x_\star)+\Theta(t,\tilde{x}_t)\right\|\\&\leq\left\|\bar{T}(\tilde{x}_t)-\bar{T}(x_\star)\right\|+\|\Theta(t,\tilde{x}_t)\|\\&\leq L\|\tilde{x}_t-x_\star\|+C\beta^t\|\tilde{x}_t-x_\star\|\\&=(L+C\beta^t)e_t\end{aligned}$$

Choosing t large enough, we get $0<L+C\beta^t\leq\tilde{L}<1$ for all $t\geq K$ because $\beta,L<1$. Then

$$e_t\leq\left(\prod_{k=1}^{K}(L+C\beta^k)\right)\tilde{L}^{t-K}e_0=:\tilde{C}\tilde{L}^{t-K}e_0$$

with $\tilde{C}$ independent of $\tilde{x}_0$. So e_t converges to 0 exponentially.

The arguments so far have only been valid if $\tilde{x}_t\in M$, i.e. the iteration is well defined. But choosing $\tilde{x}_0$ such that $e_0=\|\tilde{x}_0-x_\star\|<\frac{r}{\tilde{C}}$ small enough that we can achieve $e_t\leq\tilde{C}e_0<r$. □

Theorem V.2. ***Determinants of Block Matrices [24]***
Let $M=\begin{pmatrix}AB\\CD\end{pmatrix}\in\mathrm{Mat}_{2n}$ be a block matrix with $A,B,C,D\in\mathrm{Mat}_n$. If C and D commute, then $\det(M)=\det(AD-BC)$ holds.

Lemma V.3. *Let $f\in C^2(\mathbb{R}^n,\mathbb{R})$, $x_\star\in\mathbb{R}^n$ with $\nabla f(x_\star)=0$ and $\nabla^2 f(x_\star)$ invertible. Then there exist $\epsilon>0$ and $C>0$ with $\|\nabla f(x)\|\geq C\|x-x_\star\|$ for all $x\in B_\epsilon(x_\star)$.*

Proof. As f is C^2 we have $\nabla f(x)-\nabla f(x_\star)-\nabla^2 f(x_\star)=o(\|x-x_\star\|)$. So for each $\delta>0$ there exists $\epsilon>0$ with $\left\|\nabla f(x)-\nabla f(x_\star)-\nabla^2 f(x_\star)\right\|\leq\delta\|x-x_\star\|$ for all $x\in B_\epsilon(x_\star)$. Assume w.l.o.g. $\delta<\frac{1}{\|\nabla^2 f(x_\star)^{-1}\|}$. Then we have

$$\begin{aligned}\|\nabla f(x)\|&\geq\left\|\nabla^2 f(x_\star)(x-x_\star)\right\|\\&\quad-\left\|\nabla f(x)-\nabla f(x_\star)-\nabla^2 f(x_\star)(x-x_\star)\right\|\\&\geq\frac{1}{\|\nabla^2 f(x_\star)^{-1}\|}\|x-x_\star\|-\delta\|x-x_\star\|\\&=:C\|x-x_\star\|\end{aligned}$$

with $C=\frac{1}{\|\nabla^2 f(x_\star)^{-1}\|}-\delta>0$ by choice of δ. □