Supplementary Material for the Manuscript "On the Implicit Bias of Adam" by Matias D. Cattaneo, Jason M. Klusowski, and Boris Shigida

August 31, 2023

Contents

1 Overview

SA-1.1. This appendix provides some omitted details and proofs.

We consider two algorithms: RMSProp and Adam, and two versions of each algorithm (with the numerical stability ε parameter inside and outside of the square root in the denominator). This means there are four main theorems: Theorem SA-2.4, Theorem SA-3.4, Theorem SA-4.4 and Theorem SA-5.4, each residing in the section completely devoted to one algorithm. The simple induction argument taken from [1], essentially the same for each of these theorems, is based on an auxiliary result whose corresponding versions are Theorem SA-2.3, Theorem SA-3.3, Theorem SA-4.3 and Theorem SA-5.3. The proof of this result is also elementary but long, and it is done by a series of lemmas in Section 6 and Section 7, culminating in Section SA-7.6. Out of these four, we only prove Theorem SA-2.3 since the other three results are proven in the same way with obvious changes.

Section 8 contains some details about the numerical experiments.

SA-1.2 Notation. We denote the loss of the kth minibatch as a function of the network parameters $\boldsymbol{\theta} \in \mathbb{R}^p$ by $E_k(\boldsymbol{\theta})$, and in the full-batch setting we omit the index and write $E(\boldsymbol{\theta})$. As usual, ∇E means the gradient of E, and nabla with indices means partial derivatives, e. g. $\nabla_{ijs} E$ is a shortcut for $\frac{\partial^3 E}{\partial\theta_i \partial\theta_j \partial\theta_s}$.

The letter $T > 0$ will always denote a finite time horizon of the ODEs, h will always denote the training step size, and we will replace nh with t_n when convenient, where $n \in \{0, 1, \dots\}$ is the step number. *We will use the same notation* for the iteration of the discrete algorithm $\left\{\boldsymbol{\theta}^{(k)}\right\}_{k \in \mathbb{Z}_{\geq 0}}$, the piecewise ODE solution $\tilde{\boldsymbol{\theta}}(t)$ and some auxiliary terms for each of the four algorithms: see Definition SA-2.1, Definition SA-

3.1, Definition SA-4.1, Definition SA-5.1. This way, we avoid cluttering the notation significantly. We are careful to reference the relevant definition in all theorem statements.

2 RMSProp with ε outside the square root

Definition SA-2.1. In this section, for some $\boldsymbol{\theta}^{(0)} \in \mathbb{R}^p$, $\nu^{(0)} = \mathbf{0} \in \mathbb{R}^p$, $\rho \in (0,1)$, let the sequence of p-vectors $\left\{\boldsymbol{\theta}^{(k)}\right\}_{k \in \mathbb{Z}_{\geq 0}}$ be defined for $n \geq 0$ by

$$\begin{aligned} \nu_j^{(n+1)} &= \rho \nu_j^{(n)} + (1-\rho)\left(\nabla_j E_n\left(\boldsymbol{\theta}^{(n)}\right)\right)^2, \\ \theta_j^{(n+1)} &= \theta_j^{(n)} - \frac{h}{\sqrt{\nu_j^{(n+1)}} + \varepsilon} \nabla_j E_n\left(\boldsymbol{\theta}^{(n)}\right). \end{aligned} \tag{2.1}$$

Let $\tilde{\boldsymbol{\theta}}(t)$ be defined as a continuous solution to the piecewise ODE

$$\begin{aligned} \dot{\tilde{\theta}}_j(t) &= -\frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \varepsilon} \\ &+ h\left(\frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\left(2P_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \bar{P}_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)}{2\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \varepsilon\right)^2 R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)} - \frac{\sum_{i=1}^p \nabla_{ij} E_n\left(\tilde{\boldsymbol{\theta}}(t)\right) \frac{\nabla_i E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \varepsilon}}{2\left(R_j^{(n)}(\tilde{\boldsymbol{\theta}}(t)) + \varepsilon\right)}\right) \end{aligned} \tag{2.2}$$

with the initial condition $\tilde{\boldsymbol{\theta}}(0) = \boldsymbol{\theta}^{(0)}$, where $\mathbf{R}^{(n)}(\boldsymbol{\theta})$, $\mathbf{P}^{(n)}(\boldsymbol{\theta})$ and $\bar{\mathbf{P}}^{(n)}(\boldsymbol{\theta})$ are p-dimensional functions with components

$$\begin{aligned} R_j^{(n)}(\boldsymbol{\theta}) &:= \sqrt{\sum_{k=0}^n \rho^{n-k}(1-\rho)\left(\nabla_j E_k(\boldsymbol{\theta})\right)^2}, \\ P_j^{(n)}(\boldsymbol{\theta}) &:= \sum_{k=0}^n \rho^{n-k}(1-\rho) \nabla_j E_k\left(\boldsymbol{\theta}\right) \sum_{i=1}^p \nabla_{ij} E_k\left(\boldsymbol{\theta}\right) \sum_{l=k}^{n-1} \frac{\nabla_i E_l\left(\boldsymbol{\theta}\right)}{R_i^{(l)}\left(\boldsymbol{\theta}\right) + \varepsilon}, \\ \bar{P}_j^{(n)}(\boldsymbol{\theta}) &:= \sum_{k=0}^n \rho^{n-k}(1-\rho) \nabla_j E_k\left(\boldsymbol{\theta}\right) \sum_{i=1}^p \nabla_{ij} E_k\left(\boldsymbol{\theta}\right) \frac{\nabla_i E_n\left(\boldsymbol{\theta}\right)}{R_i^{(n)}\left(\boldsymbol{\theta}\right) + \varepsilon}. \end{aligned}$$

Assumption SA-2.2.

1. For some positive constants M_1, M_2, M_3, M_4 we have

$$\begin{aligned} &\sup_i \sup_k \sup_{\boldsymbol{\theta}} \left|\nabla_i E_k(\boldsymbol{\theta})\right| \leq M_1, \\ &\sup_{i,j} \sup_k \sup_{\boldsymbol{\theta}} \left|\nabla_{ij} E_k(\boldsymbol{\theta})\right| \leq M_2, \\ &\sup_{i,j,s} \sup_k \sup_{\boldsymbol{\theta}} \left|\nabla_{ijs} E_k(\boldsymbol{\theta})\right| \leq M_3, \\ &\sup_{i,j,s,r} \sup_k \sup_{\boldsymbol{\theta}} \left|\nabla_{ijsr} E_k(\boldsymbol{\theta})\right| \leq M_4. \end{aligned}$$

2. For some $R > 0$ we have for all $n \in \left\{0, 1, \ldots, \lfloor T/h \rfloor\right\}$

$$R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right) \geq R, \quad \sum_{k=0}^n \rho^{n-k}(1-\rho)\left(\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_k)\right)\right)^2 \geq R^2,$$

where $\tilde{\boldsymbol{\theta}}(t)$ is defined in Definition SA-2.1.

Theorem SA-2.3 (RMSProp with ε outside: local error bound)**.** *Suppose Assumption SA-2.2 holds. Then for all* $n \in \{0, 1, \ldots, \lfloor T/h \rfloor\}$

$$\left| \tilde{\theta}_j(t_{n+1}) - \tilde{\theta}_j(t_n) + h \frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t_n)\right)}{\sqrt{\sum_{k=0}^{n} \rho^{n-k}(1-\rho)\left(\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_k)\right)\right)^2} + \varepsilon} \right| \leq C_1 h^3$$

for a positive constant C_1 *depending on* ρ*.*

The proof of Theorem SA-2.3 is conceptually simple but very technical, and we delay it until Section 7. For now assuming it as given and combining it with a simple induction argument gives a global error bound which follows.

Theorem SA-2.4 (RMSProp with ε outside: global error bound)**.** *Suppose Assumption SA-2.2 holds, and*

$$\sum_{k=0}^{n} \rho^{n-k}(1-\rho)\left(\nabla_j E_k\left(\boldsymbol{\theta}^{(k)}\right)\right)^2 \geq R^2$$

for $\left\{\boldsymbol{\theta}^{(k)}\right\}_{k \in \mathbb{Z}_{\geq 0}}$ *defined in Definition SA-2.1. Then there exist positive constants* d_1*,* d_2*,* d_3 *such that for all* $n \in \{0, 1, \ldots, \lfloor T/h \rfloor\}$

$$\|\mathbf{e}_n\| \leq d_1 e^{d_2 n h} h^2 \quad \textit{and} \quad \|\mathbf{e}_{n+1} - \mathbf{e}_n\| \leq d_3 e^{d_2 n h} h^3,$$

where $\mathbf{e}_n := \tilde{\boldsymbol{\theta}}(t_n) - \boldsymbol{\theta}^{(n)}$*. The constants can be defined as*

$$\begin{aligned}
d_1 &:= C_1, \\
d_2 &:= \left[1 + \frac{M_2\sqrt{p}}{R+\varepsilon}\left(\frac{M_1^2}{R(R+\varepsilon)} + 1\right) d_1\right]\sqrt{p}, \\
d_3 &:= C_1 d_2.
\end{aligned}$$

Proof. We will show this by induction over n, the same way an analogous bound is shown in [1].

The base case is $n = 0$. Indeed, $\mathbf{e}_0 = \tilde{\boldsymbol{\theta}}(0) - \boldsymbol{\theta}^{(0)} = \mathbf{0}$. Then the jth component of $\mathbf{e}_1 - \mathbf{e}_0$ is

$$\begin{aligned}
[\mathbf{e}_1 - \mathbf{e}_0]_j = [\mathbf{e}_1]_j = \tilde{\theta}_j(t_1) - \theta_j^{(0)} + \frac{h \nabla_j E_0\left(\boldsymbol{\theta}^{(0)}\right)}{\sqrt{(1-\rho)\left(\nabla_j E_0\left(\boldsymbol{\theta}^{(0)}\right)\right)^2} + \varepsilon} \\
= \tilde{\theta}_j(t_1) - \tilde{\theta}_j(t_0) + \frac{h \nabla_j E_0\left(\tilde{\boldsymbol{\theta}}(t_0)\right)}{\sqrt{(1-\rho)\left(\nabla_j E_0\left(\tilde{\boldsymbol{\theta}}(t_0)\right)\right)^2} + \varepsilon}.
\end{aligned}$$

By Theorem SA-2.3, the absolute value of the right-hand side does not exceed $C_1 h^3$, which means $\|\mathbf{e}_1 - \mathbf{e}_0\| \leq C_1 h^3 \sqrt{p}$. Since $C_1\sqrt{p} \leq d_3$, the base case is proven.

Now suppose that for all $k = 0, 1, \ldots, n-1$ the claim

$$\|\mathbf{e}_k\| \leq d_1 e^{d_2 k h} h^2 \quad \text{and} \quad \|\mathbf{e}_{k+1} - \mathbf{e}_k\| \leq d_3 e^{d_2 k h} h^3$$

is proven. Then

$$\begin{aligned}
\|\mathbf{e}_n\| &\overset{\text{(a)}}{\leq} \|\mathbf{e}_{n-1}\| + \|\mathbf{e}_n - \mathbf{e}_{n-1}\| \leq d_1 e^{d_2(n-1)h} h^2 + d_3 e^{d_2(n-1)h} h^3 \\
&= d_1 e^{d_2(n-1)h} h^2 \left(1 + \frac{d_3}{d_1} h\right) \overset{\text{(b)}}{\leq} d_1 e^{d_2(n-1)h} h^2 (1 + d_2 h)
\end{aligned}$$

$$\overset{(c)}{\leq} d_1 e^{d_2(n-1)h} h^2 \cdot e^{d_2 h} = d_1 e^{d_2 nh} h^2,$$

where (a) is by the triangle inequality, (b) is by $d_3/d_1 \leq d_2$, in (c) we used $1 + x \leq e^x$ for all $x \geq 0$.

Next, combining Theorem SA-2.3 with (2.1), we have

$$\left| [\mathbf{e}_{n+1} - \mathbf{e}_n]_j \right| \leq C_1 h^3 + h \left| \frac{\nabla_j E_n \left(\tilde{\boldsymbol{\theta}}(t_n) \right)}{\sqrt{A} + \varepsilon} - \frac{\nabla_j E_n \left(\boldsymbol{\theta}^{(n)} \right)}{\sqrt{B} + \varepsilon} \right|, \tag{2.3}$$

where to simplify notation we put

$$A := \sum_{k=0}^{n} \rho^{n-k} (1 - \rho) \left(\nabla_j E_k \left(\tilde{\boldsymbol{\theta}}(t_k) \right) \right)^2,$$
$$B := \sum_{k=0}^{n} \rho^{n-k} (1 - \rho) \left(\nabla_j E_k \left(\boldsymbol{\theta}^{(k)} \right) \right)^2.$$

Using $A \geq R^2$, $B \geq R^2$, we have

$$\left| \frac{1}{\sqrt{A} + \varepsilon} - \frac{1}{\sqrt{B} + \varepsilon} \right| = \frac{|A - B|}{\left(\sqrt{A} + \varepsilon \right) \left(\sqrt{B} + \varepsilon \right) \left(\sqrt{A} + \sqrt{B} \right)} \leq \frac{|A - B|}{2R \left(R + \varepsilon \right)^2}. \tag{2.4}$$

But since

$$\begin{aligned}
&\left| \left(\nabla_j E_k \left(\tilde{\boldsymbol{\theta}}(t_k) \right) \right)^2 - \left(\nabla_j E_k \left(\boldsymbol{\theta}^{(k)} \right) \right)^2 \right| \\
&= \left| \nabla_j E_k \left(\tilde{\boldsymbol{\theta}}(t_k) \right) - \nabla_j E_k \left(\boldsymbol{\theta}^{(k)} \right) \right| \cdot \left| \nabla_j E_k \left(\tilde{\boldsymbol{\theta}}(t_k) \right) + \nabla_j E_k \left(\boldsymbol{\theta}^{(k)} \right) \right| \\
&\leq 2M_1 \left| \nabla_j E_k \left(\tilde{\boldsymbol{\theta}}(t_k) \right) - \nabla_j E_k \left(\boldsymbol{\theta}^{(k)} \right) \right| \leq 2M_1 M_2 \sqrt{p} \left\| \tilde{\boldsymbol{\theta}}(t_k) - \boldsymbol{\theta}^{(k)} \right\|,
\end{aligned}$$

we have

$$|A - B| \leq 2M_1 M_2 \sqrt{p} \sum_{k=0}^{n} \rho^{n-k} (1 - \rho) \left\| \tilde{\boldsymbol{\theta}}(t_k) - \boldsymbol{\theta}^{(k)} \right\|. \tag{2.5}$$

Combining (2.4) and (2.5), we obtain

$$\begin{aligned}
&\left| \frac{\nabla_j E_n \left(\tilde{\boldsymbol{\theta}}(t_n) \right)}{\sqrt{A} + \varepsilon} - \frac{\nabla_j E_n \left(\boldsymbol{\theta}^{(n)} \right)}{\sqrt{B} + \varepsilon} \right| \\
&\leq \left| \nabla_j E_n \left(\tilde{\boldsymbol{\theta}}(t_n) \right) \right| \cdot \left| \frac{1}{\sqrt{A} + \varepsilon} - \frac{1}{\sqrt{B} + \varepsilon} \right| + \frac{\left| \nabla_j E_n \left(\tilde{\boldsymbol{\theta}}(t_n) \right) - \nabla_j E_n \left(\boldsymbol{\theta}^{(n)} \right) \right|}{\sqrt{B} + \varepsilon} \\
&\leq M_1 \cdot \frac{2M_1 M_2 \sqrt{p} \sum_{k=0}^{n} \rho^{n-k} (1 - \rho) \left\| \tilde{\boldsymbol{\theta}}(t_k) - \boldsymbol{\theta}^{(k)} \right\|}{2R(R + \varepsilon)^2} + \frac{M_2 \sqrt{p} \left\| \tilde{\boldsymbol{\theta}}(t_n) - \boldsymbol{\theta}^{(n)} \right\|}{R + \varepsilon} \\
&= \frac{M_1^2 M_2 \sqrt{p}}{R(R + \varepsilon)^2} \sum_{k=0}^{n} \rho^{n-k} (1 - \rho) \left\| \tilde{\boldsymbol{\theta}}(t_k) - \boldsymbol{\theta}^{(k)} \right\| + \frac{M_2 \sqrt{p}}{R + \varepsilon} \left\| \tilde{\boldsymbol{\theta}}(t_n) - \boldsymbol{\theta}^{(n)} \right\| \\
&\overset{(a)}{\leq} \frac{M_1^2 M_2 \sqrt{p}}{R(R + \varepsilon)^2} \sum_{k=0}^{n} \rho^{n-k} (1 - \rho) d_1 e^{d_2 kh} h^2 + \frac{M_2 \sqrt{p}}{R + \varepsilon} d_1 e^{d_2 nh} h^2,
\end{aligned} \tag{2.6}$$

where in (a) we used the induction hypothesis and that the bound on $\|\mathbf{e}_n\|$ is already proven.

Now note that since $0 < \rho e^{-d_2 h} \leq \rho$, we have $\sum_{k=0}^{n} \left(\rho e^{-d_2 h} \right)^k \leq \sum_{k=0}^{\infty} \rho^k = \frac{1}{1-\rho}$, which is rewritten as

$$\sum_{k=0}^{n} \rho^{n-k} (1 - \rho) e^{d_2 kh} \leq e^{d_2 nh}.$$

Then we can continue (2.6):

$$\left| \frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t_n)\right)}{\sqrt{A}+\varepsilon} - \frac{\nabla_j E_n\left(\boldsymbol{\theta}^{(n)}\right)}{\sqrt{B}+\varepsilon} \right| \leq \frac{M_2\sqrt{p}}{R+\varepsilon}\left(\frac{M_1^2}{R(R+\varepsilon)}+1\right) d_1 e^{d_2 nh} h^2 \tag{2.7}$$

Again using $1 \leq e^{d_2 nh}$, we conclude from (2.3) and (2.7) that

$$\|\mathbf{e}_{n+1}-\mathbf{e}_n\| \leq \underbrace{\left(C_1 + \frac{M_2\sqrt{p}}{R+\varepsilon}\left(\frac{M_1^2}{R(R+\varepsilon)}+1\right) d_1\right)}_{\leq d_3} \sqrt{p}\, e^{d_2 nh} h^3,$$

finishing the induction step. $\square$

SA-2.5 RMSProp with ε outside: full-batch. In the full-batch setting $E_k \equiv E$, the terms in (2.2) simplify to

$$\begin{aligned}
R_j^{(n)}(\boldsymbol{\theta}) &= \left|\nabla_j E(\boldsymbol{\theta})\right| \sqrt{1-\rho^{n+1}}, \\
P_j^{(n)}(\boldsymbol{\theta}) &= \sum_{k=0}^{n} \rho^{n-k}(1-\rho)\nabla_j E(\boldsymbol{\theta}) \sum_{i=1}^{p} \nabla_{ij} E(\boldsymbol{\theta}) \sum_{l=k}^{n-1} \frac{\nabla_i E(\boldsymbol{\theta})}{\left|\nabla_i E(\boldsymbol{\theta})\right| \sqrt{1-\rho^{l+1}}+\varepsilon}, \\
\bar{P}_j^{(n)}(\boldsymbol{\theta}) &= \left(1-\rho^{n+1}\right) \nabla_j E(\boldsymbol{\theta}) \sum_{i=1}^{p} \nabla_{ij} E(\boldsymbol{\theta}) \frac{\nabla_i E(\boldsymbol{\theta})}{\left|\nabla_i E(\boldsymbol{\theta})\right| \sqrt{1-\rho^{n+1}}+\varepsilon}.
\end{aligned}$$

If ε is small and the iteration number n is large, (2.2) simplifies to

$$\begin{aligned}
\dot{\tilde{\theta}}_j(t) &= -\operatorname{sign} \nabla_j E(\tilde{\boldsymbol{\theta}}(t)) + h\frac{\rho}{1-\rho} \cdot \frac{\sum_{i=1}^{p} \nabla_{ij} E(\tilde{\boldsymbol{\theta}}(t)) \operatorname{sign} \nabla_i E(\tilde{\boldsymbol{\theta}}(t))}{\left|\nabla_j E(\tilde{\boldsymbol{\theta}}(t))\right|} \\
&= \left|\nabla_j E(\tilde{\boldsymbol{\theta}}(t))\right|^{-1} \left[-\nabla_j E(\tilde{\boldsymbol{\theta}}(t)) + h\frac{\rho}{1-\rho}\nabla_j \left\|\nabla E(\tilde{\boldsymbol{\theta}}(t))\right\|_1\right].
\end{aligned}$$

3 RMSProp with ε inside the square root

Definition SA-3.1. In this section, for some $\boldsymbol{\theta}^{(0)} \in \mathbb{R}^p$, $\nu^{(0)} = \mathbf{0} \in \mathbb{R}^p$, $\rho \in (0,1)$, let the sequence of p-vectors $\left\{\boldsymbol{\theta}^{(k)}\right\}_{k\in\mathbb{Z}_{\geq 0}}$ be defined for $n \geq 0$ by

$$\begin{aligned}
\nu_j^{(n+1)} &= \rho\nu_j^{(n)} + (1-\rho)\left(\nabla_j E_n\left(\boldsymbol{\theta}^{(n)}\right)\right)^2, \\
\theta_j^{(n+1)} &= \theta_j^{(n)} - \frac{h}{\sqrt{\nu_j^{(n+1)}+\varepsilon}} \nabla_j E_n\left(\boldsymbol{\theta}^{(n)}\right).
\end{aligned} \tag{3.1}$$

Let $\tilde{\boldsymbol{\theta}}(t)$ be defined as a continuous solution to the piecewise ODE

$$\begin{aligned}
\dot{\tilde{\theta}}_j(t) &= -\frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)} \\
&\quad + h\left(\frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\left(2P_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\bar{P}_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)}{2R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)^3} - \frac{\sum_{i=1}^{p} \nabla_{ij} E_n\left(\tilde{\boldsymbol{\theta}}(t)\right) \frac{\nabla_i E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)}}{2R_j^{(n)}(\tilde{\boldsymbol{\theta}}(t))}\right).
\end{aligned} \tag{3.2}$$

with the initial condition $\tilde{\boldsymbol{\theta}}(0) = \boldsymbol{\theta}^{(0)}$, where $\mathbf{R}^{(n)}(\boldsymbol{\theta})$, $\mathbf{P}^{(n)}(\boldsymbol{\theta})$ and $\bar{\mathbf{P}}^{(n)}(\boldsymbol{\theta})$ are p-dimensional functions with components

$$\begin{aligned} R_j^{(n)}(\boldsymbol{\theta}) &:= \sqrt{\sum_{k=0}^{n} \rho^{n-k}(1-\rho)\left(\nabla_j E_k(\boldsymbol{\theta})\right)^2 + \varepsilon}, \\ P_j^{(n)}(\boldsymbol{\theta}) &:= \sum_{k=0}^{n} \rho^{n-k}(1-\rho)\nabla_j E_k\left(\boldsymbol{\theta}\right) \sum_{i=1}^{p} \nabla_{ij} E_k\left(\boldsymbol{\theta}\right) \sum_{l=k}^{n-1} \frac{\nabla_i E_l\left(\boldsymbol{\theta}\right)}{R_i^{(l)}\left(\boldsymbol{\theta}\right)}, \\ \bar{P}_j^{(n)}(\boldsymbol{\theta}) &:= \sum_{k=0}^{n} \rho^{n-k}(1-\rho)\nabla_j E_k\left(\boldsymbol{\theta}\right) \sum_{i=1}^{p} \nabla_{ij} E_k\left(\boldsymbol{\theta}\right) \frac{\nabla_i E_n\left(\boldsymbol{\theta}\right)}{R_i^{(n)}\left(\boldsymbol{\theta}\right)}. \end{aligned} \tag{3.3}$$

Assumption SA-3.2. For some positive constants M_1, M_2, M_3, M_4 we have

$$\begin{aligned} &\sup_i \sup_k \sup_{\boldsymbol{\theta}} \left|\nabla_i E_k(\boldsymbol{\theta})\right| \le M_1, \\ &\sup_{i,j} \sup_k \sup_{\boldsymbol{\theta}} \left|\nabla_{ij} E_k(\boldsymbol{\theta})\right| \le M_2, \\ &\sup_{i,j,s} \sup_k \sup_{\boldsymbol{\theta}} \left|\nabla_{ijs} E_k(\boldsymbol{\theta})\right| \le M_3, \\ &\sup_{i,j,s,r} \sup_k \sup_{\boldsymbol{\theta}} \left|\nabla_{ijsr} E_k(\boldsymbol{\theta})\right| \le M_4. \end{aligned}$$

Theorem SA-3.3 (RMSProp with ε inside: local error bound)**.** *Suppose Assumption SA-3.2 holds. Then for all $n \in \{0, 1, \dots, \lfloor T/h \rfloor\}$*

$$\left| \tilde{\theta}_j(t_{n+1}) - \tilde{\theta}_j(t_n) + h \frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t_n)\right)}{\sqrt{\sum_{k=0}^{n} \rho^{n-k}(1-\rho)\left(\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_k)\right)\right)^2 + \varepsilon}} \right| \le C_2 h^3$$

for a positive constant C_2 depending on ρ, where $\tilde{\boldsymbol{\theta}}(t)$ is defined in Definition SA-3.1.

We omit the proof since it is essentially the same argument as for Theorem SA-2.3.

Theorem SA-3.4 (RMSProp with ε inside: global error bound)**.** *Suppose Assumption SA-3.2 holds. Then there exist positive constants d_4, d_5, d_6 such that for all $n \in \{0, 1, \dots, \lfloor T/h \rfloor\}$*

$$\|\mathbf{e}_n\| \le d_4 e^{d_5 n h} h^2 \quad \text{and} \quad \|\mathbf{e}_{n+1} - \mathbf{e}_n\| \le d_6 e^{d_5 n h} h^3,$$

where $\mathbf{e}_n := \tilde{\boldsymbol{\theta}}(t_n) - \boldsymbol{\theta}^{(n)}$; $\tilde{\boldsymbol{\theta}}(t)$ and $\left\{\boldsymbol{\theta}^{(k)}\right\}_{k \in \mathbb{Z}_{\ge 0}}$ are defined in Definition SA-3.1. The constants can be defined as

$$\begin{aligned} d_4 &:= C_2, \\ d_5 &:= \left[1 + \frac{M_2 \sqrt{p}}{\sqrt{\varepsilon}}\left(\frac{M_1^2}{\varepsilon} + 1\right) d_4\right] \sqrt{p}, \\ d_6 &:= C_2 d_5. \end{aligned}$$

We omit the proof since it is essentially the same argument as for Theorem SA-2.4.

4 Adam with ε outside the square root

Definition SA-4.1. In this section, for some $\boldsymbol{\theta}^{(0)} \in \mathbb{R}^p$, $\nu^{(0)} = \mathbf{0} \in \mathbb{R}^p$, $\beta, \rho \in (0,1)$, let the sequence of p-vectors $\left\{\boldsymbol{\theta}^{(k)}\right\}_{k \in \mathbb{Z}_{\geq 0}}$ be defined for $n \geq 0$ by

$$\begin{aligned}
\nu_j^{(n+1)} &= \rho \nu_j^{(n)} + (1-\rho)\left(\nabla_j E_n\left(\boldsymbol{\theta}^{(n)}\right)\right)^2, \\
m_j^{(n+1)} &= \beta m_j^{(n)} + (1-\beta)\nabla_j E_n\left(\boldsymbol{\theta}^{(n)}\right), \\
\theta_j^{(n+1)} &= \theta_j^{(n)} - h\frac{m_j^{(n+1)}/\left(1-\beta^{n+1}\right)}{\sqrt{\nu_j^{(n+1)}/\left(1-\rho^{n+1}\right)} + \varepsilon}
\end{aligned}$$

or, rewriting,

$$\theta_j^{(n+1)} = \theta_j^{(n)} - h\frac{\frac{1}{1-\beta^{n+1}}\sum_{k=0}^n \beta^{n-k}\left(1-\beta\right)\nabla_j E_k\left(\boldsymbol{\theta}^{(k)}\right)}{\sqrt{\frac{1}{1-\rho^{n+1}}\sum_{k=0}^n \rho^{n-k}(1-\rho)\left(\nabla_j E_k\left(\boldsymbol{\theta}^{(k)}\right)\right)^2} + \varepsilon}. \tag{4.1}$$

Let $\tilde{\boldsymbol{\theta}}(t)$ be defined as a continuous solution to the piecewise ODE

$$\begin{aligned}
\dot{\tilde{\theta}}_j(t) = &-\frac{M_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \varepsilon} \\
&+ h\left(\frac{M_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\left(2P_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \bar{P}_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)}{2\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \varepsilon\right)^2 R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)} - \frac{2L_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \bar{L}_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)}{2\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \varepsilon\right)}\right).
\end{aligned} \tag{4.2}$$

with the initial condition $\tilde{\boldsymbol{\theta}}(0) = \boldsymbol{\theta}^{(0)}$, where $\mathbf{R}^{(n)}(\boldsymbol{\theta})$, $\mathbf{P}^{(n)}(\boldsymbol{\theta})$, $\bar{\mathbf{P}}^{(n)}(\boldsymbol{\theta})$, $\mathbf{M}^{(n)}(\boldsymbol{\theta})$, $\mathbf{L}^{(n)}(\boldsymbol{\theta})$, $\bar{\mathbf{L}}^{(n)}(\boldsymbol{\theta})$ are p-dimensional functions with components

$$\begin{aligned}
R_j^{(n)}(\boldsymbol{\theta}) &:= \sqrt{\sum_{k=0}^n \rho^{n-k}(1-\rho)\left(\nabla_j E_k(\boldsymbol{\theta})\right)^2/\left(1-\rho^{n+1}\right)}, \\
M_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\beta^{n+1}}\sum_{k=0}^n \beta^{n-k}\left(1-\beta\right)\nabla_j E_k\left(\boldsymbol{\theta}\right), \\
L_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\beta^{n+1}}\sum_{k=0}^n \beta^{n-k}(1-\beta)\sum_{i=1}^p \nabla_{ij}E_k(\boldsymbol{\theta})\sum_{l=k}^{n-1}\frac{M_i^{(l)}(\boldsymbol{\theta})}{R_i^{(l)}(\boldsymbol{\theta}) + \varepsilon}, \\
\bar{L}_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\beta^{n+1}}\sum_{k=0}^n \beta^{n-k}(1-\beta)\sum_{i=1}^p \nabla_{ij}E_k(\boldsymbol{\theta})\frac{M_i^{(n)}(\boldsymbol{\theta})}{R_i^{(n)}(\boldsymbol{\theta}) + \varepsilon}, \\
P_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\rho^{n+1}}\sum_{k=0}^n \rho^{n-k}(1-\rho)\nabla_j E_k(\boldsymbol{\theta})\sum_{i=1}^p \nabla_{ij}E_k(\boldsymbol{\theta})\sum_{l=k}^{n-1}\frac{M_i^{(l)}(\boldsymbol{\theta})}{R_i^{(l)}(\boldsymbol{\theta}) + \varepsilon}, \\
\bar{P}_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\rho^{n+1}}\sum_{k=0}^n \rho^{n-k}(1-\rho)\nabla_j E_k(\boldsymbol{\theta})\sum_{i=1}^p \nabla_{ij}E_k(\boldsymbol{\theta})\frac{M_i^{(n)}(\boldsymbol{\theta})}{R_i^{(n)}(\boldsymbol{\theta}) + \varepsilon}.
\end{aligned} \tag{4.3}$$

Assumption SA-4.2.

1. For some positive constants M_1, M_2, M_3, M_4 we have

$$\sup_i \sup_k \sup_{\boldsymbol{\theta}} \left|\nabla_i E_k(\boldsymbol{\theta})\right| \leq M_1,$$

$$\sup_{i,j} \sup_k \sup_{\boldsymbol{\theta}} \left|\nabla_{ij} E_k(\boldsymbol{\theta})\right| \leq M_2,$$
$$\sup_{i,j,s} \sup_k \sup_{\boldsymbol{\theta}} \left|\nabla_{ijs} E_k(\boldsymbol{\theta})\right| \leq M_3,$$
$$\sup_{i,j,s,r} \sup_k \sup_{\boldsymbol{\theta}} \left|\nabla_{ijsr} E_k(\boldsymbol{\theta})\right| \leq M_4.$$

2. For some $R > 0$ we have for all $n \in \{0, 1, \dots, \lfloor T/h \rfloor\}$

$$R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right) \geq R, \quad \frac{1}{1-\rho^{n+1}} \sum_{k=0}^{n} \rho^{n-k}(1-\rho)\left(\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_k)\right)\right)^2 \geq R^2,$$

where $\tilde{\boldsymbol{\theta}}(t)$ is defined in Definition SA-4.1.

Theorem SA-4.3 (Adam with ε outside: local error bound)**.** *Suppose Assumption SA-4.2 holds. Then for all $n \in \{0, 1, \dots, \lfloor T/h \rfloor\}$*

$$\left|\tilde{\theta}_j(t_{n+1}) - \tilde{\theta}_j(t_n) + h \frac{\frac{1}{1-\beta^{n+1}} \sum_{k=0}^{n} \beta^{n-k}(1-\beta) \nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_k)\right)}{\sqrt{\frac{1}{1-\rho^{n+1}} \sum_{k=0}^{n} \rho^{n-k}(1-\rho)\left(\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_k)\right)\right)^2} + \varepsilon}\right| \leq C_3 h^3$$

for a positive constant C_3 depending on β and ρ.

We omit the proof since it is essentially the same argument as for Theorem SA-2.3.

Theorem SA-4.4 (Adam with ε outside: global error bound)**.** *Suppose Assumption SA-4.2 holds, and*

$$\frac{1}{1-\rho^{n+1}} \sum_{k=0}^{n} \rho^{n-k}(1-\rho)\left(\nabla_j E_k\left(\boldsymbol{\theta}^{(k)}\right)\right)^2 \geq R^2$$

for $\left\{\boldsymbol{\theta}^{(k)}\right\}_{k \in \mathbb{Z}_{\geq 0}}$ defined in Definition SA-4.1. Then there exist positive constants d_7, d_8, d_9 such that for all $n \in \{0, 1, \dots, \lfloor T/h \rfloor\}$

$$\|\mathbf{e}_n\| \leq d_7 e^{d_8 n h} h^2 \quad \text{and} \quad \|\mathbf{e}_{n+1} - \mathbf{e}_n\| \leq d_9 e^{d_8 n h} h^3,$$

where $\mathbf{e}_n := \tilde{\boldsymbol{\theta}}(t_n) - \boldsymbol{\theta}^{(n)}$. The constants can be defined as

$$d_7 := C_3,$$
$$d_8 := \left[1 + \frac{M_2\sqrt{p}}{R+\varepsilon}\left(\frac{M_1^2}{R(R+\varepsilon)} + 1\right) d_7\right]\sqrt{p},$$
$$d_9 := C_3 d_8.$$

Proof. Analogously to Theorem SA-2.4, we will prove this by induction over n.

The base case is $n = 0$. Indeed, $\mathbf{e}_0 = \tilde{\boldsymbol{\theta}}(0) - \boldsymbol{\theta}^{(0)} = \mathbf{0}$. Then the jth component of $\mathbf{e}_1 - \mathbf{e}_0$ is

$$\begin{aligned}
[\mathbf{e}_1 - \mathbf{e}_0]_j = [\mathbf{e}_1]_j &= \tilde{\theta}_j(t_1) - \theta_j^{(0)} + \frac{h \nabla_j E_0\left(\boldsymbol{\theta}^{(0)}\right)}{\left|\nabla_j E_0\left(\boldsymbol{\theta}^{(0)}\right)\right| + \varepsilon} \\
&= \tilde{\theta}_j(t_1) - \tilde{\theta}_j(t_0) + \frac{h \nabla_j E_0\left(\tilde{\boldsymbol{\theta}}(t_0)\right)}{\sqrt{\left(\nabla_j E_0\left(\tilde{\boldsymbol{\theta}}(t_0)\right)\right)^2} + \varepsilon}.
\end{aligned}$$

By Theorem SA-4.3, the absolute value of the right-hand side does not exceed $C_3 h^3$, which means $\|\mathbf{e}_1 - \mathbf{e}_0\| \le C_3 h^3 \sqrt{p}$. Since $C_3\sqrt{p} \le d_9$, the base case is proven.

Now suppose that for all $k = 0, 1, \ldots, n-1$ the claim

$$\|\mathbf{e}_k\| \le d_7 e^{d_8 kh} h^2 \quad \text{and} \quad \|\mathbf{e}_{k+1} - \mathbf{e}_k\| \le d_9 e^{d_8 kh} h^3$$

is proven. Then

$$\begin{aligned}
\|\mathbf{e}_n\| &\overset{\text{(a)}}{\le} \|\mathbf{e}_{n-1}\| + \|\mathbf{e}_n - \mathbf{e}_{n-1}\| \le d_7 e^{d_8(n-1)h} h^2 + d_9 e^{d_8(n-1)h} h^3 \\
&= d_7 e^{d_8(n-1)h} h^2 \left(1 + \frac{d_9}{d_7} h\right) \overset{\text{(b)}}{\le} d_7 e^{d_8(n-1)h} h^2 \left(1 + d_8 h\right) \\
&\overset{\text{(c)}}{\le} d_7 e^{d_8(n-1)h} h^2 \cdot e^{d_8 h} = d_7 e^{d_8 nh} h^2,
\end{aligned}$$

where (a) is by the triangle inequality, (b) is by $d_9/d_7 \le d_8$, in (c) we used $1 + x \le e^x$ for all $x \ge 0$.

Next, combining Theorem SA-4.3 with (4.1), we have

$$\left|[\mathbf{e}_{n+1} - \mathbf{e}_n]_j\right| \le C_3 h^3 + h \left|\frac{N'}{\sqrt{D'} + \varepsilon} - \frac{N''}{\sqrt{D''} + \varepsilon}\right|, \tag{4.4}$$

where to simplify notation we put

$$\begin{aligned}
N' &:= \frac{1}{1 - \beta^{n+1}} \sum_{k=0}^{n} \beta^{n-k} (1 - \beta) \nabla_j E_k \left(\boldsymbol{\theta}^{(k)}\right), \\
N'' &:= \frac{1}{1 - \beta^{n+1}} \sum_{k=0}^{n} \beta^{n-k} (1 - \beta) \nabla_j E_k \left(\tilde{\boldsymbol{\theta}}(t_k)\right), \\
D' &:= \frac{1}{1 - \rho^{n+1}} \sum_{k=0}^{n} \rho^{n-k} (1 - \rho) \left(\nabla_j E_k \left(\boldsymbol{\theta}^{(k)}\right)\right)^2, \\
D'' &:= \frac{1}{1 - \rho^{n+1}} \sum_{k=0}^{n} \rho^{n-k} (1 - \rho) \left(\nabla_j E_k \left(\tilde{\boldsymbol{\theta}}(t_k)\right)\right)^2.
\end{aligned}$$

Using $D' \ge R^2$, $D'' \ge R^2$, we have

$$\left|\frac{1}{\sqrt{D'} + \varepsilon} - \frac{1}{\sqrt{D''} + \varepsilon}\right| = \frac{\left|D' - D''\right|}{\left(\sqrt{D'} + \varepsilon\right)\left(\sqrt{D''} + \varepsilon\right)\left(\sqrt{D'} + \sqrt{D''}\right)} \le \frac{\left|D' - D''\right|}{2R\left(R + \varepsilon\right)^2}. \tag{4.5}$$

But since

$$\begin{aligned}
&\left|\left(\nabla_j E_k \left(\boldsymbol{\theta}^{(k)}\right)\right)^2 - \left(\nabla_j E_k \left(\tilde{\boldsymbol{\theta}}(t_k)\right)\right)^2\right| \\
&\quad = \left|\nabla_j E_k \left(\boldsymbol{\theta}^{(k)}\right) - \nabla_j E_k \left(\tilde{\boldsymbol{\theta}}(t_k)\right)\right| \cdot \left|\nabla_j E_k \left(\boldsymbol{\theta}^{(k)}\right) + \nabla_j E_k \left(\tilde{\boldsymbol{\theta}}(t_k)\right)\right| \\
&\quad \le 2M_1 \left|\nabla_j E_k \left(\boldsymbol{\theta}^{(k)}\right) - \nabla_j E_k \left(\tilde{\boldsymbol{\theta}}(t_k)\right)\right| \le 2M_1 M_2 \sqrt{p} \left\|\boldsymbol{\theta}^{(k)} - \tilde{\boldsymbol{\theta}}(t_k)\right\|,
\end{aligned}$$

we have

$$\left|D' - D''\right| \le \frac{2M_1 M_2 \sqrt{p}}{1 - \rho^{n+1}} \sum_{k=0}^{n} \rho^{n-k} (1 - \rho) \left\|\boldsymbol{\theta}^{(k)} - \tilde{\boldsymbol{\theta}}(t_k)\right\|. \tag{4.6}$$

Similarly,

$$\begin{aligned}
\left|N' - N''\right| &\le \frac{1}{1 - \beta^{n+1}} \sum_{k=0}^{n} \beta^{n-k} (1 - \beta) \left|\nabla_j E_k \left(\boldsymbol{\theta}^{(k)}\right) - \nabla_j E_k \left(\tilde{\boldsymbol{\theta}}(t_k)\right)\right| \\
&\le \frac{1}{1 - \beta^{n+1}} \sum_{k=0}^{n} \beta^{n-k} (1 - \beta) M_2 \sqrt{p} \left\|\boldsymbol{\theta}^{(k)} - \tilde{\boldsymbol{\theta}}(t_k)\right\|.
\end{aligned} \tag{4.7}$$

Combining (4.5), (4.6) and (4.7), we get

$$
\begin{aligned}
&\left|\frac{N'}{\sqrt{D'}+\varepsilon}-\frac{N''}{\sqrt{D''}+\varepsilon}\right| \le \left|N'\right| \cdot \left|\frac{1}{\sqrt{D'}+\varepsilon}-\frac{1}{\sqrt{D''}+\varepsilon}\right| + \frac{\left|N'-N''\right|}{\sqrt{D''}+\varepsilon} \\
&\le \frac{1}{1-\beta^{n+1}}\sum_{k=0}^{n}\beta^{n-k}(1-\beta)M_1 \cdot \frac{2M_1M_2\sqrt{p}}{2R(R+\varepsilon)^2\left(1-\rho^{n+1}\right)}\sum_{k=0}^{n}\rho^{n-k}(1-\rho)\left\|\boldsymbol{\theta}^{(k)}-\tilde{\boldsymbol{\theta}}(t_k)\right\| \\
&\quad + \frac{M_2\sqrt{p}}{(R+\varepsilon)\left(1-\beta^{n+1}\right)}\sum_{k=0}^{n}\beta^{n-k}(1-\beta)\left\|\boldsymbol{\theta}^{(k)}-\tilde{\boldsymbol{\theta}}(t_k)\right\| \\
&= \frac{M_1^2M_2\sqrt{p}}{R(R+\varepsilon)^2\left(1-\rho^{n+1}\right)}\sum_{k=0}^{n}\rho^{n-k}(1-\rho)\left\|\boldsymbol{\theta}^{(k)}-\tilde{\boldsymbol{\theta}}(t_k)\right\| \\
&\quad + \frac{M_2\sqrt{p}}{(R+\varepsilon)\left(1-\beta^{n+1}\right)}\sum_{k=0}^{n}\beta^{n-k}(1-\beta)\left\|\boldsymbol{\theta}^{(k)}-\tilde{\boldsymbol{\theta}}(t_k)\right\| \\
&\overset{\text{(a)}}{\le} \frac{M_1^2M_2\sqrt{p}}{R(R+\varepsilon)^2\left(1-\rho^{n+1}\right)}\sum_{k=0}^{n}\rho^{n-k}(1-\rho)d_7e^{d_8kh}h^2 \\
&\quad + \frac{M_2\sqrt{p}}{(R+\varepsilon)\left(1-\beta^{n+1}\right)}\sum_{k=0}^{n}\beta^{n-k}(1-\beta)d_7e^{d_8kh}h^2,
\end{aligned} \tag{4.8}
$$

where in (a) we used the induction hypothesis and that the bound on $\|\mathbf{e}_n\|$ is already proven.

Now note that since $0<\rho e^{-d_8h}<\rho$, we have $\sum_{k=0}^{n}\left(\rho e^{-d_8h}\right)^k \le \sum_{k=0}^{n}\rho^k = \left(1-\rho^{n+1}\right)/(1-\rho)$, which is rewritten as

$$
\frac{1}{1-\rho^{n+1}}\sum_{k=0}^{n}\rho^{n-k}(1-\rho)e^{d_8kh} \le e^{d_8nh}.
$$

By the same logic,

$$
\frac{1}{1-\beta^{n+1}}\sum_{k=0}^{n}\beta^{n-k}(1-\beta)e^{d_8kh} \le e^{d_8nh}.
$$

Then we can continue (4.8):

$$
\left|\frac{N'}{\sqrt{D'}+\varepsilon}-\frac{N''}{\sqrt{D''}+\varepsilon}\right| \le \frac{M_2\sqrt{p}}{R+\varepsilon}\left(\frac{M_1^2}{R(R+\varepsilon)}+1\right)d_7e^{d_8nh}h^2 \tag{4.9}
$$

Again using $1\le e^{d_8nh}$, we conclude from (4.4) and (4.9) that

$$
\|\mathbf{e}_{n+1}-\mathbf{e}_n\| \le \underbrace{\left(C_3+\frac{M_2\sqrt{p}}{R+\varepsilon}\left(\frac{M_1^2}{R(R+\varepsilon)}+1\right)d_7\right)\sqrt{p}}_{\le d_9}\, e^{d_8nh}h^3,
$$

finishing the induction step. ∎

5 Adam with ε inside the square root

Definition SA-5.1. In this section, for some $\boldsymbol{\theta}^{(0)}\in\mathbb{R}^p$, $\nu^{(0)}=\mathbf{0}\in\mathbb{R}^p$, $\beta,\rho\in(0,1)$, let the sequence of p-vectors $\left\{\boldsymbol{\theta}^{(k)}\right\}_{k\in\mathbb{Z}_{\ge0}}$ be defined for $n\ge0$ by

$$
\begin{aligned}
\nu_j^{(n+1)} &= \rho\nu_j^{(n)}+(1-\rho)\left(\nabla_jE_n\left(\boldsymbol{\theta}^{(n)}\right)\right)^2, \\
m_j^{(n+1)} &= \beta m_j^{(n)}+(1-\beta)\nabla_jE_n\left(\boldsymbol{\theta}^{(n)}\right), \\
\theta_j^{(n+1)} &= \theta_j^{(n)}-h\frac{m_j^{(n+1)}/\left(1-\beta^{n+1}\right)}{\sqrt{\nu_j^{(n+1)}/\left(1-\rho^{n+1}\right)+\varepsilon}}.
\end{aligned} \tag{5.1}
$$

Let $\tilde{\boldsymbol{\theta}}(t)$ be defined as a continuous solution to the piecewise ODE

$$\begin{aligned}\dot{\tilde{\theta}}_j(t) = &-\frac{M_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)} \\ &+ h\left(\frac{M_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\left(2P_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\bar{P}_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)}{2R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)^3} - \frac{2L_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\bar{L}_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)}{2R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)}\right).\end{aligned} \tag{5.2}$$

with the initial condition $\tilde{\boldsymbol{\theta}}(0) = \boldsymbol{\theta}^{(0)}$, where $\mathbf{R}^{(n)}(\boldsymbol{\theta})$, $\mathbf{P}^{(n)}(\boldsymbol{\theta})$, $\bar{\mathbf{P}}^{(n)}(\boldsymbol{\theta})$, $\mathbf{M}^{(n)}(\boldsymbol{\theta})$, $\mathbf{L}^{(n)}(\boldsymbol{\theta})$, $\bar{\mathbf{L}}^{(n)}(\boldsymbol{\theta})$ are p-dimensional functions with components

$$\begin{aligned}
R_j^{(n)}(\boldsymbol{\theta}) &:= \sqrt{\sum_{k=0}^{n}\rho^{n-k}(1-\rho)\left(\nabla_j E_k(\boldsymbol{\theta})\right)^2/(1-\rho^{n+1})+\varepsilon},\\
M_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\beta^{n+1}}\sum_{k=0}^{n}\beta^{n-k}(1-\beta)\nabla_j E_k(\boldsymbol{\theta}),\\
L_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\beta^{n+1}}\sum_{k=0}^{n}\beta^{n-k}(1-\beta)\sum_{i=1}^{p}\nabla_{ij}E_k(\boldsymbol{\theta})\sum_{l=k}^{n-1}\frac{M_i^{(l)}(\boldsymbol{\theta})}{R_i^{(l)}(\boldsymbol{\theta})},\\
\bar{L}_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\beta^{n+1}}\sum_{k=0}^{n}\beta^{n-k}(1-\beta)\sum_{i=1}^{p}\nabla_{ij}E_k(\boldsymbol{\theta})\frac{M_i^{(n)}(\boldsymbol{\theta})}{R_i^{(n)}(\boldsymbol{\theta})},\\
P_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\rho^{n+1}}\sum_{k=0}^{n}\rho^{n-k}(1-\rho)\nabla_j E_k(\boldsymbol{\theta})\sum_{i=1}^{p}\nabla_{ij}E_k(\boldsymbol{\theta})\sum_{l=k}^{n-1}\frac{M_i^{(l)}(\boldsymbol{\theta})}{R_i^{(l)}(\boldsymbol{\theta})},\\
\bar{P}_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\rho^{n+1}}\sum_{k=0}^{n}\rho^{n-k}(1-\rho)\nabla_j E_k(\boldsymbol{\theta})\sum_{i=1}^{p}\nabla_{ij}E_k(\boldsymbol{\theta})\frac{M_i^{(n)}(\boldsymbol{\theta})}{R_i^{(n)}(\boldsymbol{\theta})}.
\end{aligned} \tag{5.3}$$

Assumption SA-5.2. For some positive constants M_1, M_2, M_3, M_4 we have

$$\begin{aligned}
&\sup_i \sup_k \sup_{\boldsymbol{\theta}} \left|\nabla_i E_k(\boldsymbol{\theta})\right| \le M_1,\\
&\sup_{i,j} \sup_k \sup_{\boldsymbol{\theta}} \left|\nabla_{ij} E_k(\boldsymbol{\theta})\right| \le M_2,\\
&\sup_{i,j,s} \sup_k \sup_{\boldsymbol{\theta}} \left|\nabla_{ijs} E_k(\boldsymbol{\theta})\right| \le M_3,\\
&\sup_{i,j,s,r} \sup_k \sup_{\boldsymbol{\theta}} \left|\nabla_{ijsr} E_k(\boldsymbol{\theta})\right| \le M_4.
\end{aligned}$$

Theorem SA-5.3 (Adam with ε inside: local error bound)**.** *Suppose Assumption SA-5.2 holds. Then for all $n \in \{0, 1, \ldots, \lfloor T/h \rfloor\}$*

$$\left|\tilde{\theta}_j(t_{n+1}) - \tilde{\theta}_j(t_n) + h\frac{\frac{1}{1-\beta^{n+1}}\sum_{k=0}^{n}\beta^{n-k}(1-\beta)\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_k)\right)}{\sqrt{\frac{1}{1-\rho^{n+1}}\sum_{k=0}^{n}\rho^{n-k}(1-\rho)\left(\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_k)\right)\right)^2+\varepsilon}}\right| \le C_4 h^3$$

for a positive constant C_4 depending on β and ρ.

We omit the proof since it is essentially the same argument as for Theorem SA-2.3.

Theorem SA-5.4 (Adam with ε inside: global error bound)**.** *Suppose Assumption SA-5.2 holds for $\left\{\boldsymbol{\theta}^{(k)}\right\}_{k\in\mathbb{Z}_{\ge 0}}$ defined in Definition SA-5.1. Then there exist positive constants d_{10}, d_{11}, d_{12} such that for all $n \in \{0, 1, \ldots, \lfloor T/h \rfloor\}$*

$$\|\mathbf{e}_n\| \le d_{10}e^{d_{11}nh}h^2 \quad \textit{and} \quad \|\mathbf{e}_{n+1}-\mathbf{e}_n\| \le d_{12}e^{d_{11}nh}h^3,$$

where $\mathbf{e}_n := \tilde{\boldsymbol{\theta}}(t_n) - \boldsymbol{\theta}^{(n)}$. *The constants can be defined as*

$$
\begin{aligned}
d_{10} &:= C_4, \\
d_{11} &:= \left[1 + \frac{M_2\sqrt{p}}{\sqrt{\varepsilon}}\left(\frac{M_1^2}{\varepsilon} + 1\right) d_{10}\right]\sqrt{p}, \\
d_{12} &:= C_4 d_{11}.
\end{aligned}
$$

6 Technical bounding lemmas

We will need the following lemmas to prove Theorem SA-2.3.

Lemma SA-6.1. *Suppose Assumption SA-2.2 holds. Then*

$$\sup_{\boldsymbol{\theta}} \left| P_j^{(n)}(\boldsymbol{\theta}) \right| \le C_5, \tag{6.1}$$

$$\sup_{\boldsymbol{\theta}} \left| \bar{P}_j^{(n)}(\boldsymbol{\theta}) \right| \le C_6, \tag{6.2}$$

with constants C_5, C_6 *defined as follows:*

$$
\begin{aligned}
C_5 &:= p\frac{M_1^2 M_2}{R+\varepsilon} \cdot \frac{\rho}{1-\rho}, \\
C_6 &:= p\frac{M_1^2 M_2}{R+\varepsilon}.
\end{aligned}
$$

Proof of Lemma SA-6.1. The proof is done in the following simple steps.

SA-6.2 Proof of (6.1). This bound is straightforward:

$$
\begin{aligned}
\sup_{\boldsymbol{\theta}} \left| P_j^{(n)}(\boldsymbol{\theta}) \right| &= \sup_{\boldsymbol{\theta}} \left| \sum_{k=0}^{n} \rho^{n-k}(1-\rho)\nabla_j E_k(\boldsymbol{\theta}) \sum_{i=1}^{p} \nabla_{ij} E_k(\boldsymbol{\theta}) \sum_{l=k}^{n-1} \frac{\nabla_i E_l(\boldsymbol{\theta})}{R_i^{(l)}(\boldsymbol{\theta}) + \varepsilon} \right| \\
&\le p\frac{M_1^2 M_2}{R+\varepsilon}(1-\rho)\sum_{k=0}^{n} \rho^{n-k}(n-k) \le p\frac{M_1^2 M_2}{R+\varepsilon}(1-\rho)\sum_{k=0}^{\infty} \rho^k k = C_5.
\end{aligned}
$$

SA-6.3 Proof of (6.2). This bound is straightforward:

$$
\begin{aligned}
\sup_{\boldsymbol{\theta}} \left| \bar{P}_j^{(n)}(\boldsymbol{\theta}) \right| &= \sup_{\boldsymbol{\theta}} \left| \sum_{k=0}^{n} \rho^{n-k}(1-\rho)\nabla_j E_k(\boldsymbol{\theta}) \sum_{i=1}^{p} \nabla_{ij} E_k(\boldsymbol{\theta}) \frac{\nabla_i E_n(\boldsymbol{\theta})}{R_i^{(n)}(\boldsymbol{\theta}) + \varepsilon} \right| \\
&\le p\frac{M_1^2 M_2}{R+\varepsilon}(1-\rho)\sum_{k=0}^{n} \rho^{n-k} \le p\frac{M_1^2 M_2}{R+\varepsilon} = C_6.
\end{aligned}
$$

This concludes the proof of Lemma SA-6.1. □

Lemma SA-6.4. *Suppose Assumption SA-2.2 holds. Then the first derivative of* $t \mapsto \tilde{\theta}_j(t)$ *is uniformly over* j *and* $t \in [0, T]$ *bounded in absolute value by some positive constant, say* D_1.

Proof. This follows immediately from $h \le T$, (6.1), (6.2) and the definition of $\tilde{\boldsymbol{\theta}}(t)$ given in (2.2). □

Lemma SA-6.5. *Suppose Assumption SA-2.2 holds. Then*

$$\sup_{t \in [0,T]} \sup_j \left| \left(\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)^{\cdot} \right| \le C_7, \tag{6.3}$$

$$\sup_{n,k}\sup_{t\in[t_n,t_{n+1}]}\left|\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\left[\dot{\tilde{\theta}}_i(t)+\frac{\nabla_i E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon}\right]\right|\leq C_8 h, \tag{6.4}$$

$$\sup_{k\leq n}\sup_{t\in[0,T]}\left|\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\sum_{l=k}^{n-1}\frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon}\right|\leq (n-k)C_9, \tag{6.5}$$

$$\left|\left(P_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)^{\cdot}\right|\leq C_{10}+C_{14}, \tag{6.6}$$

$$\left|\left(\bar{P}_j^{(n)}(\tilde{\boldsymbol{\theta}}(t))\right)^{\cdot}\right|\leq C_{15}, \tag{6.7}$$

$$\left|\left(\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\frac{\nabla_i E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon}\right)^{\cdot}\right|\leq C_{13}, \tag{6.8}$$

$$\left|\left(\frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\left(2P_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\bar{P}_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)}{2\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^2 R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)}\right)^{\cdot}\right|\leq C_{17}, \tag{6.9}$$

$$\left|\left(\frac{\sum_{i=1}^{p}\nabla_{ij}E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\frac{\nabla_i E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon}}{2\left(R_j^{(n)}(\tilde{\boldsymbol{\theta}}(t))+\varepsilon\right)}\right)^{\cdot}\right|\leq C_{18}, \tag{6.10}$$

with constants C_7, C_8, C_9, C_{10}, C_{11}, C_{12}, C_{13}, C_{14}, C_{15}, C_{16}, C_{17}, C_{18} defined as follows:

$$\begin{aligned}
C_7 &:= pM_2D_1,\\
C_8 &:= pM_2\left[\frac{M_1\left(2C_5+C_6\right)}{2(R+\varepsilon)^2R}+\frac{pM_1M_2}{2(R+\varepsilon)^2}\right],\\
C_9 &:= p\frac{M_1M_2}{R+\varepsilon},\\
C_{10} &:= D_1p^2\frac{M_1M_2^2}{R+\varepsilon}\cdot\frac{\rho}{1-\rho},\\
C_{11} &:= \frac{D_1pM_1M_2}{R},\\
C_{12} &:= D_1p^2\frac{M_1M_3}{R+\varepsilon},\\
C_{13} &:= C_{12}+pM_2\left(\frac{D_1pM_2}{R+\varepsilon}+\frac{M_1}{(R+\varepsilon)^2}C_{11}\right)\\
&= \frac{D_1p^2}{R+\varepsilon}\left(M_1M_3+M_2^2+\frac{M_1^2M_2^2}{(R+\varepsilon)R}\right),\\
C_{14} &:= M_1C_{13}\frac{\rho}{1-\rho},\\
C_{15} &:= \frac{D_1p^2M_1M_2^2}{R+\varepsilon}+\frac{D_1p^2M_1^2M_3}{R+\varepsilon}+\frac{D_1p^2M_1M_2^2}{R+\varepsilon}+\frac{pM_1^2M_2C_{11}}{(R+\varepsilon)^2},\\
C_{16} &:= \frac{2C_{11}}{R(R+\varepsilon)^3}+\frac{C_{11}}{(R+\varepsilon)^4},\\
C_{17} &:= \frac{D_1pM_2\cdot(2C_5+C_6)}{2\left(R+\varepsilon\right)^2R}+\frac{M_1\left(2\left(C_{10}+C_{14}\right)+C_{15}\right)}{2\left(R+\varepsilon\right)^2R}+\frac{M_1\left(2C_5+C_6\right)C_{16}}{2},
\end{aligned}$$

$$C_{18} := \frac{1}{2(R+\varepsilon)} \left(\frac{p^2 D_1 M_1 M_3}{R+\varepsilon} + \frac{p^2 D_1 M_2^2}{R+\varepsilon} + \frac{p M_1 M_2 C_{11}}{(R+\varepsilon)^2} \right) + \frac{1}{2} \cdot \frac{p M_1 M_2}{R+\varepsilon} \cdot \frac{C_{11}}{(R+\varepsilon)^2}.$$

Proof of Lemma SA-6.5. We divide this argument in several steps.

SA-6.6 Proof of (6.3). This bound is straightforward:

$$\left| \left(\nabla_j E_n \left(\tilde{\boldsymbol{\theta}}(t) \right) \right)^{\cdot} \right| = \left| \sum_{i=1}^{p} \nabla_{ij} E_n \left(\tilde{\boldsymbol{\theta}}(t) \right) \dot{\tilde{\theta}}_i(t) \right| \le C_7.$$

SA-6.7 Proof of (6.4). By (2.2) we have for $t = t_{n+1}^-$

$$\left| \dot{\tilde{\theta}}_j(t) + \frac{\nabla_j E_n \left(\tilde{\boldsymbol{\theta}}(t) \right)}{R_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon} \right| \le h \left[\frac{M_1 \left(2C_5 + C_6 \right)}{2(R+\varepsilon)^2 R} + \frac{p M_1 M_2}{2(R+\varepsilon)^2} \right],$$

giving (6.4) immediately.

SA-6.8 Proof of (6.5). This bound follows from the assumptions immediately.

SA-6.9 Proof of (6.6). We will prove this by bounding the two terms in the expression

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t} P_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t) \right) \\ &\quad = \sum_{k=0}^{n} \rho^{n-k} (1-\rho) \sum_{u=1}^{p} \nabla_{ju} E_k \left(\tilde{\boldsymbol{\theta}}(t) \right) \dot{\tilde{\theta}}_u(t) \sum_{i=1}^{p} \nabla_{ij} E_k \left(\tilde{\boldsymbol{\theta}}(t) \right) \sum_{l=k}^{n-1} \frac{\nabla_i E_l \left(\tilde{\boldsymbol{\theta}}(t) \right)}{R_i^{(l)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon} \\ &\qquad + \sum_{k=0}^{n} \rho^{n-k} (1-\rho) \nabla_j E_k \left(\tilde{\boldsymbol{\theta}}(t) \right) \sum_{i=1}^{p} \frac{\mathrm{d}}{\mathrm{d}t} \left\{ \nabla_{ij} E_k \left(\tilde{\boldsymbol{\theta}}(t) \right) \sum_{l=k}^{n-1} \frac{\nabla_i E_l \left(\tilde{\boldsymbol{\theta}}(t) \right)}{R_i^{(l)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon} \right\}. \end{aligned} \tag{6.11}$$

It is easily shown that the first term in (6.11) is bounded in absolute value by C_{10}:

$$\begin{aligned} &\left| \sum_{k=0}^{n} \rho^{n-k} (1-\rho) \sum_{u=1}^{p} \nabla_{ju} E_k \left(\tilde{\boldsymbol{\theta}}(t) \right) \dot{\tilde{\theta}}_u(t) \sum_{i=1}^{p} \nabla_{ij} E_k \left(\tilde{\boldsymbol{\theta}}(t) \right) \sum_{l=k}^{n-1} \frac{\nabla_i E_l \left(\tilde{\boldsymbol{\theta}}(t) \right)}{R_i^{(l)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon} \right| \\ &\quad \le D_1 p^2 \frac{M_1 M_2^2}{R+\varepsilon} (1-\rho) \sum_{k=0}^{n} \rho^k k \\ &\quad \le D_1 p^2 \frac{M_1 M_2^2}{R+\varepsilon} (1-\rho) \sum_{k=0}^{\infty} \rho^k k \\ &\quad = C_{10}. \end{aligned}$$

For the proof of (6.6), it is left to show that the second term in (6.11) is bounded in absolute value by C_{14}.

To bound $\sum_{i=1}^{p} \frac{\mathrm{d}}{\mathrm{d}t} \left\{ \nabla_{ij} E_k \left(\tilde{\boldsymbol{\theta}}(t) \right) \sum_{l=k}^{n-1} \frac{\nabla_i E_l \left(\tilde{\boldsymbol{\theta}}(t) \right)}{R_i^{(l)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon} \right\}$, we can use

$$\left| \sum_{i=1}^{p} \frac{\mathrm{d}}{\mathrm{d}t} \left\{ \nabla_{ij} E_k \left(\tilde{\boldsymbol{\theta}}(t) \right) \sum_{l=k}^{n-1} \frac{\nabla_i E_l \left(\tilde{\boldsymbol{\theta}}(t) \right)}{R_i^{(l)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon} \right\} \right|$$

$$
\leq \left| \sum_{i=1}^{p} \frac{\mathrm{d}}{\mathrm{d}t} \left\{ \nabla_{ij} E_k \left(\tilde{\boldsymbol{\theta}}(t) \right) \right\} \sum_{l=k}^{n-1} \frac{\nabla_i E_l \left(\tilde{\boldsymbol{\theta}}(t) \right)}{R_i^{(l)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon} \right|
$$

$$
+ \left| \sum_{i=1}^{p} \nabla_{ij} E_k \left(\tilde{\boldsymbol{\theta}}(t) \right) \sum_{l=k}^{n-1} \frac{\mathrm{d}}{\mathrm{d}t} \left\{ \frac{\nabla_i E_l \left(\tilde{\boldsymbol{\theta}}(t) \right)}{R_i^{(l)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon} \right\} \right|
$$

By the Cauchy-Schwarz inequality applied twice,

$$
\begin{aligned}
&\left| \sum_{i=1}^{p} \frac{\mathrm{d}}{\mathrm{d}t} \left\{ \nabla_{ij} E_k \left(\tilde{\boldsymbol{\theta}}(t) \right) \right\} \sum_{l=k}^{n-1} \frac{\nabla_i E_l \left(\tilde{\boldsymbol{\theta}}(t) \right)}{R_i^{(l)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon} \right| \\
&\leq \sqrt{\sum_{i=1}^{p} \sum_{s=1}^{p} \left(\nabla_{ijs} E_k \left(\tilde{\boldsymbol{\theta}}(t) \right) \right)^2} \sqrt{\sum_{u=1}^{p} \dot{\tilde{\theta}}_u(t)^2} \sqrt{\sum_{i=1}^{p} \left| \sum_{l=k}^{n-1} \frac{\nabla_i E_l \left(\tilde{\boldsymbol{\theta}}(t) \right)}{R_i^{(l)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon} \right|^2} \\
&\leq M_3 p \cdot D_1 \sqrt{p} \cdot \sqrt{\sum_{i=1}^{p} \left| \sum_{l=k}^{n-1} \frac{\nabla_i E_l \left(\tilde{\boldsymbol{\theta}}(t) \right)}{R_i^{(l)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon} \right|^2} \leq (n-k) C_{12}.
\end{aligned}
$$

Next, for any n and j

$$
\begin{aligned}
\left| \frac{\mathrm{d}}{\mathrm{d}t} R_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t) \right) \right| &= \frac{1}{R_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t) \right)} \left| \sum_{k=0}^{n} \rho^{n-k} (1-\rho) \nabla_j E_k \left(\tilde{\boldsymbol{\theta}}(t) \right) \sum_{i=1}^{p} \nabla_{ij} E_k \left(\tilde{\boldsymbol{\theta}}(t) \right) \dot{\tilde{\theta}}_i(t) \right| \\
&\leq \frac{1}{R_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t) \right)} D_1 p M_1 M_2 \sum_{k=0}^{n} \rho^{n-k} (1-\rho) \leq C_{11}.
\end{aligned} \tag{6.12}
$$

This gives

$$
\begin{aligned}
\left| \frac{\mathrm{d}}{\mathrm{d}t} \left\{ \frac{\nabla_i E_l \left(\tilde{\boldsymbol{\theta}}(t) \right)}{R_i^{(l)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon} \right\} \right| &\leq \frac{\left| \sum_{s=1}^{p} \nabla_{is} E_l \left(\tilde{\boldsymbol{\theta}}(t) \right) \dot{\tilde{\theta}}_s(t) \right|}{R_i^{(l)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon} + \frac{\left| \nabla_i E_l \left(\tilde{\boldsymbol{\theta}}(t) \right) \right| \cdot \left| \frac{\mathrm{d}}{\mathrm{d}t} R_i^{(l)} \left(\tilde{\boldsymbol{\theta}}(t) \right) \right|}{\left(R_i^{(l)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon \right)^2} \\
&\leq \frac{D_1 p M_2}{R + \varepsilon} + \frac{M_1}{(R+\varepsilon)^2} C_{11}.
\end{aligned}
$$

We have obtained

$$
\left| \sum_{i=1}^{p} \frac{\mathrm{d}}{\mathrm{d}t} \left\{ \nabla_{ij} E_k \left(\tilde{\boldsymbol{\theta}}(t) \right) \sum_{l=k}^{n-1} \frac{\nabla_i E_l \left(\tilde{\boldsymbol{\theta}}(t) \right)}{R_i^{(l)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon} \right\} \right| \leq (n-k) C_{13}. \tag{6.13}
$$

This gives a bound on the second term in (6.11):

$$
\begin{aligned}
&\left| \sum_{k=0}^{n} \rho^{n-k} (1-\rho) \nabla_j E_k \left(\tilde{\boldsymbol{\theta}}(t) \right) \sum_{i=1}^{p} \frac{\mathrm{d}}{\mathrm{d}t} \left\{ \nabla_{ij} E_k \left(\tilde{\boldsymbol{\theta}}(t) \right) \sum_{l=k}^{n-1} \frac{\nabla_i E_l \left(\tilde{\boldsymbol{\theta}}(t) \right)}{R_i^{(l)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon} \right\} \right| \\
&\leq M_1 \sum_{k=0}^{n} \rho^{n-k} (1-\rho)(n-k) C_{13} \leq C_{14},
\end{aligned}
$$

concluding the proof of (6.6).

SA-6.10 Proof of (6.7). We will prove this by bounding the four terms in the expression

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\{\sum_{k=0}^{n}\rho^{n-k}(1-\rho)\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\frac{\nabla_i E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon}\right\}$$
$$= \text{Term1} + \text{Term2} + \text{Term3} + \text{Term4},$$

where

$$\text{Term1} := \sum_{k=0}^{n}\rho^{n-k}(1-\rho)\frac{\mathrm{d}}{\mathrm{d}t}\left\{\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\right\}\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\frac{\nabla_i E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon},$$

$$\text{Term2} := \sum_{k=0}^{n}\rho^{n-k}(1-\rho)\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\sum_{i=1}^{p}\frac{\mathrm{d}}{\mathrm{d}t}\left\{\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\right\}\frac{\nabla_i E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon},$$

$$\text{Term3} := \sum_{k=0}^{n}\rho^{n-k}(1-\rho)\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\frac{\frac{\mathrm{d}}{\mathrm{d}t}\left\{\nabla_i E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\right\}}{R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon},$$

$$\text{Term4} := -\sum_{k=0}^{n}\rho^{n-k}(1-\rho)\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\frac{\nabla_i E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\frac{\mathrm{d}}{\mathrm{d}t}R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)}{\left(R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^2}.$$

To bound Term1, use $\left|\frac{\mathrm{d}}{\mathrm{d}t}\left\{\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\right\}\right| \le D_1 p M_2$, giving

$$|\text{Term1}| \le \frac{D_1 p^2 M_1 M_2^2}{R+\varepsilon}\sum_{k=0}^{n}\rho^{n-k}(1-\rho) \le \frac{D_1 p^2 M_1 M_2^2}{R+\varepsilon}.$$

To bound Term2, use $\left|\frac{\mathrm{d}}{\mathrm{d}t}\left\{\nabla_{ij} E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\right\}\right| \le D_1 p M_3$, giving

$$|\text{Term2}| \le \frac{D_1 p^2 M_1^2 M_3}{R+\varepsilon}\sum_{k=0}^{n}\rho^{n-k}(1-\rho) \le \frac{D_1 p^2 M_1^2 M_3}{R+\varepsilon}.$$

To bound Term3, use $\left|\frac{\mathrm{d}}{\mathrm{d}t}\left\{\nabla_i E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\right\}\right| \le D_1 p M_2$, giving

$$|\text{Term3}| \le \frac{D_1 p^2 M_1 M_2^2}{R+\varepsilon}\sum_{k=0}^{n}\rho^{n-k}(1-\rho) \le \frac{D_1 p^2 M_1 M_2^2}{R+\varepsilon}.$$

To bound Term4, use (6.12), giving

$$|\text{Term4}| \le \frac{p M_1^2 M_2 C_{11}}{(R+\varepsilon)^2}\sum_{k=0}^{n}\rho^{n-k}(1-\rho) \le \frac{p M_1^2 M_2 C_{11}}{(R+\varepsilon)^2}.$$

SA-6.11 Proof of (6.8). This is proven in (6.13).

SA-6.12 Proof of (6.9)**.** (6.12) gives

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}\left\{\frac{1}{R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)}\right\}\right| = \frac{\left|\frac{\mathrm{d}}{\mathrm{d}t}R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right|}{R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)^2} \leq \frac{C_{11}}{R^2}, \tag{6.14}$$

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}\left\{\frac{1}{R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon}\right\}\right| = \frac{\left|\frac{\mathrm{d}}{\mathrm{d}t}R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right|}{\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^2} \leq \frac{C_{11}}{(R+\varepsilon)^2}, \tag{6.15}$$

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}\left\{\frac{1}{\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^2}\right\}\right| = \frac{2\left|\frac{\mathrm{d}}{\mathrm{d}t}R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right|}{\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^3} \leq \frac{2C_{11}}{(R+\varepsilon)^3}. \tag{6.16}$$

Combining two bounds above, we have

$$\begin{aligned}
&\left|\frac{\mathrm{d}}{\mathrm{d}t}\left\{\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^{-2}R_j^{(n)}(\tilde{\boldsymbol{\theta}}(t))^{-1}\right\}\right| \\
&\quad\leq \frac{\left|\frac{\mathrm{d}}{\mathrm{d}t}\left\{\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^{-2}\right\}\right|}{R_j^{(n)}(\tilde{\boldsymbol{\theta}}(t))} + \frac{\left|\frac{\mathrm{d}}{\mathrm{d}t}\left\{R_j^{(n)}(\tilde{\boldsymbol{\theta}}(t))^{-1}\right\}\right|}{\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^2} \leq C_{16}.
\end{aligned}$$

We are ready to bound

$$\begin{aligned}
&\left|\left(\frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\left(2P_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\bar{P}_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)}{2\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^2 R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)}\right)^{\cdot}\right| \\
&\quad\leq \left|\frac{\left(\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)^{\cdot}\left(2P_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\bar{P}_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)}{2\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^2 R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)}\right| + \\
&\quad\quad + \left|\frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\left(2P_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\bar{P}_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)^{\cdot}}{2\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^2 R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)}\right| \\
&\quad\quad + \left|\frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\left(2P_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\bar{P}_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)}{2}\right. \\
&\quad\quad\quad \left.\times\left(\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^{-2}R_j^{(n)}(\tilde{\boldsymbol{\theta}}(t))^{-1}\right)^{\cdot}\right| \leq C_{17}.
\end{aligned}$$

SA-6.13 Proof of (6.10)**.** Since

$$\left|\sum_{i=1}^{p} \nabla_{ij} E_n\left(\tilde{\boldsymbol{\theta}}(t)\right) \frac{\nabla_i E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon}\right| \leq \frac{pM_1M_2}{R+\varepsilon}$$

and, as we have already seen in the argument for (6.7),

$$\left|\left(\sum_{i=1}^{p} \nabla_{ij} E_n\left(\tilde{\boldsymbol{\theta}}(t)\right) \frac{\nabla_i E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon}\right)^{\cdot}\right| \leq \frac{p^2D_1M_1M_3}{R+\varepsilon}+\frac{p^2D_1M_2^2}{R+\varepsilon}+\frac{pM_1M_2C_{11}}{(R+\varepsilon)^2},$$

we are ready to bound

$$\left|\left(\frac{\sum_{i=1}^{p} \nabla_{ij} E_n\left(\tilde{\boldsymbol{\theta}}(t)\right) \frac{\nabla_i E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon}}{2\left(R_j^{(n)}(\tilde{\boldsymbol{\theta}}(t))+\varepsilon\right)}\right)^{\cdot}\right| \leq C_{18}.$$

The proof of Lemma SA-6.5 is concluded. □

Lemma SA-6.14. *Suppose Assumption SA-2.2 holds. Then the second derivative of* $t \mapsto \tilde{\theta}_j(t)$ *is uniformly over* j *and* $t \in [0,T]$ *bounded in absolute value by some positive constant, say* D_2*.*

Proof. This follows from the definition of $\tilde{\boldsymbol{\theta}}(t)$ given in (2.2), $h \leq T$ and that the first derivatives of all three terms in (2.2) are bounded by Lemma SA-6.5. □

Lemma SA-6.15. *Suppose Assumption SA-2.2 holds. Then*

$$\left|\left(\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)^{\cdot\cdot}\right| \leq C_{19}, \tag{6.17}$$

$$\left|\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)^{\cdot\cdot}\right| \leq C_{20}, \tag{6.18}$$

$$\left|\left(\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^{-2}\right)^{\cdot\cdot}\right| \leq C_{21}, \tag{6.19}$$

$$\left|\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)^{-1}\right)^{\cdot\cdot}\right| \leq C_{22}, \tag{6.20}$$

$$\left|\left(\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^{-2} R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)^{-1}\right)^{\cdot\cdot}\right| \leq C_{23}, \tag{6.21}$$

$$\left|\left(\sum_{i=1}^{p} \nabla_{ij} E_k\left(\tilde{\boldsymbol{\theta}}(t)\right) \sum_{l=k}^{n-1} \frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon}\right)^{\cdot\cdot}\right| \leq (n-k)C_{24}, \tag{6.22}$$

with constants C_{19}*,* C_{20}*,* C_{21}*,* C_{22}*,* C_{23}*,* C_{24} *defined as follows:*

$$\begin{aligned}
C_{19} &:= p^2M_3D_1^2 + pM_2D_2,\\
C_{20} &:= \frac{C_{11}}{R^2}pM_1M_2D_1 + \frac{1}{R}p^2M_2^2D_1^2 + \frac{1}{R}p^2M_1M_3D_1^2 + \frac{1}{R}pM_1M_2D_2,\\
C_{21} &:= \frac{6C_{11}^2}{(R+\varepsilon)^4} + \frac{2C_{20}}{(R+\varepsilon)^3},\\
C_{22} &:= \frac{2C_{11}^2}{R^3} + \frac{C_{20}}{R^2},
\end{aligned}$$

$$C_{23} := \frac{C_{21}}{R} + \frac{4C_{11}^2}{R^2(R+\varepsilon)^3} + \frac{C_{22}}{(R+\varepsilon)^2},$$

$$C_{24} := p\left[\frac{2C_{11}\left(D_1M_2^2p + D_1M_1M_3p\right)}{(R+\varepsilon)^2} + M_1M_2\left(\frac{2C_{11}^2}{(R+\varepsilon)^3} + \frac{C_{20}}{(R+\varepsilon)^2}\right)\right.$$
$$\left. + \frac{2D_1^2M_2M_3p^2 + M_2\left(D_1^2M_3p^2 + D_2M_2p\right) + M_1\left(D_1^2M_4p^2 + D_2M_3p\right)}{R+\varepsilon}\right].$$

Proof of Lemma SA-6.15. We divide this argument in several steps.

SA-6.16 Proof of (6.17). This bound is straightforward:

$$\left|\left(\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)^{\cdot\cdot}\right| = \left|\sum_{i=1}^{p}\sum_{s=1}^{p}\nabla_{ijs}E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\dot{\tilde{\theta}}_s(t)\dot{\tilde{\theta}}_i(t) + \sum_{i=1}^{p}\nabla_{ij}E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\ddot{\tilde{\theta}}_t(t)\right| \le C_{19}.$$

SA-6.17 Proof of (6.18). Note that

$$\begin{aligned}
\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)^{\cdot\cdot} &= \left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)^{-1}\right)^{\cdot}\sum_{k=0}^{n}\rho^{n-k}(1-\rho)\nabla_jE_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\dot{\tilde{\theta}}_i(t)\\
&+ R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)^{-1}\sum_{k=0}^{n}\rho^{n-k}(1-\rho)\left(\nabla_jE_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)^{\cdot}\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\dot{\tilde{\theta}}_i(t)\\
&+ R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)^{-1}\sum_{k=0}^{n}\rho^{n-k}(1-\rho)\nabla_jE_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\sum_{i=1}^{p}\left(\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)^{\cdot}\dot{\tilde{\theta}}_i(t)\\
&+ R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)^{-1}\sum_{k=0}^{n}\rho^{n-k}(1-\rho)\nabla_jE_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\ddot{\tilde{\theta}}_i(t),
\end{aligned}$$

giving by (6.14)

$$\begin{aligned}
\left|\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)^{\cdot\cdot}\right| &\le \frac{C_{11}}{R^2}pM_1M_2D_1\sum_{k=0}^{n}\rho^{n-k}(1-\rho) + \frac{1}{R}p^2M_2^2D_1^2\sum_{k=0}^{n}\rho^{n-k}(1-\rho)\\
&+ \frac{1}{R}p^2M_1M_3D_1^2\sum_{k=0}^{n}\rho^{n-k}(1-\rho) + \frac{1}{R}pM_1M_2D_2\sum_{k=0}\rho^{n-k}(1-\rho)\\
&\le C_{20}.
\end{aligned}$$

SA-6.18 Proof of (6.19). Note that

$$\left(\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^{-2}\right)^{\cdot\cdot} = \frac{6\left(\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)^{\cdot}\right)^2}{\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^4} - \frac{2\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)^{\cdot\cdot}}{\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^3},$$

giving by (6.12) and (6.18)

$$\left|\left(\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^{-2}\right)^{\cdot\cdot}\right| \le C_{21}.$$

SA-6.19 Proof of (6.20). The bound follows from (6.12), (6.18) and

$$\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)^{-1}\right)^{\cdot\cdot} = \frac{2\left(\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)^{\cdot}\right)^2}{R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)^3} - \frac{\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)^{\cdot\cdot}}{R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)^2}.$$

SA-6.20 Proof of (6.21). Putting $a := \left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^{-2}$, $b := R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)^{-1}$, use

$$\begin{aligned}
|a| &\leq \frac{1}{(R+\varepsilon)^2}, \quad |b| \leq \frac{1}{R}, \\
|\dot{a}| &\leq \frac{2C_{11}}{(R+\varepsilon)^3}, \quad \left|\dot{b}\right| \leq \frac{C_{11}}{R^2}, \\
|\ddot{a}| &\leq C_{21}, \quad \left|\ddot{b}\right| \leq C_{22},
\end{aligned}$$

and

$$(ab)^{\cdot\cdot} = \ddot{a}b + 2\dot{a}\dot{b} + a\ddot{b}.$$

SA-6.21 Proof of (6.22). Putting

$$\begin{aligned}
a &:= \nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t)\right), \\
b &:= \nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t)\right), \\
c &:= \left(R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^{-1},
\end{aligned}$$

we have

$$\begin{aligned}
&|a| \leq M_2, \quad |\dot{a}| \leq pM_3D_1, \quad |\ddot{a}| \leq p^2M_4D_1^2 + pM_3D_2, \\
&|b| \leq M_1, \quad \left|\dot{b}\right| \leq pM_2D_1, \quad \left|\ddot{b}\right| \leq p^2M_3D_1^2 + pM_2D_2, \\
&|c| \leq \frac{1}{R+\varepsilon}, \quad |\dot{c}| \leq \frac{C_{11}}{(R+\varepsilon)^2}, \quad |\ddot{c}| \leq \frac{2C_{11}^2}{(R+\varepsilon)^3} + \frac{C_{20}}{(R+\varepsilon)^2}.
\end{aligned}$$

(6.22) follows.

The proof of Lemma SA-6.15 is concluded. ☐

Lemma SA-6.22. *Suppose Assumption SA-2.2 holds. Then the third derivative of $t \mapsto \tilde{\theta}_j(t)$ is uniformly over j and $t \in [0,T]$ bounded in absolute value by some positive constant, say D_3.*

Proof. By (6.5), (6.13) and (6.22)

$$\begin{aligned}
&\left|\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\sum_{l=k}^{n-1}\frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon}\right| \leq (n-k)C_9, \\
&\left|\left(\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\sum_{l=k}^{n-1}\frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon}\right)^{\cdot}\right| \leq (n-k)C_{13}, \\
&\left|\left(\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\sum_{l=k}^{n-1}\frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon}\right)^{\cdot\cdot}\right| \leq (n-k)C_{24}.
\end{aligned}$$

From the definition of $t \mapsto P_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)$, it means that its derivatives up to order two are bounded. Similarly, the same is true for $t \mapsto \bar{P}_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)$.

It follows from (6.19) and its proof that the derivatives up to order two of

$$t \mapsto \left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^{-2} R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)^{-1}$$

are also bounded.

These considerations give the boundedness of the second derivative of the term

$$t \mapsto \frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\left(2P_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\bar{P}_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)}{2\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^2 R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)}$$

in (2.2). The boundedness of the second derivatives of the other two terms is shown analogously. By (2.2) and since $h \leq T$, this means

$$\sup_j \sup_{t\in[0,T]} \left|\dddot{\tilde{\theta}}_j(t)\right| \leq D_3$$

for some positive constant D_3. $\square$

7 Proof of Theorem SA-2.3

Lemma SA-7.1. *Suppose Assumption SA-2.2 holds. Then for all $n \in \{0,1,\ldots,\lfloor T/h\rfloor\}$, $k \in \{0,1,\ldots,n-1\}$ we have*

$$\left|\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_k)\right)-\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_n)\right)\right| \leq C_7(n-k)h \tag{7.1}$$

Proof. (7.1) follows from the mean value theorem applied $n-k$ times. $\square$

Lemma SA-7.2. *In the setting of Lemma SA-7.1, for any $l \in \{k,k+1,\ldots,n-1\}$ we have*

$$\begin{aligned}&\left|\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_l)\right)-\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_{l+1})\right)-h\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t_n)\right)\frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t_n)\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)+\varepsilon}\right|\\&\quad\leq\left(C_{19}/2+C_8+(n-l-1)C_{13}\right)h^2.\end{aligned}$$

Proof. By the Taylor expansion of $t \mapsto \nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)$ on the segment $[t_l, t_{l+1}]$ at t_{l+1} on the left

$$\left|\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_l)\right)-\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_{l+1})\right)+h\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t_{l+1})\right)\dot{\tilde{\boldsymbol{\theta}}}_i\left(t_{l+1}^-\right)\right| \leq \frac{C_{19}}{2}h^2.$$

Combining this with (6.4) gives

$$\begin{aligned}&\left|\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_l)\right)-\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_{l+1})\right)-h\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t_{l+1})\right)\frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t_{l+1})\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t_{l+1})\right)+\varepsilon}\right|\\&\quad\leq\left(C_{19}/2+C_8\right)h^2.\end{aligned}\tag{7.2}$$

Now applying the mean-value theorem $n-l-1$ times, we have

$$\left|\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t_{l+1})\right)\frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t_{l+1})\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t_{l+1})\right)+\varepsilon}-\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t_{l+2})\right)\frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t_{l+2})\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t_{l+2})\right)+\varepsilon}\right| \leq C_{13}h,$$

$$\ldots$$

$$\left|\sum_{i=1}^{p}\nabla_{ij}E_l\left(\tilde{\boldsymbol{\theta}}(t_{n-1})\right)\frac{\nabla_i E_k\left(\tilde{\boldsymbol{\theta}}(t_{n-1})\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t_{n-1})\right)+\varepsilon}-\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t_n)\right)\frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t_n)\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)+\varepsilon}\right| \leq C_{13}h,$$

and in particular

$$\left|\sum_{i=1}^{p} \nabla_{ij} E_k\left(\tilde{\boldsymbol{\theta}}(t_{l+1})\right) \frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t_{l+1})\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t_{l+1})\right)+\varepsilon} - \sum_{i=1}^{p} \nabla_{ij} E_k\left(\tilde{\boldsymbol{\theta}}(t_n)\right) \frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t_n)\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)+\varepsilon}\right|$$
$$\leq (n-l-1) C_{13} h.$$

Combining this with (7.2), we conclude the proof of Lemma SA-7.2. $\square$

Lemma SA-7.3. *In the setting of Lemma SA-7.1,*

$$\left|\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_k)\right) - \nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_n)\right) - h \sum_{i=1}^{p} \nabla_{ij} E_k\left(\tilde{\boldsymbol{\theta}}(t_n)\right) \sum_{l=k}^{n-1} \frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t_n)\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)+\varepsilon}\right|$$
$$\leq \left((n-k)(C_{19}/2 + C_8) + \frac{(n-k)(n-k-1)}{2} C_{13}\right) h^2.$$

Proof. Fix $n \in \mathbb{Z}_{\geq 0}$.

Note that

$$\begin{aligned}
&\left|\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_k)\right) - \nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_n)\right) - h \sum_{i=1}^{p} \nabla_{ij} E_k\left(\tilde{\boldsymbol{\theta}}(t_n)\right) \sum_{l=k}^{n-1} \frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t_n)\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)+\varepsilon}\right| \\
&= \left|\sum_{l=k}^{n-1} \left\{\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_l)\right) - \nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_{l+1})\right) - h \sum_{i=1}^{p} \nabla_{ij} E_k\left(\tilde{\boldsymbol{\theta}}(t_n)\right) \frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t_n)\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)+\varepsilon}\right\}\right| \\
&\leq \sum_{l=k}^{n-1} \left|\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_l)\right) - \nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_{l+1})\right) - h \sum_{i=1}^{p} \nabla_{ij} E_k\left(\tilde{\boldsymbol{\theta}}(t_n)\right) \frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t_n)\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)+\varepsilon}\right| \\
&\overset{\text{(a)}}{\leq} \sum_{l=k}^{n-1} \left(C_{19}/2 + C_8 + (n-l-1) C_{13}\right) h^2 = \left((n-k)(C_{19}/2 + C_8) + \frac{(n-k)(n-k-1)}{2} C_{13}\right) h^2,
\end{aligned}$$

where (a) is by Lemma SA-7.2. $\square$

Lemma SA-7.4. *Suppose Assumption SA-2.2 holds. Then for all $n \in \{0, 1, \ldots, \lfloor T/h \rfloor\}$*

$$\left|\sum_{k=0}^{n} \rho^{n-k}(1-\rho)\left(\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_k)\right)\right)^2 - R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)^2\right| \leq C_{25} h \tag{7.3}$$

and

$$\left|\sum_{k=0}^{n} \rho^{n-k}(1-\rho)\left(\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_k)\right)\right)^2 - R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)^2 - 2h P_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)\right| \leq C_{26} h^2 \tag{7.4}$$

with C_{25} and C_{26} defined as follows:

$$\begin{aligned}
C_{25}(\rho) &:= 2 M_1 C_7 \frac{\rho}{1-\rho}, \\
C_{26}(\rho) &:= M_1 \left|C_{19} + 2C_8 - C_{13}\right| \frac{\rho}{1-\rho} \\
&\quad + \left(M_1 C_{13} + \left|C_{19} + 2C_8 - C_{13}\right| C_9 + \frac{\left(C_{19} + 2C_8 - C_{13}\right)^2}{4}\right) \frac{\rho(1+\rho)}{(1-\rho)^2} \\
&\quad + \left(C_{13} C_9 + \frac{C_{13}}{2}\left|C_{19} + 2C_8 - C_{13}\right|\right) \frac{\rho\left(1 + 4\rho + \rho^2\right)}{(1-\rho)^3} + \frac{C_{13}^2}{4} \cdot \frac{\rho\left(1 + 11\rho + 11\rho^2 + \rho^3\right)}{(1-\rho)^4}.
\end{aligned}$$

Proof. Note that

$$\begin{aligned}
&\left|\left(\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_k)\right)\right)^2-\left(\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_n)\right)\right)^2\right| \\
&\quad\le\left|\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_k)\right)-\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_n)\right)\right|\cdot\left|\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_k)\right)+\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_n)\right)\right| \\
&\quad\overset{\text{(a)}}{\le} C_7(n-k)h\cdot 2M_1,
\end{aligned}$$

where (a) is by (7.1). Using the triangle inequality, we can conclude

$$\begin{aligned}
&\left|\sum_{k=0}^{n}\rho^{n-k}(1-\rho)\left(\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_k)\right)\right)^2-R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)^2\right| \\
&\le 2M_1C_7h(1-\rho)\sum_{k=0}^{n}(n-k)\rho^{n-k}=2M_1C_7h(1-\rho)\sum_{k=0}^{n}k\rho^k=2M_1C_7\frac{\rho}{1-\rho}h.
\end{aligned}$$

(7.3) is proven.

We continue by showing

$$\begin{aligned}
&\Bigg|\left(\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_k)\right)\right)^2-\left(\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_n)\right)\right)^2 \\
&\qquad -2\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_n)\right)h\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t_n)\right)\sum_{l=k}^{n-1}\frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t_n)\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)+\varepsilon}\Bigg| \\
&\le 2M_1\left((n-k)(C_{19}/2+C_8)+\frac{(n-k)(n-k-1)}{2}C_{13}\right)h^2 \\
&\quad +2(n-k)C_9\left((n-k)(C_{19}/2+C_8)+\frac{(n-k)(n-k-1)}{2}C_{13}\right)h^3 \\
&\quad +\left((n-k)(C_{19}/2+C_8)+\frac{(n-k)(n-k-1)}{2}C_{13}\right)^2h^4.
\end{aligned}\tag{7.5}$$

To prove this, use

$$\left|a^2-b^2-2bKh\right|\le 2|b|\cdot|a-b-Kh|+2|K|\cdot h\cdot|a-b-Kh|+(a-b-Kh)^2$$

with

$$a:=\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_k)\right),\quad b:=\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_n)\right),\quad K:=\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t_n)\right)\sum_{l=k}^{n-1}\frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t_n)\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)+\varepsilon},$$

and bounding

$$\begin{aligned}
&|a-b-Kh|\overset{\text{(a)}}{\le}\left((n-k)(C_{19}/2+C_8)+\frac{(n-k)(n-k-1)}{2}C_{13}\right)h^2, \\
&|b|\le M_1,\quad |K|\le(n-k)C_9,
\end{aligned}$$

where (a) is by Lemma SA-7.3. (7.5) is proven.

We turn to the proof of (7.4). By (7.5) and the triangle inequality

$$\left|\sum_{k=0}^{n}\rho^{n-k}(1-\rho)\left(\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_k)\right)\right)^2-R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)^2-2hP_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)\right|$$

$$
\begin{aligned}
&\leq (1-\rho)\sum_{k=0}^{n}\rho^{n-k}\left(\mathrm{Poly}_1(n-k)h^2+\mathrm{Poly}_2(n-k)h^3+\mathrm{Poly}_3(n-k)h^4\right)\\
&=(1-\rho)\sum_{k=0}^{n}\rho^{k}\left(\mathrm{Poly}_1(k)h^2+\mathrm{Poly}_2(k)h^3+\mathrm{Poly}_3(k)h^4\right),
\end{aligned}
$$

where

$$
\begin{aligned}
\mathrm{Poly}_1(k) &:= 2M_1\left(k(C_{19}/2+C_8)+\frac{k(k-1)}{2}C_{13}\right)=M_1C_{13}k^2+M_1(C_{19}+2C_8-C_{13})k,\\
\mathrm{Poly}_2(k) &:= 2kC_9\left(k(C_{19}/2+C_8)+\frac{k(k-1)}{2}C_{13}\right)=C_{13}C_9k^3+\left(C_{19}+2C_8-C_{13}\right)C_9k^2,\\
\mathrm{Poly}_3(k) &:= \left(k(C_{19}/2+C_8)+\frac{k(k-1)}{2}C_{13}\right)^2\\
&=\frac{C_{13}^2}{4}k^4+\frac{C_{13}}{2}\left(C_{19}+2C_8-C_{13}\right)k^3+\frac{1}{4}\left(C_{19}+2C_8-C_{13}\right)^2k^2.
\end{aligned}
$$

It is left to combine this with

$$
\begin{aligned}
\sum_{k=0}^{n}k\rho^k &\leq \sum_{k=0}^{\infty}k\rho^k=\frac{\rho}{(1-\rho)^2},\\
\sum_{k=0}^{n}k^2\rho^k &\leq \sum_{k=0}^{\infty}k^2\rho^k=\frac{\rho(1+\rho)}{(1-\rho)^3},\\
\sum_{k=0}^{n}k^3\rho^k &\leq \sum_{k=0}^{\infty}k^3\rho^k=\frac{\rho\left(1+4\rho+\rho^2\right)}{(1-\rho)^4},\\
\sum_{k=0}^{n}k^4\rho^k &\leq \sum_{k=0}^{\infty}k^4\rho^k=\frac{\rho\left(1+11\rho+11\rho^2+\rho^3\right)}{(1-\rho)^5}.
\end{aligned}
$$

This gives

$$
\begin{aligned}
&\left|\sum_{k=0}^{n}\rho^{n-k}(1-\rho)\left(\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_k)\right)\right)^2-R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)^2-2hP_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)\right|\\
&\leq\left(M_1C_{13}\frac{\rho(1+\rho)}{(1-\rho)^2}+M_1\left|C_{19}+2C_8-C_{13}\right|\frac{\rho}{1-\rho}\right)h^2\\
&\quad+\left(C_{13}C_9\frac{\rho\left(1+4\rho+\rho^2\right)}{(1-\rho)^3}+\left|C_{19}+2C_8-C_{13}\right|C_9\frac{\rho(1+\rho)}{(1-\rho)^2}\right)h^3\\
&\quad+\left(\frac{C_{13}^2}{4}\cdot\frac{\rho\left(1+11\rho+11\rho^2+\rho^3\right)}{(1-\rho)^4}+\frac{C_{13}}{2}\left|C_{19}+2C_8-C_{13}\right|\frac{\rho\left(1+4\rho+\rho^2\right)}{(1-\rho)^3}\right.\\
&\qquad\left.+\frac{1}{4}\left(C_{19}+2C_8-C_{13}\right)^2\frac{\rho(1+\rho)}{(1-\rho)^2}\right)h^4\\
&\overset{\text{(a)}}{\leq}\left[M_1\left|C_{19}+2C_8-C_{13}\right|\frac{\rho}{1-\rho}\right.\\
&\quad+\left(M_1C_{13}+\left|C_{19}+2C_8-C_{13}\right|C_9+\frac{\left(C_{19}+2C_8-C_{13}\right)^2}{4}\right)\frac{\rho(1+\rho)}{(1-\rho)^2}\\
&\quad+\left(C_{13}C_9+\frac{C_{13}}{2}\left|C_{19}+2C_8-C_{13}\right|\right)\frac{\rho\left(1+4\rho+\rho^2\right)}{(1-\rho)^3}\\
&\quad\left.+\frac{C_{13}^2}{4}\cdot\frac{\rho\left(1+11\rho+11\rho^2+\rho^3\right)}{(1-\rho)^4}\right]h^2,
\end{aligned}
$$

where in (a) we used that $h<1$. (7.4) is proven. $\square$

Lemma SA-7.5. *Suppose Assumption SA-2.2 holds. Then*

$$\left| \left(\sqrt{\sum_{k=0}^{n} \rho^{n-k}(1-\rho) \left(\nabla_j E_k \left(\tilde{\boldsymbol{\theta}}(t_k) \right) \right)^2} + \varepsilon \right)^{-1} - \left(R_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t_n) \right) + \varepsilon \right)^{-1} \right.$$

$$\left. + h \frac{P_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t_n) \right)}{\left(R_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t_n) \right) + \varepsilon \right)^2 R_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t_n) \right)} \right| \leq \frac{C_{25}(\rho)^2 + R^2 C_{26}(\rho)}{2R^3(R+\varepsilon)^2} h^2.$$

Proof. Note that if $a \geq R^2$, $b \geq R^2$, we have

$$\begin{aligned}
&\left| \frac{1}{\sqrt{a}+\varepsilon} - \frac{1}{\sqrt{b}+\varepsilon} + \frac{a-b}{2\left(\sqrt{b}+\varepsilon\right)^2 \sqrt{b}} \right| \\
&\quad = \frac{(a-b)^2}{2\sqrt{b}\left(\sqrt{b}+\varepsilon\right)\left(\sqrt{a}+\varepsilon\right)\left(\sqrt{a}+\sqrt{b}\right)} \underbrace{\left\{ \frac{1}{\sqrt{b}+\varepsilon} + \frac{1}{\sqrt{a}+\sqrt{b}} \right\}}_{\leq 2/R} \\
&\quad \leq \frac{(a-b)^2}{2R^3(R+\varepsilon)^2}.
\end{aligned}$$

By the triangle inequality,

$$\begin{aligned}
&\left| \frac{1}{\sqrt{a}+\varepsilon} - \frac{1}{\sqrt{b}+\varepsilon} + \frac{c}{2\left(\sqrt{b}+\varepsilon\right)^2 \sqrt{b}} \right| \leq \frac{(a-b)^2}{2R^3(R+\varepsilon)^2} + \frac{|a-b-c|}{2\left(\sqrt{b}+\varepsilon\right)^2\sqrt{b}} \\
&\quad \leq \frac{(a-b)^2}{2R^3(R+\varepsilon)^2} + \frac{|a-b-c|}{2R\left(R+\varepsilon\right)^2}
\end{aligned}$$

Apply this with

$$\begin{aligned}
a &:= \sum_{k=0}^{n} \rho^{n-k}(1-\rho) \left(\nabla_j E_k \left(\tilde{\boldsymbol{\theta}}(t_k) \right) \right)^2, \\
b &:= R_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t_n) \right)^2, \\
c &:= 2h P_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t_n) \right)
\end{aligned}$$

and use bounds

$$|a-b| \leq 2M_1 C_7 \frac{\rho}{1-\rho} h, \quad |a-b-c| \leq C_{26}(\rho) h^2$$

by Lemma SA-7.4. □

SA-7.6. We are finally ready to prove Theorem SA-2.3.

Proof of Theorem SA-2.3. By (6.9) and (6.10), the first derivative of the function

$$t \mapsto \left(\frac{\nabla_j E_n \left(\tilde{\boldsymbol{\theta}}(t) \right) \left(2P_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \bar{P}_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t) \right) \right)}{2\left(R_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon \right)^2 R_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t) \right)} - \frac{\sum_{i=1}^{p} \nabla_{ij} E_n \left(\tilde{\boldsymbol{\theta}}(t) \right) \frac{\nabla_i E_n \left(\tilde{\boldsymbol{\theta}}(t) \right)}{R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon}}{2\left(R_j^{(n)}(\tilde{\boldsymbol{\theta}}(t)) + \varepsilon \right)} \right)$$

is bounded in absolute value by a positive constant $C_{27} = C_{17} + C_{18}$. By (2.2), this means

$$\left| \ddot{\tilde{\theta}}_j(t) + \frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{\nabla_j E_n \left(\tilde{\boldsymbol{\theta}}(t) \right)}{R_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon} \right) \right| \leq C_{27} h.$$

Combining this with

$$\left| \tilde{\theta}_j(t_{n+1}) - \tilde{\theta}_j(t_n) - \dot{\tilde{\theta}}_j \left(t_n^+ \right) h - \frac{\ddot{\tilde{\theta}}_j \left(t_n^+ \right)}{2} h^2 \right| \leq \frac{D_3}{6}$$

by Taylor expansion, we get

$$\begin{aligned} &\left| \tilde{\theta}_j(t_{n+1}) - \tilde{\theta}_j(t_n) - \dot{\tilde{\theta}}_j \left(t_n^+ \right) h + \frac{h^2}{2} \cdot \frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{\nabla_j E_n \left(\tilde{\boldsymbol{\theta}}(t) \right)}{R_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon} \right) \Bigg|_{t=t_n^+} \right| \\ &\quad \leq \left(\frac{D_3}{6} + \frac{C_{27}}{2} \right) h^3. \end{aligned} \tag{7.6}$$

Using

$$\left| \dot{\tilde{\theta}}_j(t_n) + \frac{\nabla_j E_n \left(\tilde{\boldsymbol{\theta}}(t_n) \right)}{R_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t_n) \right) + \varepsilon} \right| \leq C_{28} h$$

with C_{28} defined as

$$C_{28} := \frac{M_1 \left(2C_5 + C_6 \right)}{2(R + \varepsilon)^2 R} + \frac{p M_1 M_2}{2(R + \varepsilon)^2}$$

by (2.2), and calculating the derivative, it is easy to show

$$\left| \frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{\nabla_j E_n \left(\tilde{\boldsymbol{\theta}}(t) \right)}{R_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon} \right) \Bigg|_{t=t_n^+} - \mathrm{FrDer} \right| \leq C_{29} h \tag{7.7}$$

for a positive constant C_{29}, where

$$\begin{aligned} \mathrm{FrDer} &:= \frac{\mathrm{FrDerNum}}{\left(R_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t_n) \right) + \varepsilon \right)^2 R_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t_n) \right)} \\ \mathrm{FrDerNum} &:= \nabla_j E_n \left(\tilde{\boldsymbol{\theta}}(t_n) \right) \bar{P}_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t_n) \right) \\ &\quad - \left(R_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t_n) \right) + \varepsilon \right) R_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t_n) \right) \sum_{i=1}^{p} \nabla_{ij} E_n \left(\tilde{\boldsymbol{\theta}}(t_n) \right) \frac{\nabla_i E_n \left(\tilde{\boldsymbol{\theta}}(t_n) \right)}{R_i^{(n)} \left(\tilde{\boldsymbol{\theta}}(t_n) \right) + \varepsilon}, \\ C_{29} &:= \left\{ \frac{p M_2}{R + \varepsilon} + \frac{M_1^2 M_2 p}{(R + \varepsilon)^2 R} \right\} C_{28}. \end{aligned}$$

From (7.6) and (7.7), by the triangle inequality

$$\left| \tilde{\theta}_j(t_{n+1}) - \tilde{\theta}_j(t_n) - \dot{\tilde{\theta}}_j \left(t_n^+ \right) h + \frac{h^2}{2} \mathrm{FrDer} \right| \leq \left(\frac{D_3}{6} + \frac{C_{27} + C_{29}}{2} \right) h^3,$$

which, using (2.2), is rewritten as

$$\left| \tilde{\theta}_j(t_{n+1}) - \tilde{\theta}_j(t_n) + h \frac{\nabla_j E_n \left(\tilde{\boldsymbol{\theta}}(t_n) \right)}{R_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t_n) \right) + \varepsilon} - h^2 \frac{\nabla_j E_n \left(\tilde{\boldsymbol{\theta}}(t_n) \right) P_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t_n) \right)}{\left(R_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t_n) \right) + \varepsilon \right)^2 R_j^{(n)} \left(\tilde{\boldsymbol{\theta}}(t_n) \right)} \right|$$

$$\leq \left(\frac{D_3}{6} + \frac{C_{27} + C_{29}}{2} \right) h^3.$$

It is left to combine this with Lemma SA-7.5, giving the assertion of the theorem with

$$C_1 = \frac{D_3}{6} + \frac{C_{27} + C_{29}}{2} + M_1 \frac{C_{25}^2 + R^2 C_{26}}{2R^3 (R+\varepsilon)^2}. \qquad \square$$

8 Numerical experiments

SA-8.1 Models. We use small modifications of default Keras Resnet-50 and Resnet-101 architectures[1] for training on CIFAR-10 and CIFAR-100 (since image sizes are not the same as Imagenet), after verifying their correctness. The first convolution layer `conv1` has 3×3 kernel, stride 1 and "same" padding. Then comes batch normalization, and relu. Max pooling is removed, and otherwise `conv2_x` to `conv5_x` are as described in [2], see Table 1 there (downsampling is performed by the first convolution of each bottleneck block, same as in this original paper, not the middle one as in version 1.5[2]; all convolution layers have learned biases). After `conv5` there is global average pooling, 10 or 100-way fully connected layer (for CIFAR-10 and CIFAR-100 respectively), and softmax.

SA-8.2 Data augmentation. We subtract the per-pixel mean and divide by standard deviation, and we use the data augmentation scheme from [3], following [2], section 4.2. We take inspiration and some code snippets from [4] (though we do not use their models). During each pass over the training dataset, each 32×32 initial image is padded evenly with zeros so that it becomes 36×36, then random crop is applied so that the picture becomes 32×32 again, and finally random (probability 0.5) horizontal (left to right) flip is used.

References

[1] Avrajit Ghosh, He Lyu, Xitong Zhang, and Rongrong Wang. "Implicit regularization in Heavy-ball momentum accelerated stochastic gradient descent". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=ZzdBhtEH9yB.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[3] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. "Deeply-supervised nets". In: *Artificial intelligence and statistics*. Pmlr. 2015, pp. 562–570.

[4] Chia-Hung Yuan. *Training CIFAR-10 with TensorFlow2(TF2)*. https://github.com/lionelmessi6410/tensorflow2-cifar. 2021.

[1] https://github.com/keras-team/keras/blob/v2.13.1/keras/applications/resnet.py

[2] https://catalog.ngc.nvidia.com/orgs/nvidia/resources/resnet_50_v1_5_for_pytorch