On the Implicit Bias of Adam

Matias D. Cattaneo [*1] **Jason M. Klusowski** [*1] **Boris Shigida** [*1]

Abstract

In previous literature, backward error analysis was used to find ordinary differential equations (ODEs) approximating the gradient descent trajectory. It was found that finite step sizes implicitly regularize solutions because terms appearing in the ODEs penalize the two-norm of the loss gradients. We prove that the existence of similar implicit regularization in RMSProp and Adam depends on their hyperparameters and the training stage, but with a different "norm" involved: the corresponding ODE terms either penalize the (perturbed) one-norm of the loss gradients or, conversely, impede its reduction (the latter case being typical). We also conduct numerical experiments and discuss how the proven facts can influence generalization.

1. Introduction

Gradient descent (GD) can be seen as a numerical method solving the ordinary differential equation (ODE) $\dot{\boldsymbol{\theta}} = -\nabla E(\boldsymbol{\theta})$, where $E(\cdot)$ is the loss function and $\nabla E(\boldsymbol{\theta})$ is its gradient. Starting at $\boldsymbol{\theta}^{(0)}$, it creates a sequence of guesses $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$, which lie close to the solution trajectory $\boldsymbol{\theta}(t)$ governed by the aforementioned ODE. Since the step size h is finite, one could search for a modified differential equation $\dot{\tilde{\boldsymbol{\theta}}} = -\nabla \widetilde{E}(\tilde{\boldsymbol{\theta}})$ such that $\boldsymbol{\theta}^{(n)} - \tilde{\boldsymbol{\theta}}(nh)$ is exactly zero, or at least closer to zero than $\boldsymbol{\theta}^{(n)} - \boldsymbol{\theta}(nh)$, that is, all the guesses of the descent lie exactly on the new solution curve or closer compared to the original curve. This approach to analysing properties of a numerical method is sometimes called backward error analysis in the numerical integration literature (see Chapter IX in Ernst Hairer & Wanner (2006) and references therein).

Barrett & Dherin (2021) used this idea for full-batch GD and found that the modified loss function $\widetilde{E}(\tilde{\boldsymbol{\theta}}) = E(\tilde{\boldsymbol{\theta}}) + (h/4)\|\nabla E(\tilde{\boldsymbol{\theta}})\|^2$ makes the trajectory of the solution to $\dot{\tilde{\boldsymbol{\theta}}} = -\nabla \widetilde{E}(\tilde{\boldsymbol{\theta}})$ approximate the sequence $\{\boldsymbol{\theta}^{(n)}\}_{n=0}^{\infty}$ one order of h better than the original ODE, where $\|\cdot\|$ is the Euclidean norm. In related work, Miyagawa (2022) obtained the correction term for full-batch GD up to any chosen order, also studying the global error (uniform in the iteration number) as opposed to the local (one-step) error.

The analysis was later extended to mini-batch GD in Smith et al. (2021). Assume that the training set is split into batches of size B and there are m batches per epoch (so the training set size is mB). The cost function is rewritten as $E(\boldsymbol{\theta}) = (1/m)\sum_{k=0}^{m-1} \hat{E}_k(\boldsymbol{\theta})$ with mini-batch costs denoted $\hat{E}_k(\boldsymbol{\theta}) = (1/B)\sum_{j=kB+1}^{kB+B} E_j(\boldsymbol{\theta})$. It was obtained in that work that after one epoch, the mean iterate of the algorithm, averaged over all possible shuffles of the batch indices, is close to the solution to $\dot{\boldsymbol{\theta}} = -\nabla \widetilde{E}_{SGD}(\boldsymbol{\theta})$, where the modified loss is given by $\widetilde{E}_{SGD}(\boldsymbol{\theta}) = E(\boldsymbol{\theta}) + h/(4m) \cdot \sum_{k=0}^{m-1} \left\|\nabla \hat{E}(\boldsymbol{\theta})\right\|^2$.

Modified equations have also been derived for GD with heavy-ball momentum $\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)} - h\nabla E(\boldsymbol{\theta}^{(n)}) + \beta(\boldsymbol{\theta}^{(n)} - \boldsymbol{\theta}^{(n-1)})$, where β is the momentum parameter. In the full-batch setting, it turns out that for n large enough it is close to the continuous trajectory solving

$$\dot{\boldsymbol{\theta}} = -\frac{\nabla E(\boldsymbol{\theta})}{1-\beta} - \underbrace{h\frac{1+\beta}{(1-\beta)^3}\frac{\nabla\|\nabla E(\boldsymbol{\theta})\|^2}{4}}_{\text{implicit regularization}}. \tag{1}$$

Versions of this general result were proven in Farazmand (2020), Kovachki & Stuart (2021), Ghosh et al. (2023) under different assumptions. The focus of the latter work is the closest to ours since they interpret the correction term as implicit regularization. Their main theorem also provides the analysis for the general mini-batch case.

In another recent work, Zhao et al. (2022) introduce a regularization term $\lambda \cdot \|\nabla E(\boldsymbol{\theta})\|$ to the loss function as a way to ensure finding flatter minima, improving generalization. The only difference between their term and the first-order correction coming from backward error analysis (up to a coefficient) is that the norm is not squared and regularization is applied on a per-batch basis.

The application of backward error analysis for approximating the discrete dynamics of adaptive algorithms such as

[*] Equal contribution [1] Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ, USA. Correspondence to: Boris Shigida <bs1624@princeton.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

RMSProp (Tieleman et al., 2012) and Adam (Kingma & Ba, 2015) is currently missing in the literature. Barrett & Dherin (2021) note that "it would be interesting to use backward error analysis to calculate the modified loss and implicit regularization for other widely used optimizers such as momentum, Adam and RMSprop". Smith et al. (2021) reiterate that they "anticipate that backward error analysis could also be used to clarify the role of finite learning rates in adaptive optimizers like Adam". Ghosh et al. (2023) agree that "RMSProp ... and Adam ..., albeit being powerful alternatives to SGD with faster convergence rates, are far from well-understood in the aspect of implicit regularization". In a similar context, in Appendix G to Miyagawa (2022) it is mentioned that "its [Adam's] counter term and discretization error are open questions".

This work fills the gap by conducting backward error analysis for (mini-batch, and full-batch as a special case) Adam and RMSProp. Our main contributions are listed below.

- In Theorem 3.1, we provide a global second-order in h continuous piecewise ODE approximation to Adam in the general mini-batch setting. (A similar result for RMSProp is moved to Appendix C.) For the full-batch special case, it was shown in prior work Ma et al. (2022) that the continuous-time limit of both these algorithms is a (perturbed by the numerical stability parameter ε) signGD flow $\dot{\boldsymbol{\theta}} = -\nabla E(\boldsymbol{\theta})/(|\nabla E(\boldsymbol{\theta})| + \varepsilon)$ component-wise; we make this more precise by finding a linear in h correction term on the right.

- We analyze the full-batch case in the context of regularization (see the summary in Section 2). In contrast to the case of GD, where the two-norm of the loss gradient is implicitly penalized, Adam typically *anti*-penalizes the perturbed one-norm of the loss gradient $\|\mathbf{v}\|_{1,\varepsilon} = \sum_{i=1}^{p} \sqrt{v_i^2 + \varepsilon}$ (i. e., penalizes the negative norm), as specified in (5). Thus, the implicit bias of Adam that we identify serves as *anti-regularization* (except for the unusual case $\beta \geq \rho$, large ε or very late at training).

- We provide numerical evidence consistent with our theoretical results by training various vision models on CIFAR-10 using full-batch Adam. In particular, we observe that the stronger the implicit anti-regularization effect predicted by our theory, the worse the generalization. This pattern holds across different architectures: ResNets, simple convolutional neural networks (CNNs) and Vision Transformers. Thus, we propose a novel possible explanation for often-reported poor generalization of adaptive gradient algorithms. The code used for training the models is available at `https://github.com/borshigida/implicit-bias-of-adam`.

1.1. Related Work

Backward error analysis of first-order methods. We outlined the history of finding ODEs approximating different algorithms above in the introduction. Recently, there have been other applications of backward error analysis related to machine learning. Kunin et al. (2020) show that the approximating continuous-time trajectories satisfy conservation laws that are broken in discrete time. França et al. (2021) use backward error analysis while studying how to discretize continuous-time dynamical systems preserving stability and convergence rates. Rosca et al. (2021) find continuous-time approximations of discrete two-player differential games.

Approximating gradient methods by differential equation trajectories. Under the assumption that the hyperparameters β, ρ of the Adam algorithm (see Definition 1.1) tend to 1 at a certain rate as $h \to 0$, a first-order continuous ODE approximation to this algorithm was derived in Barakat & Bianchi (2021). On the other hand, if β, ρ are kept fixed, Ma et al. (2022) prove that the trajectories of Adam and RMSProp are close to signGD dynamics, and investigate different training regimes of these algorithms empirically. SGD is approximated by stochastic differential equations and novel adaptive parameter adjustment policies are devised in Li et al. (2017). Malladi et al. (2022) derive stochastic differential equations that are order-1 weak approximations of RMSProp and Adam. We go in a different direction: instead of clarifying the previously obtained continuous ODE approximations by taking gradient noise into account, we take a deterministic approach but go one order of h further. In particular, we keep β, ρ fixed (thus generalizing the analysis for SGD with momentum), whereas Malladi et al. (2022) take $\beta, \rho \to 1$.

Connection with signGD. The connection of adaptive gradient methods with sign(S)GD is extensively discussed in Bernstein et al. (2018). Balles et al. (2020) study a version of signGD with an update proportional to $-\|\nabla E(\boldsymbol{\theta})\|_1 \operatorname{sign} \nabla E(\boldsymbol{\theta})$ as a special case of steepest descent, and discuss when sign-based methods are preferable to GD.

Implicit bias of first-order methods. Soudry et al. (2018) prove that GD trained to classify linearly separable data with logistic loss converges to the direction of the max-margin vector (the solution to the hard margin SVM). This result has been extended to different loss functions in Nacson et al. (2019b), to SGD in Nacson et al. (2019c), AdaGrad in Qian & Qian (2019), (S)GD with momentum, deterministic Adam and stochastic RMSProp in Wang et al. (2022), more generic optimization methods in Gunasekar et al. (2018a), to the nonseparable case in Ji & Telgarsky (2018b), Ji & Telgarsky (2019). This line of research has been generalized to studying implicit biases of linear networks (Ji & Telgarsky, 2018a; Gunasekar et al., 2018b), homogeneous neural networks (Ji & Telgarsky, 2020; Nacson et al., 2019a; Lyu & Li, 2019). Woodworth et al. (2020) study the gradient flow of a diagonal linear network with squared loss and show that large initializations lead to minimum two-norm solutions while small initializations lead to minimum one-norm solutions. Even et al. (2023) extend this work to the case of non-zero step sizes and mini-batch training. Wang et al. (2021) prove that Adam and RMSProp maximize the margin of homogeneous neural networks. Our perspective on the implicit bias is different since we are considering a generic loss function without any assumptions on the network architecture. Beneventano (2023) proves that in expectation over batch sampling the trajectory of SGD without replacement differs from that of SGD with replacement by an additional step on a regularizer. As opposed to the work on backward error analysis for SGD discussed above, they do not assume the largest eigenvalue of the hessian to be bounded.

Generalization of adaptive methods. Cohen et al. (2022) investigate the edge-of-stability regime of adaptive gradient algorithms and the effect of sharpness (the largest eigenvalue of the hessian) on generalization. Granziol (2020); Chen et al. (2021) observe that adaptive methods find sharper minima than SGD and Zhou et al. (2020); Xie et al. (2022) argue theoretically that it is the case. Jiang et al. (2022) introduce a statistic that measures the uniformity of the hessian diagonal and argue that adaptive gradient algorithms are biased towards making this statistic smaller. Keskar & Socher (2017) propose to improve generalization of adaptive methods by switching to SGD in the middle of training.

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 $E(\boldsymbol{\theta})$. ∇E means the gradient of E, and ∇ with indices denotes 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 norm notation without indices $\|\cdot\|$ is the two-norm of a vector, $\|\cdot\|_1$ is the one-norm and $\|\cdot\|_{1,\varepsilon}$ is the perturbed one-norm defined as $\|\mathbf{v}\|_{1,\varepsilon} = \sum_{i=1}^p \sqrt{v_i^2 + \varepsilon}$. (Of course, if $\varepsilon > 0$ the perturbed one-norm is not really a norm, but taking $\varepsilon = 0$ makes it the one-norm.) For a real number a the floor $\lfloor a \rfloor$ is the largest integer not exceeding a.

To provide the names and notations for hyperparameters, we define the algorithm below.

Definition 1.1. The *Adam* algorithm (Kingma & Ba, 2015) is an optimization algorithm with numerical stability hyperparameter $\varepsilon > 0$, squared gradient momentum hyperparameter $\rho \in (0, 1)$, gradient momentum hyperparameter $\beta \in (0, 1)$, initialization $\boldsymbol{\theta}^{(0)} \in \mathbb{R}^p$, $\boldsymbol{\nu}^{(0)} = \mathbf{0} \in \mathbb{R}^p$, $\mathbf{m}^{(0)} = \mathbf{0} \in \mathbb{R}^p$ and the following update rule: for each $n \geq 0$, $j \in \{1, \dots, p\}$

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

Remark 1.2. Note that the numerical stability hyperparameter $\varepsilon > 0$, which is introduced in these algorithms to avoid division by zero, is inside the square root in our definition. This way we avoid division by zero in the derivative too: the first derivative of $x \mapsto \big(\sqrt{x + \varepsilon}\big)^{-1}$ is bounded for $x \geq 0$. This is useful for our analysis. In Theorems B.4 and D.4, the original versions of RMSProp and Adam are also tackled, though with an additional assumption which requires that no component of the gradient can come very close to zero in the region of interest. This is true only for the initial period of learning (whereas Theorem 3.1 tackles the whole period). Practitioners do not seem to make a distinction between the version with ε inside vs. outside the square root: tutorials with both versions abound on machine learning related websites. Moreover, the popular Tensorflow and Optax variants of RMSProp have ε inside the square root. Empirically we also observed that moving ε inside or outside the square root does not change the behavior of Adam or RMSProp qualitatively.

2. Implicit Bias of Full-Batch Adam: an Informal Summary

We are ready to describe our theoretical result (Theorem 3.1 below) in the full-batch special case. Assume $E(\boldsymbol{\theta})$ is the loss, whose partial derivatives up to the fourth order are bounded. Let $\{\boldsymbol{\theta}^{(n)}\}$ be iterations of Adam as defined in Definition 1.1. We find an ODE whose solution trajectory $\tilde{\boldsymbol{\theta}}(t)$ is h^2-close to $\{\boldsymbol{\theta}^{(n)}\}$, meaning that for any time horizon $T > 0$ there is a constant C such that for any step size $h \in (0, T)$ we have $\|\tilde{\boldsymbol{\theta}}(nh) - \boldsymbol{\theta}^{(n)}\| \leq Ch^2$ (for n between 0 and $\lfloor T/h \rfloor$). The ODE is written the following way (up to terms that rapidly go to zero as n grows): for the component number $j \in \{1, \dots, p\}$

$$\dot{\tilde{\theta}}_j(t) = -\frac{\nabla_j E\big(\tilde{\boldsymbol{\theta}}(t)\big) + \mathrm{correction}_j\big(\tilde{\boldsymbol{\theta}}(t)\big)}{\sqrt{\big|\nabla_j E\big(\tilde{\boldsymbol{\theta}}(t)\big)\big|^2 + \varepsilon}} \qquad (2)$$

with initial conditions $\tilde{\boldsymbol{\theta}}_j(0) = \boldsymbol{\theta}_j^{(0)}$ for all j, where the correction term is

$$\begin{aligned}
&\mathrm{correction}_j(\boldsymbol{\theta}) \\
&:= \frac{h}{2}\left\{\frac{1+\beta}{1-\beta} - \frac{1+\rho}{1-\rho} + \frac{1+\rho}{1-\rho} \cdot \frac{\varepsilon}{|\nabla_j E(\boldsymbol{\theta})|^2 + \varepsilon}\right\}
\end{aligned}$$

$$\times \nabla_j \big\|\nabla E(\boldsymbol{\theta})\big\|_{1,\varepsilon}. \tag{3}$$

Depending on hyperparameters and the training stage, the correction term can take two extreme forms listed below. The reality is in between, but typically much closer to the first case.

- If $\sqrt{\varepsilon}$ is **small** compared to all components of $\nabla E\big(\tilde{\boldsymbol{\theta}}(t)\big)$, i. e. $\min_j \big|\nabla_j E\big(\tilde{\boldsymbol{\theta}}(t)\big)\big| \gg \sqrt{\varepsilon}$, which **is usually the case during most of the training**, then we can write

  $$\text{correction}_j(\boldsymbol{\theta}) \approx \frac{h}{2}\left\{\frac{1+\beta}{1-\beta} - \frac{1+\rho}{1-\rho}\right\}\nabla_j \|\nabla E(\boldsymbol{\theta})\|_{1,\varepsilon}. \tag{4}$$

  For small ε, the perturbed one-norm is indistinguishable from the usual one-norm, and for $\beta > \rho$ it is penalized (in much the same way as the squared two-norm is implicitly penalized in the case of GD), but for the typical case $\rho > \beta$ its decrease is actually hindered by this term (so the bias is *anti*-regularization). The ODE in (2) approximately becomes

  $$\dot{\tilde{\theta}}_j(t) = -\frac{\nabla_j \widetilde{E}\big(\tilde{\boldsymbol{\theta}}(t)\big)}{\big|\nabla_j E\big(\tilde{\boldsymbol{\theta}}(t)\big)\big|}, \quad \text{with}$$
  $$\widetilde{E}(\boldsymbol{\theta}) = E(\boldsymbol{\theta}) + \underbrace{\frac{h}{2}\left\{\frac{1+\beta}{1-\beta} - \frac{1+\rho}{1-\rho}\right\}\|\nabla E(\boldsymbol{\theta})\|_1}_{\text{implicit } \textbf{anti-regularization} \text{ (if } \rho > \beta)}. \tag{5}$$

- If $\sqrt{\varepsilon}$ is **large** compared to all gradient components, i. e. $\max_j \big|\nabla_j E\big(\tilde{\boldsymbol{\theta}}(t)\big)\big| \ll \sqrt{\varepsilon}$ (which may happen during the late learning stage, or if non-standard hyperparameter values are chosen), the fraction in (3) with ε in the numerator approaches one, the dependence on ρ cancels out, and

  $$\big\|\nabla E\big(\tilde{\boldsymbol{\theta}}(t)\big)\big\|_{1,\varepsilon} \approx \sum_{i=1}^{p} \sqrt{\varepsilon}\big(1 + \big|\nabla_i E\big(\tilde{\boldsymbol{\theta}}(t)\big)\big|^2/(2\varepsilon)\big)$$
  $$= p\sqrt{\varepsilon} + \frac{1}{2\sqrt{\varepsilon}}\big\|\nabla E\big(\tilde{\boldsymbol{\theta}}(t)\big)\big\|^2. \tag{6}$$

  In other words, $\|\cdot\|_{1,\varepsilon}$ becomes $\|\cdot\|^2/(2\sqrt{\varepsilon})$ up to an additive constant, giving

  $$\text{correction}_j(\boldsymbol{\theta})$$
  $$\approx \big(4\sqrt{\varepsilon}\big)^{-1}(1-\beta)^{-1}(1+\beta)\nabla_j \|\nabla E(\boldsymbol{\theta})\|^2.$$

  The form of the ODE in this case is

  $$\dot{\tilde{\theta}}_j(t) = -\nabla_j \widetilde{E}\big(\tilde{\boldsymbol{\theta}}(t)\big), \quad \text{with}$$
  $$\widetilde{E}(\boldsymbol{\theta}) = \frac{1}{\sqrt{\varepsilon}}\left(E(\boldsymbol{\theta}) + \frac{h}{4\sqrt{\varepsilon}}\frac{1+\beta}{1-\beta}\|\nabla E(\boldsymbol{\theta})\|^2\right). \tag{7}$$

These two extreme cases are summarized in Table 1. In Figure 1, we use the one-dimensional ($p = 1$) case to illustrate what kind of term is being implicitly penalized.

Table 1. Implicit bias of Adam: special cases. "Small" and "large" are in relation to squared gradient components (Adam in the latter case is close to GD with momentum).

	ε "small"	ε "large"
$\rho > \beta$	$-\|\nabla E(\boldsymbol{\theta})\|_1$**-penalized**	$\|\nabla E(\boldsymbol{\theta})\|_2^2$-penalized
$\beta \geq \rho$	$\|\nabla E(\boldsymbol{\theta})\|_1$-penalized	$\|\nabla E(\boldsymbol{\theta})\|_2^2$-penalized

Usually, ε is chosen to be small, and during most of the training Adam is much better described by the first extreme case. It is clear from (5) that, if $\rho > \beta$, the correction term provides the opposite of regularization, in contrast to (1). The larger ρ compared to β, the stronger the anti-regularization effect is.

This finding may partially explain why adaptive gradient methods have been reported to generalize worse than non-adaptive ones (Chen et al., 2018; Wilson et al., 2017), as it offers a previously unknown perspective on why they are biased towards "higher-curvature" regions and find "sharper" minima. Indeed, note that standard (non-adaptive) ℓ_∞-sharpness at $\boldsymbol{\theta}$ can be defined by $\max_{\|\boldsymbol{\delta}\|_\infty \leq r} E(\boldsymbol{\theta} + \boldsymbol{\delta}) - E(\boldsymbol{\theta})$ for some radius r. This or similar definitions have been considered often in literature, see, e. g., Andriushchenko et al. (2023), Foret et al. (2021). Replacing the difference of the losses with its first-order approximation under the maximum (Foret et al., 2021; Ghosh et al., 2023)

$$\max_{\|\boldsymbol{\delta}\|_\infty \leq r} E(\boldsymbol{\theta} + \boldsymbol{\delta}) - E(\boldsymbol{\theta})$$
$$\approx \max_{\|\boldsymbol{\delta}\|_\infty \leq r} \nabla E(\boldsymbol{\theta})^\intercal \boldsymbol{\delta} = r\|\nabla E(\boldsymbol{\theta})\|_1,$$

we see that Adam typically anti-penalizes the approximation of ℓ_∞-sharpness. Although the connection between sharpness and generalization is not clear-cut (Andriushchenko et al., 2023), our empirical results (Section 5) are consistent with this theory.

This overview also applies to RMSProp by setting $\beta = 0$; see Theorem C.4 for the formal result.

Example 2.1 (Backward Error Analysis for GD with Heavy-ball Momentum)**.** Assume ε is large compared to all squared gradient components during the whole training process, so that the form of the ODE is approximated by (7). Since Adam with a large ε and after a certain number of iterations approximates SGD with heavy-ball momentum with step size $h(1-\beta)/\sqrt{\varepsilon}$, a linear step size change (and corresponding time change) gives exactly the equations in Theorem 4.1 of Ghosh et al. (2023). Taking $\beta = 0$ (no momentum), we get the implicit regularization of GD from Barrett & Dherin (2021).

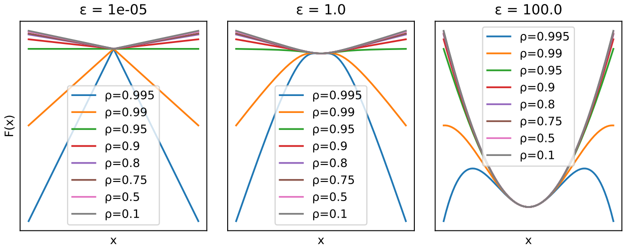

Figure 1. To illustrate what term is being implicitly penalized in the simple case $p = 1$, we plot the graphs of $x \mapsto F(x) := \frac{h}{2}\int_0^x \left\{\frac{1+\beta}{1-\beta} - \frac{1+\rho}{1-\rho} + \frac{1+\rho}{1-\rho}\cdot\frac{\varepsilon}{y^2+\varepsilon}\right\} \mathrm{d}\sqrt{\varepsilon + y^2}$ with $\beta = 0.95$. In this case, the correction term in (3) is itself the gradient of the function $F(E'(\theta))$, where E' is the derivative (=gradient) of the loss: specifically, correction $= \frac{\mathrm{d}}{\mathrm{d}\theta}F(E'(\theta))$. Hence, Adam's iteration penalizes $F(E'(\theta))$. If ε is small and $\rho > \beta$, the *negative* one-norm of the gradient is penalized (leftmost picture, highest values of ρ); in other words, the one-norm is *anti*-penalized.

3. Main Result: ODE Approximating Mini-Batch Adam

We only make one assumption, which is standard in the literature: the loss E_k for each mini-batch is 4 times continuously differentiable, and partial derivatives of E_k up to order 4 are bounded, i. e. there is a positive constant M such that for $\boldsymbol{\theta}$ in the region of interest

$$\sup_k \left\{ \sup_i \left|\nabla_i E_k(\boldsymbol{\theta})\right| \vee \sup_{i,j} \left|\nabla_{ij} E_k(\boldsymbol{\theta})\right| \vee \sup_{i,j,s} \left|\nabla_{ijs} E_k(\boldsymbol{\theta})\right| \vee \sup_{i,j,s,r} \left|\nabla_{ijsr} E_k(\boldsymbol{\theta})\right| \right\} \le M. \tag{8}$$

Theorem 3.1. *Assume* (8) *holds. Let* $\{\boldsymbol{\theta}^{(n)}\}$ *be iterations of Adam as defined in Definition* 1.1, $\tilde{\boldsymbol{\theta}}(t)$ *be the continuous solution to the piecewise ODE*

$$\begin{aligned}\dot{\tilde{\theta}}_j(t) = &-\frac{M_j^{(n)}(\tilde{\boldsymbol{\theta}}(t))}{R_j^{(n)}(\tilde{\boldsymbol{\theta}}(t))} \\ &+ h\Bigg(\frac{M_j^{(n)}(\tilde{\boldsymbol{\theta}}(t))\left(2P_j^{(n)}(\tilde{\boldsymbol{\theta}}(t)) + \bar{P}_j^{(n)}(\tilde{\boldsymbol{\theta}}(t))\right)}{2R_j^{(n)}(\tilde{\boldsymbol{\theta}}(t))^3} \\ &\qquad - \frac{2L_j^{(n)}(\tilde{\boldsymbol{\theta}}(t)) + \bar{L}_j^{(n)}(\tilde{\boldsymbol{\theta}}(t))}{2R_j^{(n)}(\tilde{\boldsymbol{\theta}}(t))}\Bigg)\end{aligned} \tag{9}$$

for $t \in [nh, (n+1)h]$ *with the initial condition* $\tilde{\boldsymbol{\theta}}(0) = \boldsymbol{\theta}^{(0)}$, *where*

$$R_j^{(n)}(\boldsymbol{\theta}) := \sqrt{\frac{\sum_{k=0}^n \rho^{n-k}(1-\rho)(\nabla_j E_k(\boldsymbol{\theta}))^2}{1-\rho^{n+1}} + \varepsilon},$$

$$M_j^{(n)}(\boldsymbol{\theta}) := \frac{\sum_{k=0}^n \beta^{n-k}(1-\beta)\nabla_j E_k(\boldsymbol{\theta})}{1-\beta^{n+1}},$$

$$\begin{aligned}L_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\beta^{n+1}}\sum_{k=0}^n \beta^{n-k}(1-\beta) \\ &\times \sum_{i=1}^p \nabla_{ij}E_k(\boldsymbol{\theta}) \sum_{l=k}^{n-1} \frac{M_i^{(l)}(\boldsymbol{\theta})}{R_i^{(l)}(\boldsymbol{\theta})}, \\ \bar{L}_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\beta^{n+1}}\sum_{k=0}^n \beta^{n-k}(1-\beta) \\ &\times \sum_{i=1}^p \nabla_{ij}E_k(\boldsymbol{\theta}) \frac{M_i^{(n)}(\boldsymbol{\theta})}{R_i^{(n)}(\boldsymbol{\theta})}, \\ P_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\rho^{n+1}}\sum_{k=0}^n \rho^{n-k}(1-\rho)\nabla_j E_k(\boldsymbol{\theta}) \\ &\times \sum_{i=1}^p \nabla_{ij}E_k(\boldsymbol{\theta}) \sum_{l=k}^{n-1} \frac{M_i^{(l)}(\boldsymbol{\theta})}{R_i^{(l)}(\boldsymbol{\theta})}, \\ \bar{P}_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\rho^{n+1}}\sum_{k=0}^n \rho^{n-k}(1-\rho)\nabla_j E_k(\boldsymbol{\theta}) \\ &\times \sum_{i=1}^p \nabla_{ij}E_k(\boldsymbol{\theta}) \frac{M_i^{(n)}(\boldsymbol{\theta})}{R_i^{(n)}(\boldsymbol{\theta})}.\end{aligned}$$

Then, for any fixed positive time horizon $T > 0$ *there exists a constant* C *(depending on* T, ρ, β, ε*) such that for any step size* $h \in (0, T)$ *we have* $\left\|\tilde{\boldsymbol{\theta}}(nh) - \boldsymbol{\theta}^{(n)}\right\| \le Ch^2$ *for* $n \in \{0, \dots, \lfloor T/h \rfloor\}$.

Remark 3.2. In the *full-batch* setting $E_k \equiv E$, the terms above simplify to

$$\begin{aligned}R_j^{(n)}(\boldsymbol{\theta}) &= \left(|\nabla_j E(\boldsymbol{\theta})|^2 + \varepsilon\right)^{1/2}, \\ M_j^{(n)}(\boldsymbol{\theta}) &= \nabla_j E(\boldsymbol{\theta}), \\ L_j^{(n)}(\boldsymbol{\theta}) &= \left[\frac{\beta}{1-\beta} - \frac{(n+1)\beta^{n+1}}{1-\beta^{n+1}}\right]\bar{L}_j^{(n)}(\boldsymbol{\theta}), \\ \bar{L}_j^{(n)}(\boldsymbol{\theta}) &= \nabla_j \|\nabla E(\boldsymbol{\theta})\|_{1,\varepsilon}, \\ P_j^{(n)}(\boldsymbol{\theta}) &= \left[\frac{\rho}{1-\rho} - \frac{(n+1)\rho^{n+1}}{1-\rho^{n+1}}\right]\bar{P}_j^{(n)}(\boldsymbol{\theta}), \\ \bar{P}_j^{(n)}(\boldsymbol{\theta}) &= \nabla_j E(\boldsymbol{\theta}) \nabla_j \|\nabla E(\boldsymbol{\theta})\|_{1,\varepsilon}.\end{aligned}$$

If the iteration number n is large, (9) rapidly becomes as described in (2) and (3).

Derivation sketch. The proof is in the appendix (this is Theorem E.4; see Appendix A for the overview of the appendix). To help the reader understand how the ODE (9) is obtained, apart from the full proof, we include an informal derivation in Appendix I, and provide an even briefer sketch of this derivation here.

Our goal is to find such a trajectory $\tilde{\boldsymbol{\theta}}(t)$ that, denoting

$t_n := nh$, we have

$$\tilde{\theta}_j(t_{n+1}) = \tilde{\theta}_j(t_n) - h\frac{T^{(n)}_{\beta,j}}{\sqrt{T^{(n)}_{\rho,j}}} + O(h^3) \quad \text{with} \tag{10}$$

$$T^{(n)}_{\beta,j} := \frac{1}{1-\beta^{n+1}}\sum_{k=0}^{n} \beta^{n-k}(1-\beta)\nabla_j E_k\big(\tilde{\boldsymbol{\theta}}(t_k)\big),$$

$$T^{(n)}_{\rho,j} := \frac{1}{1-\rho^{n+1}}\sum_{k=0}^{n} \rho^{n-k}(1-\rho)\big(\nabla_j E_k\big(\tilde{\boldsymbol{\theta}}(t_k)\big)\big)^2 + \varepsilon.$$

Ignoring the terms of order higher than one in h, we can take a first-order approximation for granted: $\tilde{\theta}_j(t_{n+1}) = \tilde{\theta}_j(t_n) - hA^{(n)}_j(\tilde{\boldsymbol{\theta}}(t_n)) + O(h^2)$ with $A^{(n)}_j(\boldsymbol{\theta}) := M^{(n)}_j(\boldsymbol{\theta})/R^{(n)}_j(\boldsymbol{\theta})$. The challenge is to make this more precise by finding an equality of the form

$$\begin{aligned}\tilde{\theta}_j(t_{n+1}) &= \tilde{\theta}_j(t_n)\\ &\quad - hA^{(n)}_j\big(\tilde{\boldsymbol{\theta}}(t_n)\big) + h^2 B^{(n)}_j\big(\tilde{\boldsymbol{\theta}}(t_n)\big) + O(h^3),\end{aligned} \tag{11}$$

where $B^{(n)}_j(\cdot)$ is a known function. This is a numerical iteration to which standard backward error analysis (Chapter IX in Ernst Hairer & Wanner (2006)) can be applied.

Using the Taylor series, we can write

$$\begin{aligned}\nabla_j E_k\big(\tilde{\boldsymbol{\theta}}(t_{n-1})\big) &= \nabla_j E_k\big(\tilde{\boldsymbol{\theta}}(t_n)\big)\\
&\quad + \sum_{i=1}^{p} \nabla_{ij} E_k\big(\tilde{\boldsymbol{\theta}}(t_n)\big)\big\{\tilde{\theta}_i(t_{n-1}) - \tilde{\theta}_i(t_n)\big\} + O(h^2)\\
&= \nabla_j E_k\big(\tilde{\boldsymbol{\theta}}(t_n)\big)\\
&\quad + h\sum_{i=1}^{p} \nabla_{ij} E_k\big(\tilde{\boldsymbol{\theta}}(t_n)\big)\frac{M^{(n-1)}_i\big(\tilde{\boldsymbol{\theta}}(t_{n-1})\big)}{R^{(n-1)}_i\big(\tilde{\boldsymbol{\theta}}(t_{n-1})\big)} + O(h^2)\\
&= \nabla_j E_k\big(\tilde{\boldsymbol{\theta}}(t_n)\big)\\
&\quad + h\sum_{i=1}^{p} \nabla_{ij} E_k\big(\tilde{\boldsymbol{\theta}}(t_n)\big)\frac{M^{(n-1)}_i\big(\tilde{\boldsymbol{\theta}}(t_n)\big)}{R^{(n-1)}_i\big(\tilde{\boldsymbol{\theta}}(t_n)\big)} + O(h^2),\end{aligned}$$

where in the last equality we just replaced t_{n-1} with t_n in the h-term since it only affects higher-order terms. Doing this again for steps $n-2, n-3, \ldots$, and adding the resulting equations will give for $k < n$

$$\begin{aligned}\nabla_j E_k\big(\tilde{\boldsymbol{\theta}}(t_k)\big) &= \nabla_j E_k\big(\tilde{\boldsymbol{\theta}}(t_n)\big)\\
&\quad + h\sum_{i=1}^{p} \nabla_{ij} E_k\big(\tilde{\boldsymbol{\theta}}(t_n)\big)\sum_{l=k}^{n-1}\frac{M^{(l)}_i\big(\tilde{\boldsymbol{\theta}}(t_n)\big)}{R^{(l)}_i\big(\tilde{\boldsymbol{\theta}}(t_n)\big)} + O(h^2),\end{aligned}$$

where we could safely ignore that $n-k$ is not bounded because of exponential averaging. Taking the square of this formal power series in h, multiplying this square by $\rho^{n-k}(1-\rho)$ and summing up over k will give

$$\begin{aligned}&\frac{1}{1-\rho^{n+1}}\sum_{k=0}^{n} \rho^{n-k}(1-\rho)\big[\nabla_j E_k\big(\tilde{\boldsymbol{\theta}}(t_k)\big)\big]^2 + \varepsilon\\
&= R^{(n)}_j\big(\tilde{\boldsymbol{\theta}}(t_n)\big)^2 + 2hP^{(n)}_j\big(\tilde{\boldsymbol{\theta}}(t_n)\big) + O(h^2),\end{aligned}$$

which, using the expression for the inverse square root $\big(\sum_{r=0}^{\infty} a_r h^r\big)^{-1/2}$ of a formal power series $\sum_{r=0}^{\infty} a_r h^r$, gives us an expansion

$$\frac{1}{\sqrt{T^{(n)}_{\rho,j}}} = \frac{1}{R^{(n)}_j\big(\tilde{\boldsymbol{\theta}}(t_n)\big)} - h\frac{P^{(n)}_j\big(\tilde{\boldsymbol{\theta}}(t_n)\big)}{R^{(n)}_j\big(\tilde{\boldsymbol{\theta}}(t_n)\big)^3} + O(h^2).$$

A similar process provides an expansion for $T^{(n)}_{\beta,j}$:

$$T^{(n)}_{\beta,j} = M^{(n)}_j\big(\tilde{\boldsymbol{\theta}}(t_n)\big) + hL^{(n)}_j\big(\tilde{\boldsymbol{\theta}}(t_n)\big) + O(h^2).$$

Inserting these two expansions into (10) leads to an expression for $B^{(n)}_j(\cdot)$:

$$B^{(n)}_j(\boldsymbol{\theta}) = \frac{M^{(n)}_j(\boldsymbol{\theta})P^{(n)}_j(\boldsymbol{\theta})}{R^{(n)}_j(\boldsymbol{\theta})^3} - \frac{L^{(n)}_j(\boldsymbol{\theta})}{R^{(n)}_j(\boldsymbol{\theta})}.$$

We are now ready to find an ODE for $t \in [t_n, t_{n+1}]$ of the form $\dot{\tilde{\boldsymbol{\theta}}} = \tilde{\mathbf{f}}\big(\tilde{\boldsymbol{\theta}}(t)\big)$ whose discretization is (11). This is a task for standard backward error analysis: expand $\tilde{\mathbf{f}}(\cdot)$ into $\tilde{\mathbf{f}}(\boldsymbol{\theta}) = \mathbf{f}(\boldsymbol{\theta}) + h\mathbf{f}_1(\boldsymbol{\theta}) + O(h^2)$. By Taylor expansion, we have

$$\begin{aligned}\tilde{\boldsymbol{\theta}}(t_{n+1}) &= \tilde{\boldsymbol{\theta}}(t_n) + h\dot{\tilde{\boldsymbol{\theta}}}(t_n^+) + \frac{h^2}{2}\ddot{\tilde{\boldsymbol{\theta}}}(t_n^+) + O(h^3)\\
&= \tilde{\boldsymbol{\theta}}(t_n) + h\big[\mathbf{f}\big(\tilde{\boldsymbol{\theta}}(t_n)\big) + h\mathbf{f}_1\big(\tilde{\boldsymbol{\theta}}(t_n)\big) + O(h^2)\big]\\
&\quad + \frac{h^2}{2}\big[\nabla\mathbf{f}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)\mathbf{f}\big(\tilde{\boldsymbol{\theta}}(t_n)\big) + O(h)\big] + O(h^3)\\
&= \tilde{\boldsymbol{\theta}}(t_n) + h\mathbf{f}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)\\
&\quad + h^2\left[\mathbf{f}_1\big(\tilde{\boldsymbol{\theta}}(t_n)\big) + \frac{\nabla\mathbf{f}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)\mathbf{f}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)}{2}\right] + O(h^3).\end{aligned}$$

It is left to equate the terms before the corresponding powers of h here and in (11), giving $f_j(\boldsymbol{\theta}) = -A^{(n)}_j(\boldsymbol{\theta})$ and $f_{1,j}(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{p}\nabla_i f_j(\boldsymbol{\theta})f_i(\boldsymbol{\theta}) + B^{(n)}_j(\boldsymbol{\theta})$. Omitting some algebra, the piecewise ODE (9) is derived. □

4. Illustration: Simple Bilinear Model

We now analyze the effect of the first-order term for Adam in the same model as Barrett & Dherin (2021) and Ghosh et al. (2023) have studied. Namely, assume the parameter $\boldsymbol{\theta} = (\theta_1, \theta_2)^{\mathsf{T}}$ is 2-dimensional, and the loss is given by $E(\boldsymbol{\theta}) := 1/2(3/2 - 2\theta_1\theta_2)^2$. The loss is minimized on the hyperbola $\theta_1\theta_2 = 3/4$. We graph the trajectories of Adam in this case: the left part of Figure 2 shows that increasing β forces the trajectory to the region with smaller $\|\nabla E(\boldsymbol{\theta})\|_1$, and increasing ρ does the opposite. The right part shows that increasing the learning rate moves Adam towards the region with smaller $\|\nabla E(\boldsymbol{\theta})\|_1$ if $\beta > \rho$ (just like in the case of GD, except the norm is different if ε is small compared to gradient components), and does the opposite if $\rho > \beta$. All these observations are exactly what Theorem 3.1 predicts.

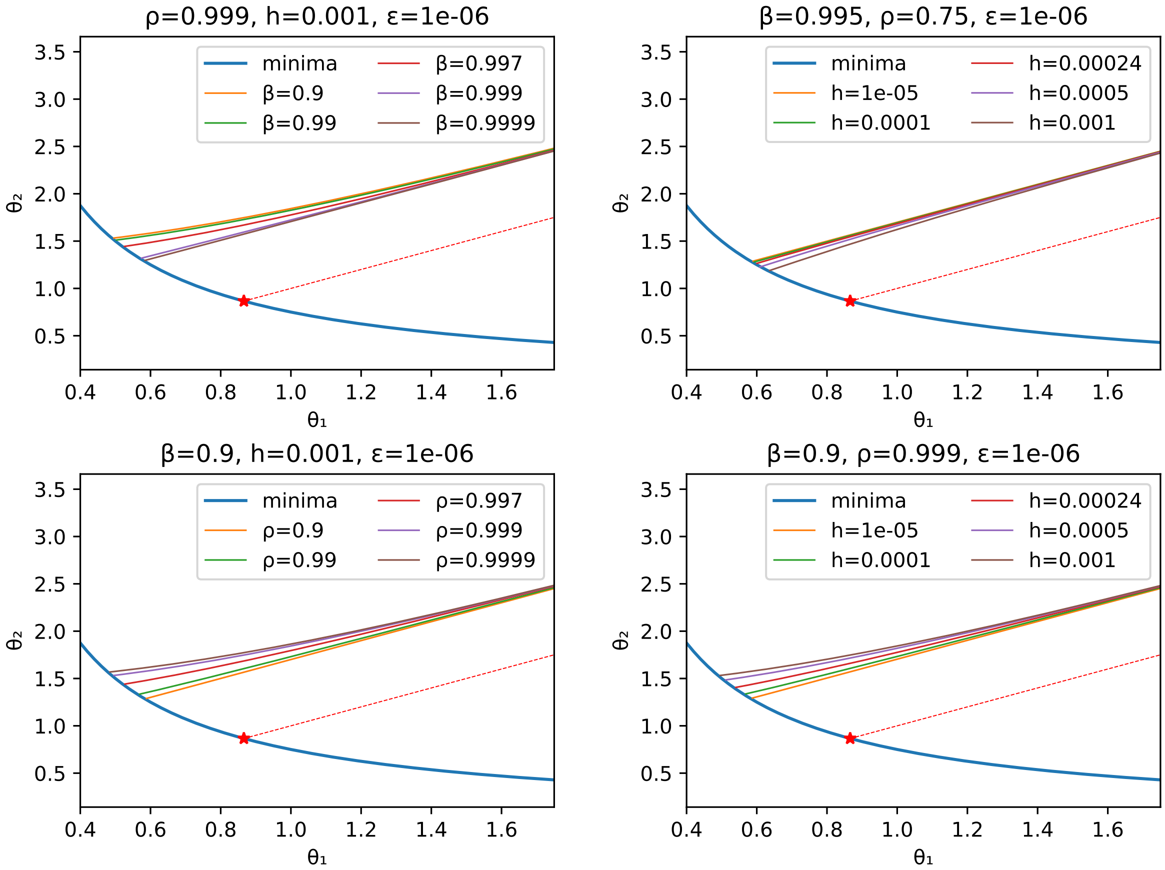

Figure 2. Left: increasing β moves the trajectory of Adam towards the regions with smaller one-norm of the gradient (if ε is sufficiently small); increasing ρ does the opposite. Right: increasing the learning rate moves the Adam trajectory towards the regions with smaller one-norm of the gradient if β is significantly larger than ρ and does the opposite if ρ is larger than β. The cross denotes the limit point of gradient one-norm minimizers on the level sets $4\theta_1\theta_2 - 3 = c$. The minimizers are drawn with a dashed line. All Adam trajectories start at $(2.8, 3.5)$.

5. Numerical Experiments

As a first sanity check, we train a relatively small fully-connected neural network with around 10^5 parameters on the first 10,000 images of MNIST with full-batch Adam for 100 epochs and plot the value $\|\boldsymbol{\theta}^{(n)} - \tilde{\boldsymbol{\theta}}(t_n)\|_\infty$, i. e. the maximal weight difference between the Adam iteration and the piecewise ODE solution.[1] We see in Figure 3 that even on this very large time horizon the trajectories are close in infinity-norm.

Further, we offer some preliminary empirical evidence that Adam (anti-)penalizes the perturbed one-norm of the gradients, as discussed in Section 2.

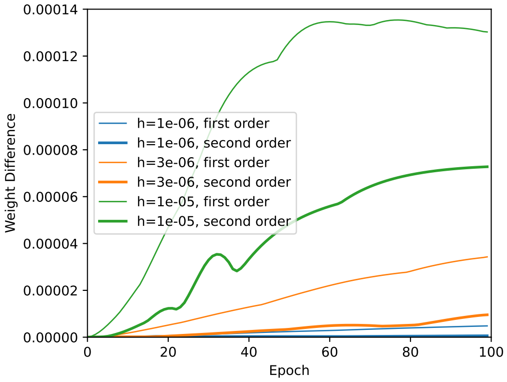

Figure 3. $\|\boldsymbol{\theta}^{(n)} - \tilde{\boldsymbol{\theta}}(t_n)\|_\infty$ for a MLP trained with full-batch Adam on truncated MNIST, where $\tilde{\boldsymbol{\theta}}(t_n)$ is either first (signGD perturbed by ε) or second order approximation to Adam; $\beta = 0.9$, $\rho = 0.95$, $\varepsilon = 10^{-6}$. Precise definitions are provided in Appendix H, specifically (63).

[1] Since it makes little sense to numerically solve an ODE by further discretization, $\tilde{\boldsymbol{\theta}}(t_n)$ is estimated using the iteration (11) with $O(h^3)$ ignored. Strictly speaking, this is not the trajectory obtained by the final backward error analysis step but rather the step immediately preceding it (after removing long-term memory but before converting the iteration to an ODE).

Ma et al. (2022) divide training regimes of Adam into three categories: the spike regime when ρ is much larger than β, in which the training loss curve contains very large spikes and the training is obviously unstable; the (stable) oscillation regime when ρ is sufficiently close to β, in which the loss curve contains fast and small oscillations; the divergence regime when β is much larger than ρ, in which Adam diverges. We exclude the last regime. In the spike regime, the loss spikes to large values at irregular intervals. This has also been observed in the context of large transformers, and mitigation strategies have been proposed in Chowdhery et al. (2022) and Molybog et al. (2023). Since it is unlikely that an unstable Adam trajectory can be meaningfully approximated by a smooth ODE solution, we reduce the incidence of large spikes by only considering β and ρ that are not too far apart, which is what Ma et al. (2022) recommend to do in practice.

We train Resnet-50, CNNs and Vision Transformers (Dosovitskiy et al., 2020) on the CIFAR-10 dataset with full-batch Adam. In this section, we provide the results for Resnet-50; the pictures for CNNs and Transformers are similar and are given in Appendix H.4. Figure 4 shows that in the stable oscillation regime increasing ρ appears to increase the perturbed one-norm (consistent with our analysis: the smaller ρ, the more this "norm" is penalized) and decrease the test accuracy. Figure 5 shows that increasing β appears to decrease the perturbed one-norm (consistent with our analysis: the larger β, the more this norm is penalized) and increase the test accuracy. The picture confirms the finding in Ghosh et al. (2023) (for the momentum parameter in momentum GD).

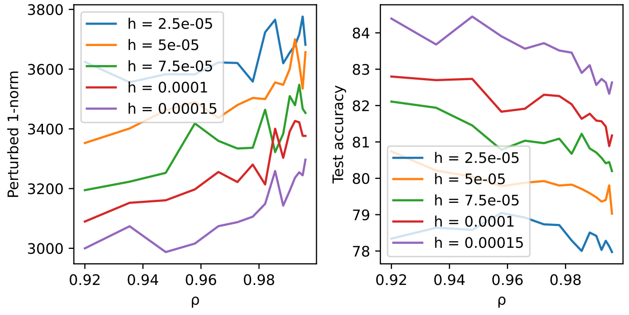

Figure 4. Resnet-50 on CIFAR-10 trained with full-batch Adam, $\varepsilon = 10^{-8}$, $\beta = 0.99$. As ρ increases, the norm rises and the test accuracy falls. We train longer than necessary for near-perfect classification on the train dataset (at least 2-3 thousand epochs), and the test accuracies plotted here are maximal. The perturbed norms are also maximal after excluding the initial training period (i. e., the plotted "norms" are at peaks of the "hills" described in Section 5). All results are averaged across five runs with different initialization seeds. Additional evidence and more details are provided in Appendix H.

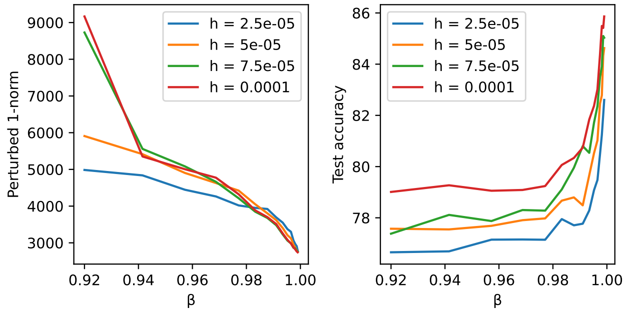

Figure 5. Resnet-50 on CIFAR-10 trained with full-batch Adam, $\rho = 0.999$, $\varepsilon = 10^{-8}$. The perturbed one-norm falls as β increases, and the test accuracy rises. Both metrics are calculated as in Figure 4. All results are averaged across three runs with different initialization seeds.

Figure 6 shows the graphs of $\|\nabla E\|_{1,\varepsilon}$ as functions of the epoch number. The "norm" decreases, then rises again, and then decreases further until it flatlines.[2] Throughout most of the training, the larger β the smaller the "norm". The "hills" of the "norm" curves are higher with smaller β and larger ρ. This is consistent with our analysis because the larger ρ compared to β, the more $\|\nabla E\|_{1,\varepsilon}$ is prevented from falling by the correction term.

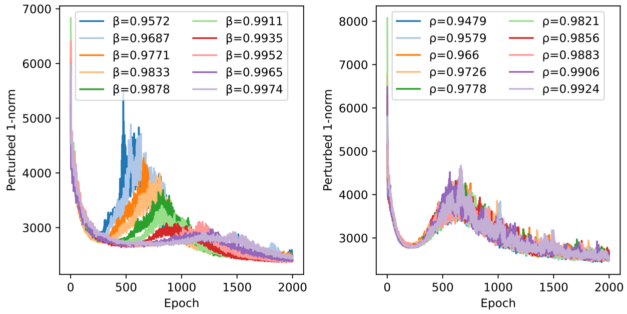

Figure 6. Plots of $\|\nabla E\|_{1,\varepsilon}$ after each epoch for full-batch Adam, $h = 10^{-4}, \varepsilon = 10^{-8}$. Resnet-50 on CIFAR-10, left: $\rho = 0.999$, right: $\beta = 0.97$.

6. Limitations and Future Directions

As far as we know, the assumption similar to (8) is explicitly or implicitly present in all previous work on backward error analysis of gradient-based machine learning algorithms. (Recently, Beneventano (2023) weakened this assumption for SGD without replacement, but their focus is somewhat different.) There is evidence that large-batch algorithms often operate near or at the edge of stability (Cohen et al., 2021; 2022), in which the largest eigenvalue of the hessian can be large, making it unclear whether the higher-order partial derivatives can safely be assumed bounded near op-

[2] Note that the perturbed one-norm cannot be near-zero at the end of training because it is bounded from below by $p\sqrt{\varepsilon}$.

timality. In addition, as Smith et al. (2021) point out, in the mini-batch setting backward error analysis can be more accurate. We leave a qualitative analysis of the behavior of first-order terms in Theorem 3.1 in the mini-batch case as a future direction.

Relatedly, Adam does not always generalize worse than SGD: for transformers, Adam often outperforms (Zhang et al., 2020; Kumar et al., 2022). Moreover, for NLP tasks a long time can be spent training close to an interpolating solution. Our analysis suggests that in the latter regime the anti-regularization effect disappears, which does indeed confirm the finding that generalization can be better. However, we believe this explanation is not complete, and more work is needed to connect the implicit bias to the training dynamics of transformers.

In addition, the constant C in Theorem 3.1 goes to infinity as ε goes to zero. Theoretically, our proof does not exclude the case where for very small ε the trajectory of the piecewise ODE is only close to the Adam trajectory for small, suboptimal learning rates, at least at later stages of learning. (For the initial learning period, this is not a problem.) It appears to also be true of Proposition 1 in Ma et al. (2022) (zeroth-order approximation by sign-GD). This is especially noticeable in the large-spike regime of training (see Section 5) which, despite being obviously unstable, can still lead to acceptable test errors. It would be worthwhile to investigate this regime in detail.

Acknowledgments

We extend our special thanks to Boris Hanin, Sam Smith, and the anonymous reviewers for their insightful comments and suggestions that greatly enhanced this work. We are also grateful to Jianqing Fan, Pier Beneventano, and Rae Yu for engaging and productive discussions. Cattaneo gratefully acknowledges financial support from the National Science Foundation through DMS-2210561 and SES-2241575. Klusowski gratefully acknowledges financial support from the National Science Foundation through CAREER DMS-2239448. Additionally, we acknowledge the Princeton Research Computing resources, coordinated by the Princeton Institute for Computational Science and Engineering (PICSciE) and the Office of Information Technology's Research Computing.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Andriushchenko, M., Croce, F., Müller, M., Hein, M., and Flammarion, N. A modern look at the relationship between sharpness and generalization. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Balles, L., Pedregosa, F., and Roux, N. L. The geometry of sign gradient descent. *arXiv preprint arXiv:2002.08056*, 2020.

Barakat, A. and Bianchi, P. Convergence and dynamical behavior of the adam algorithm for nonconvex stochastic optimization. *SIAM Journal on Optimization*, 31(1):244–274, 2021.

Barrett, D. and Dherin, B. Implicit gradient regularization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=3q5IqUrkcF.

Beneventano, P. On the trajectories of sgd without replacement. *arXiv preprint arXiv:2312.16143*, 2023.

Bernstein, J., Wang, Y.-X., Azizzadenesheli, K., and Anandkumar, A. signsgd: Compressed optimisation for nonconvex problems. In *International Conference on Machine Learning*, pp. 560–569. PMLR, 2018.

Beyer, L., Zhai, X., and Kolesnikov, A. Better plain vit baselines for imagenet-1k. *arXiv preprint arXiv:2205.01580*, 2022.

Chen, J., Zhou, D., Tang, Y., Yang, Z., Cao, Y., and Gu, Q. Closing the generalization gap of adaptive gradient methods in training deep neural networks. *arXiv preprint arXiv:1806.06763*, 2018. URL https://arxiv.org/pdf/1806.06763.

Chen, X., Hsieh, C.-J., and Gong, B. When vision transformers outperform resnets without pre-training or strong data augmentations. *arXiv preprint arXiv:2106.01548*, 2021.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Cohen, J., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jh-rTtvkGeM.

Cohen, J. M., Ghorbani, B., Krishnan, S., Agarwal, N., Medapati, S., Badura, M., Suo, D., Cardoze, D., Nado, Z., Dahl, G. E., et al. Adaptive gradient methods at the edge of stability. *arXiv preprint arXiv:2207.14484*, 2022.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Ernst Hairer, C. L. and Wanner, G. *Geometric numerical integration*. Springer-Verlag, Berlin, 2 edition, 2006. ISBN 3-540-30663-3.

Even, M., Pesme, S., Gunasekar, S., and Flammarion, N. (s) gd over diagonal linear networks: Implicit regularisation, large stepsizes and edge of stability. *arXiv preprint arXiv:2302.08982*, 2023.

Farazmand, M. Multiscale analysis of accelerated gradient methods. *SIAM Journal on Optimization*, 30(3):2337–2354, 2020.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

França, G., Jordan, M. I., and Vidal, R. On dissipative symplectic integration with applications to gradient-based optimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(4):043402, 2021.

Ghosh, A., Lyu, H., Zhang, X., and Wang, R. 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.

Granziol, D. Flatness is a false friend. *arXiv preprint arXiv:2006.09091*, 2020.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pp. 1832–1841. PMLR, 2018a.

Gunasekar, S., Lee, J. D., Soudry, D., and Srebro, N. Implicit bias of gradient descent on linear convolutional networks. *Advances in neural information processing systems*, 31, 2018b.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hoffer, E., Hubara, I., and Soudry, D. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in neural information processing systems*, 30, 2017.

Ji, Z. and Telgarsky, M. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018a.

Ji, Z. and Telgarsky, M. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018b.

Ji, Z. and Telgarsky, M. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pp. 1772–1798. PMLR, 2019.

Ji, Z. and Telgarsky, M. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020.

Jiang, K., Malik, D., and Li, Y. How does adaptive optimization impact local neural network geometry? *arXiv preprint arXiv:2211.02254*, 2022.

Keskar, N. S. and Socher, R. Improving generalization performance by switching from adam to sgd. *arXiv preprint arXiv:1712.07628*, 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Kovachki, N. B. and Stuart, A. M. Continuous time analysis of momentum methods. *Journal of Machine Learning Research*, 22(17):1–40, 2021.

Kumar, A., Shen, R., Bubeck, S., and Gunasekar, S. How to fine-tune vision models with sgd. *arXiv preprint arXiv:2211.09359*, 2022.

Kunin, D., Sagastuy-Brena, J., Ganguli, S., Yamins, D. L., and Tanaka, H. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. *arXiv preprint arXiv:2012.04728*, 2020.

Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., and Tu, Z. Deeply-supervised nets. In *Artificial intelligence and statistics*, pp. 562–570. Pmlr, 2015.

Li, Q., Tai, C., and E, W. Stochastic modified equations and adaptive stochastic gradient algorithms. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2101–2110. PMLR, 8 2017. URL https://proceedings.mlr.press/v70/li17f.html.

Lyu, K. and Li, J. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

Ma, C., Wu, L., and Weinan, E. A qualitative study of the dynamic behavior for adaptive gradient algorithms. In *Mathematical and Scientific Machine Learning*, pp. 671–692. PMLR, 2022.

Malladi, S., Lyu, K., Panigrahi, A., and Arora, S. On the sdes and scaling rules for adaptive gradient algorithms. *Advances in Neural Information Processing Systems*, 35: 7697–7711, 2022.

Miyagawa, T. Toward equation of motion for deep neural networks: Continuous-time gradient descent and discretization error analysis. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=qq84D17BPu.

Molybog, I., Albert, P., Chen, M., DeVito, Z., Esiobu, D., Goyal, N., Koura, P. S., Narang, S., Poulton, A., Silva, R., et al. A theory on adam instability in large-scale machine learning. *arXiv preprint arXiv:2304.09871*, 2023.

Nacson, M. S., Gunasekar, S., Lee, J., Srebro, N., and Soudry, D. Lexicographic and depth-sensitive margins in homogeneous and non-homogeneous deep models. In *International Conference on Machine Learning*, pp. 4683–4692. PMLR, 2019a.

Nacson, M. S., Lee, J., Gunasekar, S., Savarese, P. H. P., Srebro, N., and Soudry, D. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3420–3428. PMLR, 2019b.

Nacson, M. S., Srebro, N., and Soudry, D. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3051–3059. PMLR, 2019c.

Qian, Q. and Qian, X. The implicit bias of adagrad on separable data. *Advances in Neural Information Processing Systems*, 32, 2019.

Rosca, M. C., Wu, Y., Dherin, B., and Barrett, D. Discretization drift in two-player games. In *International Conference on Machine Learning*, pp. 9064–9074. PMLR, 2021.

Smith, S. L., Dherin, B., Barrett, D., and De, S. On the origin of implicit regularization in stochastic gradient descent. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=rq_Qr0c1Hyo.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Tieleman, T., Hinton, G., et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4 (2):26–31, 2012.

Wang, B., Meng, Q., Chen, W., and Liu, T.-Y. The implicit bias for adaptive optimization algorithms on homogeneous neural networks. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10849–10858. PMLR, 7 2021. URL https://proceedings.mlr.press/v139/wang21q.html.

Wang, B., Meng, Q., Zhang, H., Sun, R., Chen, W., Ma, Z.-M., and Liu, T.-Y. Does momentum change the implicit regularization on separable data? *Advances in Neural Information Processing Systems*, 35:26764–26776, 2022.

Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., and Recht, B. The marginal value of adaptive gradient methods in machine learning. *Advances in neural information processing systems*, 30, 2017.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

Xie, Z., Wang, X., Zhang, H., Sato, I., and Sugiyama, M. Adaptive inertia: Disentangling the effects of adaptive learning rate and momentum. In *International conference on machine learning*, pp. 24430–24459. PMLR, 2022.

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S., Kumar, S., and Sra, S. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

Zhao, Y., Zhang, H., and Hu, X. Penalizing gradient norm for efficiently improving generalization in deep learning. In *International Conference on Machine Learning*, pp. 26982–26992. PMLR, 2022.

Zhou, P., Feng, J., Ma, C., Xiong, C., Hoi, S. C. H., et al. Towards theoretically understanding why sgd generalizes better than adam in deep learning. *Advances in Neural Information Processing Systems*, 33:21285–21296, 2020.

A. Overview

The appendix provide 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 B.4, Theorem C.4, Theorem D.4 and Theorem E.4, each residing in the section completely devoted to one algorithm. The simple induction argument taken from Ghosh et al. (2023), essentially the same for each of these theorems, is based on an auxiliary result whose corresponding versions are Theorem B.3, Theorem C.3, Theorem D.3 and Theorem E.3. The proof of this result is also elementary but long, and it is done by a series of lemmas in Appendix F and Appendix G. Out of these four, we only prove Theorem B.3 since the other three results are proven in the same way with obvious changes.

Appendix H contains some details about the numerical experiments.

A.1. 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 $n \in \{0, 1, \ldots\}$ 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 B.1, Definition C.1, Definition D.1, Definition E.1. This way, we avoid cluttering the notation significantly. We are careful to reference the relevant definition in all theorem statements.

B. RMSProp with ε Outside the Square Root

Definition B.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

$$\begin{aligned} \nu_j^{(n+1)} &= \rho \nu_j^{(n)} + (1 - \rho)\left(\nabla_j E_n\left(\boldsymbol{\theta}^{(n)}\right)\right)^2, \\ \theta_j^{(n+1)} &= \theta_j^{(n)} - \frac{h}{\sqrt{\nu_j^{(n+1)}} + \varepsilon} \nabla_j E_n\left(\boldsymbol{\theta}^{(n)}\right). \end{aligned} \tag{12}$$

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

$$\begin{aligned} \dot{\tilde{\theta}}_j(t) = &-\frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \varepsilon} \\ &+ h\left(\frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\left(2P_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \bar{P}_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)}{2\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \varepsilon\right)^2 R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)} - \frac{\sum_{i=1}^p \nabla_{ij} E_n\left(\tilde{\boldsymbol{\theta}}(t)\right) \frac{\nabla_i E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \varepsilon}}{2\left(R_j^{(n)}(\tilde{\boldsymbol{\theta}}(t)) + \varepsilon\right)}\right) \end{aligned} \tag{13}$$

for $t \in [t_n, t_{n+1}]$ 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

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

$$\bar{P}_j^{(n)}(\boldsymbol{\theta}) := \sum_{k=0}^{n} \rho^{n-k}(1-\rho)\nabla_j E_k(\boldsymbol{\theta}) \sum_{i=1}^{p} \nabla_{ij} E_k(\boldsymbol{\theta}) \frac{\nabla_i E_n(\boldsymbol{\theta})}{R_i^{(n)}(\boldsymbol{\theta}) + \varepsilon}.$$

Assumption B.2.

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

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

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

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

where $\tilde{\boldsymbol{\theta}}(t)$ is defined in Definition B.1.

Theorem B.3 (RMSProp with ε outside: local error bound)**.** *Suppose Assumption B.2 holds. Then for all $n \in \{0, 1, \dots, \lfloor T/h \rfloor\}$, $j \in \{1, \dots, p\}$*

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

for a positive constant C_1 depending on ρ.

The proof of Theorem B.3 is conceptually simple but very technical, and we delay it until Appendix G. For now assuming it as given and combining it with a simple induction argument gives a global error bound which follows.

Theorem B.4 (RMSProp with ε outside: global error bound)**.** *Suppose Assumption B.2 holds, and*

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

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

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

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 Ghosh et al. (2023).

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 B.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

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

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 B.3 with (12), we have

$$\left| [\mathbf{e}_{n+1} - \mathbf{e}_n]_j \right| \leq C_1 h^3 + h \left| \frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t_n)\right)}{\sqrt{A} + \varepsilon} - \frac{\nabla_j E_n\left(\boldsymbol{\theta}^{(n)}\right)}{\sqrt{B} + \varepsilon} \right|, \tag{14}$$

where to simplify notation we put

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

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

$$\left| \frac{1}{\sqrt{A} + \varepsilon} - \frac{1}{\sqrt{B} + \varepsilon} \right| = \frac{|A - B|}{\left(\sqrt{A} + \varepsilon\right)\left(\sqrt{B} + \varepsilon\right)\left(\sqrt{A} + \sqrt{B}\right)} \leq \frac{|A - B|}{2R(R + \varepsilon)^2}. \tag{15}$$

But since

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

we have

$$|A - B| \leq 2M_1 M_2 \sqrt{p} \sum_{k=0}^{n} \rho^{n-k}(1-\rho) \left\| \tilde{\boldsymbol{\theta}}(t_k) - \boldsymbol{\theta}^{(k)} \right\|. \tag{16}$$

Combining (15) and (16), we obtain

$$\begin{aligned}
&\left| \frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t_n)\right)}{\sqrt{A} + \varepsilon} - \frac{\nabla_j E_n\left(\boldsymbol{\theta}^{(n)}\right)}{\sqrt{B} + \varepsilon} \right| \\
&\leq \left| \nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t_n)\right) \right| \cdot \left| \frac{1}{\sqrt{A} + \varepsilon} - \frac{1}{\sqrt{B} + \varepsilon} \right| + \frac{\left| \nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t_n)\right) - \nabla_j E_n\left(\boldsymbol{\theta}^{(n)}\right) \right|}{\sqrt{B} + \varepsilon} \\
&\leq M_1 \cdot \frac{2M_1 M_2 \sqrt{p} \sum_{k=0}^{n} \rho^{n-k}(1-\rho) \left\| \tilde{\boldsymbol{\theta}}(t_k) - \boldsymbol{\theta}^{(k)} \right\|}{2R(R+\varepsilon)^2} + \frac{M_2 \sqrt{p} \left\| \tilde{\boldsymbol{\theta}}(t_n) - \boldsymbol{\theta}^{(n)} \right\|}{R + \varepsilon} \\
&= \frac{M_1^2 M_2 \sqrt{p}}{R(R+\varepsilon)^2} \sum_{k=0}^{n} \rho^{n-k}(1-\rho) \left\| \tilde{\boldsymbol{\theta}}(t_k) - \boldsymbol{\theta}^{(k)} \right\| + \frac{M_2 \sqrt{p}}{R + \varepsilon} \left\| \tilde{\boldsymbol{\theta}}(t_n) - \boldsymbol{\theta}^{(n)} \right\| \\
&\overset{\text{(a)}}{\leq} \frac{M_1^2 M_2 \sqrt{p}}{R(R+\varepsilon)^2} \sum_{k=0}^{n} \rho^{n-k}(1-\rho) d_1 e^{d_2 k h} h^2 + \frac{M_2 \sqrt{p}}{R + \varepsilon} d_1 e^{d_2 n h} h^2,
\end{aligned} \tag{17}$$

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

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

Then we can continue (17):

$$\left| \frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t_n)\right)}{\sqrt{A} + \varepsilon} - \frac{\nabla_j E_n\left(\boldsymbol{\theta}^{(n)}\right)}{\sqrt{B} + \varepsilon} \right| \leq \frac{M_2 \sqrt{p}}{R + \varepsilon} \left(\frac{M_1^2}{R(R+\varepsilon)} + 1 \right) d_1 e^{d_2 n h} h^2 \tag{18}$$

Again using $1 \leq e^{d_2 n h}$, we conclude from (14) and (18) that

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

finishing the induction step. □

B.1. RMSProp with ε outside: full-batch

In the full-batch setting $E_k \equiv E$, the terms in (13) simplify to

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

If ε is small and the iteration number n is large, (13) simplifies to

$$\dot{\tilde{\theta}}_j(t) = -\operatorname{sign} \nabla_j E(\tilde{\boldsymbol{\theta}}(t)) + h \frac{\rho}{1-\rho} \cdot \frac{\sum_{i=1}^{p} \nabla_{ij} E(\tilde{\boldsymbol{\theta}}(t)) \operatorname{sign} \nabla_i E(\tilde{\boldsymbol{\theta}}(t))}{\left| \nabla_j E(\tilde{\boldsymbol{\theta}}(t)) \right|}$$

$$= \left|\nabla_j E(\tilde{\boldsymbol{\theta}}(t))\right|^{-1}\left[-\nabla_j E(\tilde{\boldsymbol{\theta}}(t)) + h\frac{\rho}{1-\rho}\nabla_j\left\|\nabla E(\tilde{\boldsymbol{\theta}}(t))\right\|_1\right].$$

C. RMSProp with ε Inside the Square Root

Definition C.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

$$\begin{aligned}\nu_j^{(n+1)} &= \rho\nu_j^{(n)} + (1-\rho)\left(\nabla_j E_n\left(\boldsymbol{\theta}^{(n)}\right)\right)^2,\\ \theta_j^{(n+1)} &= \theta_j^{(n)} - \frac{h}{\sqrt{\nu_j^{(n+1)} + \varepsilon}}\nabla_j E_n\left(\boldsymbol{\theta}^{(n)}\right).\end{aligned} \tag{19}$$

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

$$\begin{aligned}\dot{\tilde{\theta}}_j(t) = &-\frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)}\\ &+ h\left(\frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\left(2P_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \bar{P}_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)}{2R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)^3} - \frac{\sum_{i=1}^p \nabla_{ij}E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\frac{\nabla_i E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)}}{2R_j^{(n)}(\tilde{\boldsymbol{\theta}}(t))}\right).\end{aligned} \tag{20}$$

for $t \in [t_n, t_{n+1}]$ 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

$$\begin{aligned}R_j^{(n)}(\boldsymbol{\theta}) &:= \sqrt{\sum_{k=0}^n \rho^{n-k}(1-\rho)(\nabla_j E_k(\boldsymbol{\theta}))^2 + \varepsilon},\\ P_j^{(n)}(\boldsymbol{\theta}) &:= \sum_{k=0}^n \rho^{n-k}(1-\rho)\nabla_j E_k(\boldsymbol{\theta})\sum_{i=1}^p \nabla_{ij}E_k(\boldsymbol{\theta})\sum_{l=k}^{n-1}\frac{\nabla_i E_l(\boldsymbol{\theta})}{R_i^{(l)}(\boldsymbol{\theta})},\\ \bar{P}_j^{(n)}(\boldsymbol{\theta}) &:= \sum_{k=0}^n \rho^{n-k}(1-\rho)\nabla_j E_k(\boldsymbol{\theta})\sum_{i=1}^p \nabla_{ij}E_k(\boldsymbol{\theta})\frac{\nabla_i E_n(\boldsymbol{\theta})}{R_i^{(n)}(\boldsymbol{\theta})}.\end{aligned} \tag{21}$$

Assumption C.2. For some positive constants M_1, M_2, M_3, M_4 we have

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

Theorem C.3 (RMSProp with ε inside: local error bound)**.** *Suppose Assumption C.2 holds. Then for all $n \in \{0, 1, \dots, \lfloor T/h \rfloor\}$, $j \in \{1, \dots, p\}$*

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

for a positive constant C_2 depending on ρ, where $\tilde{\boldsymbol{\theta}}(t)$ is defined in Definition C.1.

The argument is the same as for Theorem B.3.

Theorem C.4 (RMSProp with ε inside: global error bound)**.** *Suppose Assumption C.2 holds. Then there exist positive constants d_4, d_5, d_6 such that for all $n \in \{0, 1, \dots, \lfloor T/h \rfloor\}$*

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

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 C.1. The constants can be defined as

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

The argument is the same as for Theorem B.4.

C.1. RMSProp with ε Inside: Full-Batch

In the full-batch setting $E_k \equiv E$, the terms in (20) simplify to

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

If the iteration number n is large, (20) rapidly becomes

$$\dot{\tilde{\theta}}_j(t) = -\frac{1}{\sqrt{|\nabla_j E(\tilde{\boldsymbol{\theta}}(t))|^2 + \varepsilon}} \left(\nabla_j E(\tilde{\boldsymbol{\theta}}(t)) + \operatorname{correction}_j\left(\tilde{\boldsymbol{\theta}}(t)\right)\right),$$

where

$$\operatorname{correction}_j\left(\tilde{\boldsymbol{\theta}}(t)\right) := \frac{h}{2} \left\{ -\frac{2\rho}{1 - \rho} + \frac{1 + \rho}{1 - \rho} \cdot \frac{\varepsilon}{|\nabla_j E(\tilde{\boldsymbol{\theta}}(t))|^2 + \varepsilon} \right\} \nabla_j \left\|\nabla E(\tilde{\boldsymbol{\theta}}(t))\right\|_{1,\varepsilon}.$$

D. Adam with ε Outside the Square Root

Definition D.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 $\left\{\boldsymbol{\theta}^{(k)}\right\}_{k \in \mathbb{Z}_{\ge 0}}$ be defined for $n \ge 0$ by

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

or, rewriting,

$$\theta_j^{(n+1)} = \theta_j^{(n)} - h \frac{\frac{1}{1 - \beta^{n+1}} \sum_{k=0}^{n} \beta^{n-k} (1 - \beta) \nabla_j E_k\left(\boldsymbol{\theta}^{(k)}\right)}{\sqrt{\frac{1}{1 - \rho^{n+1}} \sum_{k=0}^{n} \rho^{n-k} (1 - \rho) \left(\nabla_j E_k\left(\boldsymbol{\theta}^{(k)}\right)\right)^2} + \varepsilon}. \tag{22}$$

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

$$\begin{aligned}\dot{\tilde{\theta}}_j(t) = &- \frac{M_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \varepsilon} \\ &+ h \left(\frac{M_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\left(2P_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \bar{P}_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)}{2\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \varepsilon\right)^2 R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)} - \frac{2L_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \bar{L}_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)}{2\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \varepsilon\right)} \right).\end{aligned} \tag{23}$$

for $t \in [t_n, t_{n+1}]$ 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

$$\begin{aligned}
R_j^{(n)}(\boldsymbol{\theta}) &:= \sqrt{\sum_{k=0}^{n} \rho^{n-k}(1-\rho)(\nabla_j E_k(\boldsymbol{\theta}))^2/(1-\rho^{n+1})}, \\
M_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\beta^{n+1}} \sum_{k=0}^{n} \beta^{n-k}(1-\beta)\nabla_j E_k(\boldsymbol{\theta}), \\
L_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\beta^{n+1}} \sum_{k=0}^{n} \beta^{n-k}(1-\beta) \sum_{i=1}^{p} \nabla_{ij} E_k(\boldsymbol{\theta}) \sum_{l=k}^{n-1} \frac{M_i^{(l)}(\boldsymbol{\theta})}{R_i^{(l)}(\boldsymbol{\theta}) + \varepsilon}, \\
\bar{L}_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\beta^{n+1}} \sum_{k=0}^{n} \beta^{n-k}(1-\beta) \sum_{i=1}^{p} \nabla_{ij} E_k(\boldsymbol{\theta}) \frac{M_i^{(n)}(\boldsymbol{\theta})}{R_i^{(n)}(\boldsymbol{\theta}) + \varepsilon}, \\
P_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\rho^{n+1}} \sum_{k=0}^{n} \rho^{n-k}(1-\rho)\nabla_j E_k(\boldsymbol{\theta}) \sum_{i=1}^{p} \nabla_{ij} E_k(\boldsymbol{\theta}) \sum_{l=k}^{n-1} \frac{M_i^{(l)}(\boldsymbol{\theta})}{R_i^{(l)}(\boldsymbol{\theta}) + \varepsilon}, \\
\bar{P}_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\rho^{n+1}} \sum_{k=0}^{n} \rho^{n-k}(1-\rho)\nabla_j E_k(\boldsymbol{\theta}) \sum_{i=1}^{p} \nabla_{ij} E_k(\boldsymbol{\theta}) \frac{M_i^{(n)}(\boldsymbol{\theta})}{R_i^{(n)}(\boldsymbol{\theta}) + \varepsilon}.
\end{aligned} \tag{24}$$

Assumption D.2.

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

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

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 D.1.

Theorem D.3 (Adam with ε outside: local error bound)**.** *Suppose Assumption D.2 holds. Then for all* $n \in \{0, 1, \dots, \lfloor T/h \rfloor\}$, $j \in \{1, \dots, p\}$

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

for a positive constant C_3 depending on β and ρ.

The argument is the same as for Theorem B.3.

Theorem D.4 (Adam with ε outside: global error bound)**.** *Suppose Assumption D.2 holds, and*

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

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

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

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

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

Proof. Analogously to Theorem B.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

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

By Theorem D.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 $k = 0, 1, \dots, n-1$ 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

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

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 D.3 with (22), we have

$$\Big|[\mathbf{e}_{n+1}-\mathbf{e}_n]_j\Big| \le C_3 h^3 + h\left|\frac{N'}{\sqrt{D'}+\varepsilon} - \frac{N''}{\sqrt{D''}+\varepsilon}\right|, \tag{25}$$

where to simplify notation we put

$$N' := \frac{1}{1-\beta^{n+1}}\sum_{k=0}^{n}\beta^{n-k}(1-\beta)\nabla_j E_k\Big(\boldsymbol{\theta}^{(k)}\Big),$$

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

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

$$\left|\frac{1}{\sqrt{D'}+\varepsilon} - \frac{1}{\sqrt{D''}+\varepsilon}\right| = \frac{|D'-D''|}{\Big(\sqrt{D'}+\varepsilon\Big)\Big(\sqrt{D''}+\varepsilon\Big)\Big(\sqrt{D'}+\sqrt{D''}\Big)} \leq \frac{|D'-D''|}{2R(R+\varepsilon)^2}. \tag{26}$$

But since

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

we have

$$|D'-D''| \leq \frac{2M_1M_2\sqrt{p}}{1-\rho^{n+1}} \sum_{k=0}^{n} \rho^{n-k}(1-\rho)\left\|\boldsymbol{\theta}^{(k)} - \tilde{\boldsymbol{\theta}}(t_k)\right\|. \tag{27}$$

Similarly,

$$\begin{aligned}
|N'-N''| &\leq \frac{1}{1-\beta^{n+1}} \sum_{k=0}^{n} \beta^{n-k}(1-\beta)\left|\nabla_j E_k\Big(\boldsymbol{\theta}^{(k)}\Big) - \nabla_j E_k\Big(\tilde{\boldsymbol{\theta}}(t_k)\Big)\right| \\
&\leq \frac{1}{1-\beta^{n+1}} \sum_{k=0}^{n} \beta^{n-k}(1-\beta)M_2\sqrt{p}\left\|\boldsymbol{\theta}^{(k)} - \tilde{\boldsymbol{\theta}}(t_k)\right\|.
\end{aligned} \tag{28}$$

Combining (26), (27) and (28), we get

$$\begin{aligned}
&\left|\frac{N'}{\sqrt{D'}+\varepsilon} - \frac{N''}{\sqrt{D''}+\varepsilon}\right| \leq |N'| \cdot \left|\frac{1}{\sqrt{D'}+\varepsilon} - \frac{1}{\sqrt{D''}+\varepsilon}\right| + \frac{|N'-N''|}{\sqrt{D''}+\varepsilon} \\
&\quad\leq \frac{1}{1-\beta^{n+1}} \sum_{k=0}^{n} \beta^{n-k}(1-\beta)M_1 \cdot \frac{2M_1M_2\sqrt{p}}{2R(R+\varepsilon)^2(1-\rho^{n+1})} \sum_{k=0}^{n} \rho^{n-k}(1-\rho)\left\|\boldsymbol{\theta}^{(k)} - \tilde{\boldsymbol{\theta}}(t_k)\right\| \\
&\qquad+ \frac{M_2\sqrt{p}}{(R+\varepsilon)(1-\beta^{n+1})} \sum_{k=0}^{n} \beta^{n-k}(1-\beta)\left\|\boldsymbol{\theta}^{(k)} - \tilde{\boldsymbol{\theta}}(t_k)\right\| \\
&\quad= \frac{M_1^2M_2\sqrt{p}}{R(R+\varepsilon)^2(1-\rho^{n+1})} \sum_{k=0}^{n} \rho^{n-k}(1-\rho)\left\|\boldsymbol{\theta}^{(k)} - \tilde{\boldsymbol{\theta}}(t_k)\right\| \\
&\qquad+ \frac{M_2\sqrt{p}}{(R+\varepsilon)(1-\beta^{n+1})} \sum_{k=0}^{n} \beta^{n-k}(1-\beta)\left\|\boldsymbol{\theta}^{(k)} - \tilde{\boldsymbol{\theta}}(t_k)\right\| \\
&\quad\overset{\text{(a)}}{\leq} \frac{M_1^2M_2\sqrt{p}}{R(R+\varepsilon)^2(1-\rho^{n+1})} \sum_{k=0}^{n} \rho^{n-k}(1-\rho)d_7e^{d_8kh}h^2 \\
&\qquad+ \frac{M_2\sqrt{p}}{(R+\varepsilon)(1-\beta^{n+1})} \sum_{k=0}^{n} \beta^{n-k}(1-\beta)d_7e^{d_8kh}h^2,
\end{aligned} \tag{29}$$

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

Now note that since $0 < \rho e^{-d_8 h} < \rho$, we have $\sum_{k=0}^{n} \big(\rho e^{-d_8 h}\big)^k \le \sum_{k=0}^{n} \rho^k = \big(1 - \rho^{n+1}\big)/(1-\rho)$, which is rewritten as

$$\frac{1}{1-\rho^{n+1}} \sum_{k=0}^{n} \rho^{n-k}(1-\rho) e^{d_8 k h} \le e^{d_8 n h}.$$

By the same logic,

$$\frac{1}{1-\beta^{n+1}} \sum_{k=0}^{n} \beta^{n-k}(1-\beta) e^{d_8 k h} \le e^{d_8 n h}.$$

Then we can continue (29):

$$\left| \frac{N'}{\sqrt{D'} + \varepsilon} - \frac{N''}{\sqrt{D''} + \varepsilon} \right| \le \frac{M_2 \sqrt{p}}{R + \varepsilon} \left(\frac{M_1^2}{R(R+\varepsilon)} + 1 \right) d_7 e^{d_8 n h} h^2 \tag{30}$$

Again using $1 \le e^{d_8 n h}$, we conclude from (25) and (30) that

$$\|\mathbf{e}_{n+1} - \mathbf{e}_n\| \le \underbrace{\left(C_3 + \frac{M_2\sqrt{p}}{R+\varepsilon} \left(\frac{M_1^2}{R(R+\varepsilon)} + 1 \right) d_7 \right) \sqrt{p}}_{\le d_9} e^{d_8 n h} h^3,$$

finishing the induction step. □

E. Adam with ε Inside the Square Root

Definition E.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 $\left\{\boldsymbol{\theta}^{(k)}\right\}_{k \in \mathbb{Z}_{\ge 0}}$ be defined for $n \ge 0$ by

$$\begin{aligned}
\nu_j^{(n+1)} &= \rho \nu_j^{(n)} + (1-\rho)\Big(\nabla_j E_n\Big(\boldsymbol{\theta}^{(n)}\Big)\Big)^2, \\
m_j^{(n+1)} &= \beta m_j^{(n)} + (1-\beta) \nabla_j E_n\Big(\boldsymbol{\theta}^{(n)}\Big), \\
\theta_j^{(n+1)} &= \theta_j^{(n)} - h \frac{m_j^{(n+1)} / \big(1 - \beta^{n+1}\big)}{\sqrt{\nu_j^{(n+1)} / (1 - \rho^{n+1}) + \varepsilon}}.
\end{aligned} \tag{31}$$

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

$$\begin{aligned}
\dot{\tilde{\theta}}_j(t) = &- \frac{M_j^{(n)}\Big(\tilde{\boldsymbol{\theta}}(t)\Big)}{R_j^{(n)}\Big(\tilde{\boldsymbol{\theta}}(t)\Big)} \\
&+ h \left(\frac{M_j^{(n)}\Big(\tilde{\boldsymbol{\theta}}(t)\Big)\Big(2P_j^{(n)}\Big(\tilde{\boldsymbol{\theta}}(t)\Big) + \bar{P}_j^{(n)}\Big(\tilde{\boldsymbol{\theta}}(t)\Big)\Big)}{2R_j^{(n)}\Big(\tilde{\boldsymbol{\theta}}(t)\Big)^3} - \frac{2L_j^{(n)}\Big(\tilde{\boldsymbol{\theta}}(t)\Big) + \bar{L}_j^{(n)}\Big(\tilde{\boldsymbol{\theta}}(t)\Big)}{2R_j^{(n)}\Big(\tilde{\boldsymbol{\theta}}(t)\Big)} \right)
\end{aligned} \tag{32}$$

for $t \in [t_n, t_{n+1}]$ 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

$$\begin{aligned}
R_j^{(n)}(\boldsymbol{\theta}) &:= \sqrt{\sum_{k=0}^{n} \rho^{n-k}(1-\rho)(\nabla_j E_k(\boldsymbol{\theta}))^2/(1-\rho^{n+1}) + \varepsilon}, \\
M_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\beta^{n+1}} \sum_{k=0}^{n} \beta^{n-k}(1-\beta)\nabla_j E_k(\boldsymbol{\theta}), \\
L_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\beta^{n+1}} \sum_{k=0}^{n} \beta^{n-k}(1-\beta) \sum_{i=1}^{p} \nabla_{ij} E_k(\boldsymbol{\theta}) \sum_{l=k}^{n-1} \frac{M_i^{(l)}(\boldsymbol{\theta})}{R_i^{(l)}(\boldsymbol{\theta})}, \\
\bar{L}_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\beta^{n+1}} \sum_{k=0}^{n} \beta^{n-k}(1-\beta) \sum_{i=1}^{p} \nabla_{ij} E_k(\boldsymbol{\theta}) \frac{M_i^{(n)}(\boldsymbol{\theta})}{R_i^{(n)}(\boldsymbol{\theta})}, \\
P_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\rho^{n+1}} \sum_{k=0}^{n} \rho^{n-k}(1-\rho)\nabla_j E_k(\boldsymbol{\theta}) \sum_{i=1}^{p} \nabla_{ij} E_k(\boldsymbol{\theta}) \sum_{l=k}^{n-1} \frac{M_i^{(l)}(\boldsymbol{\theta})}{R_i^{(l)}(\boldsymbol{\theta})}, \\
\bar{P}_j^{(n)}(\boldsymbol{\theta}) &:= \frac{1}{1-\rho^{n+1}} \sum_{k=0}^{n} \rho^{n-k}(1-\rho)\nabla_j E_k(\boldsymbol{\theta}) \sum_{i=1}^{p} \nabla_{ij} E_k(\boldsymbol{\theta}) \frac{M_i^{(n)}(\boldsymbol{\theta})}{R_i^{(n)}(\boldsymbol{\theta})}.
\end{aligned} \tag{33}$$

Assumption E.2. For some positive constants M_1, M_2, M_3, M_4 we have

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

Theorem E.3 (Adam with ε inside: local error bound)**.** *Suppose Assumption E.2 holds. Then for all $n \in \{0, 1, \ldots, \lfloor T/h \rfloor\}$, $j \in \{1, \ldots, p\}$*

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

for a positive constant C_4 depending on β and ρ.

The argument is the same as for Theorem B.3.

Theorem E.4 (Adam with ε inside: global error bound)**.** *Suppose Assumption E.2 holds for $\left\{\boldsymbol{\theta}^{(k)}\right\}_{k \in \mathbb{Z}_{\ge 0}}$ defined in Definition E.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

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

The argument is the same as for Theorem D.4.

F. Bounding the Derivatives of the ODE Solution

Our first goal is to argue that the first derivative of $t \mapsto \tilde{\theta}_j(t)$ is uniformly bounded in absolute value. To achieve this, we just need to bound all the terms on the right-hand side of the ODE (13).

Lemma F.1. *Suppose Assumption B.2 holds. Then for all* $n \in \{0, 1, \ldots, \lfloor T/h \rfloor\}$

$$\sup_{\boldsymbol{\theta}} \left| P_j^{(n)}(\boldsymbol{\theta}) \right| \leq C_5, \tag{34}$$

$$\sup_{\boldsymbol{\theta}} \left| \bar{P}_j^{(n)}(\boldsymbol{\theta}) \right| \leq C_6, \tag{35}$$

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 F.1. Both bounds are straightforward:

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

and

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

concluding the proof of Lemma F.1. □

Lemma F.2. *Suppose Assumption B.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 $h \leq T$, (34), (35) and the definition of $\tilde{\boldsymbol{\theta}}(t)$ given in (13). □

Our next goal is to argue that the *second* derivative of $t \mapsto \tilde{\theta}_j(t)$ is bounded in absolute value. For this, we need to bound the first derivatives of all the three additive terms on the right-hand side of (13).

Lemma F.3. *Suppose Assumption B.2 holds. Then for all* $n, k \in \{0, 1, \ldots, \lfloor T/h \rfloor\}$, $j \in \{1, \ldots, p\}$ *we have*

$$\sup_{t \in [0, T]} \left| \left(\nabla_j E_n \left(\tilde{\boldsymbol{\theta}}(t) \right) \right)^{\cdot} \right| \leq C_7, \tag{36}$$

$$\sup_{t \in [t_n, t_{n+1}]} \left| \sum_{i=1}^{p} \nabla_{ij} E_k \left(\tilde{\boldsymbol{\theta}}(t) \right) \left[\dot{\tilde{\theta}}_i(t) + \frac{\nabla_i E_n \left(\tilde{\boldsymbol{\theta}}(t) \right)}{R_i^{(n)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon} \right] \right| \leq C_8 h, \tag{37}$$

$$\sup_{t \in [0, T]} \left| \sum_{i=1}^{p} \nabla_{ij} E_k \left(\tilde{\boldsymbol{\theta}}(t) \right) \sum_{l=k}^{n-1} \frac{\nabla_i E_l \left(\tilde{\boldsymbol{\theta}}(t) \right)}{R_i^{(l)} \left(\tilde{\boldsymbol{\theta}}(t) \right) + \varepsilon} \right| \leq (n - k) C_9 \quad \textit{for } k < n, \tag{38}$$

$$\sup_{t\in[0,T]}\left|\left(P_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)^{\cdot}\right| \le C_{10} + C_{14}, \tag{39}$$

$$\sup_{t\in[0,T]}\left|\left(\bar{P}_j^{(n)}(\tilde{\boldsymbol{\theta}}(t))\right)^{\cdot}\right| \le C_{15}, \tag{40}$$

$$\sup_{t\in[0,T]}\left|\left(\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\frac{\nabla_i E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon}\right)^{\cdot}\right| \le C_{13}, \tag{41}$$

$$\sup_{t\in[0,T]}\left|\left(\frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\left(2P_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\bar{P}_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)}{2\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^2 R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)}\right)^{\cdot}\right| \le C_{17}, \tag{42}$$

$$\sup_{t\in[0,T]}\left|\left(\frac{\sum_{i=1}^{p}\nabla_{ij}E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\frac{\nabla_i E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon}}{2\left(R_j^{(n)}(\tilde{\boldsymbol{\theta}}(t))+\varepsilon\right)}\right)^{\cdot}\right| \le C_{18}, \tag{43}$$

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:*

$$\begin{aligned}
C_7 &:= pM_2D_1,\\
C_8 &:= pM_2\left[\frac{M_1(2C_5+C_6)}{2(R+\varepsilon)^2R}+\frac{pM_1M_2}{2(R+\varepsilon)^2}\right],\\
C_9 &:= p\frac{M_1M_2}{R+\varepsilon},\\
C_{10} &:= D_1p^2\frac{M_1M_2^2}{R+\varepsilon}\cdot\frac{\rho}{1-\rho},\\
C_{11} &:= \frac{D_1pM_1M_2}{R},\\
C_{12} &:= D_1p^2\frac{M_1M_3}{R+\varepsilon},\\
C_{13} &:= C_{12}+pM_2\left(\frac{D_1pM_2}{R+\varepsilon}+\frac{M_1}{(R+\varepsilon)^2}C_{11}\right)\\
&= \frac{D_1p^2}{R+\varepsilon}\left(M_1M_3+M_2^2+\frac{M_1^2M_2^2}{(R+\varepsilon)R}\right),\\
C_{14} &:= M_1C_{13}\frac{\rho}{1-\rho},\\
C_{15} &:= \frac{D_1p^2M_1M_2^2}{R+\varepsilon}+\frac{D_1p^2M_1^2M_3}{R+\varepsilon}+\frac{D_1p^2M_1M_2^2}{R+\varepsilon}+\frac{pM_1^2M_2C_{11}}{(R+\varepsilon)^2},\\
C_{16} &:= \frac{2C_{11}}{R(R+\varepsilon)^3}+\frac{C_{11}}{(R+\varepsilon)^4},\\
C_{17} &:= \frac{D_1pM_2\cdot(2C_5+C_6)}{2(R+\varepsilon)^2R}+\frac{M_1(2(C_{10}+C_{14})+C_{15})}{2(R+\varepsilon)^2R}+\frac{M_1(2C_5+C_6)C_{16}}{2},\\
C_{18} &:= \frac{1}{2(R+\varepsilon)}\left(\frac{p^2D_1M_1M_3}{R+\varepsilon}+\frac{p^2D_1M_2^2}{R+\varepsilon}+\frac{pM_1M_2C_{11}}{(R+\varepsilon)^2}\right)+\frac{1}{2}\cdot\frac{pM_1M_2}{R+\varepsilon}\cdot\frac{C_{11}}{(R+\varepsilon)^2}.
\end{aligned}$$

Proof of Lemma F.3. We prove the inequalities one by one.

The bound (36) is straightforward:

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

The inequality (37) follows immediately from the fact that by (13) we have for $t \in [t_n, t_{n+1}]$

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

The bound (38) follows from the assumptions immediately.

We will prove (39) by bounding the two additive terms on the right-hand side of the equality

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t} P_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) \\
&\quad = \sum_{k=0}^{n} \rho^{n-k}(1-\rho) \sum_{u=1}^{p} \nabla_{ju} E_k\left(\tilde{\boldsymbol{\theta}}(t)\right) \dot{\tilde{\theta}}_u(t) \sum_{i=1}^{p} \nabla_{ij} E_k\left(\tilde{\boldsymbol{\theta}}(t)\right) \sum_{l=k}^{n-1} \frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \varepsilon} \\
&\qquad + \sum_{k=0}^{n} \rho^{n-k}(1-\rho) \nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t)\right) \sum_{i=1}^{p} \frac{\mathrm{d}}{\mathrm{d}t} \left\{ \nabla_{ij} E_k\left(\tilde{\boldsymbol{\theta}}(t)\right) \sum_{l=k}^{n-1} \frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \varepsilon} \right\}.
\end{aligned} \tag{44}$$

It is easily shown that the first term in (44) is bounded in absolute value by C_{10}:

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

For the proof of (39), it is left to show that the second term in (44) 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

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

By the Cauchy-Schwarz inequality applied twice,

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

$$\leq \sqrt{\sum_{i=1}^{p}\sum_{s=1}^{p}\left(\nabla_{ijs}E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\right)^2}\sqrt{\sum_{u=1}^{p}\dot{\tilde{\theta}}_u(t)^2}\sqrt{\sum_{i=1}^{p}\left|\sum_{l=k}^{n-1}\frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon}\right|^2}$$

$$\leq M_3 p\cdot D_1\sqrt{p}\cdot\sqrt{\sum_{i=1}^{p}\left|\sum_{l=k}^{n-1}\frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon}\right|^2}\leq (n-k)C_{12}.$$

Next, for any n and j

$$\begin{aligned}\left|\frac{\mathrm{d}}{\mathrm{d}t}R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right| &= \frac{1}{R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)}\left|\sum_{k=0}^{n}\rho^{n-k}(1-\rho)\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\sum_{i=1}^{p}\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\dot{\tilde{\theta}}_i(t)\right| \\ &\leq \frac{1}{R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)}D_1 p M_1 M_2\sum_{k=0}^{n}\rho^{n-k}(1-\rho)\leq C_{11}.\end{aligned} \tag{45}$$

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

$$\left|\sum_{i=1}^{p}\frac{\mathrm{d}}{\mathrm{d}t}\left\{\nabla_{ij}E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\sum_{l=k}^{n-1}\frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon}\right\}\right|\leq (n-k)C_{13}. \tag{46}$$

This gives a bound on the second term in (44):

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

concluding the proof of (39).

We will prove (40) by bounding the four terms in the expression

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

where

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

$$\begin{aligned}
&\mathrm{Term2}\\
&:= \sum_{k=0}^{n} \rho^{n-k}(1-\rho)\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t)\right) \sum_{i=1}^{p} \frac{\mathrm{d}}{\mathrm{d}t}\left\{\nabla_{ij} E_k\left(\tilde{\boldsymbol{\theta}}(t)\right)\right\} \frac{\nabla_i E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon},\\
&\mathrm{Term3}\\
&:= \sum_{k=0}^{n} \rho^{n-k}(1-\rho)\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t)\right) \sum_{i=1}^{p} \nabla_{ij} E_k\left(\tilde{\boldsymbol{\theta}}(t)\right) \frac{\frac{\mathrm{d}}{\mathrm{d}t}\left\{\nabla_i E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\right\}}{R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon},\\
&\mathrm{Term4}\\
&:= -\sum_{k=0}^{n} \rho^{n-k}(1-\rho)\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t)\right) \sum_{i=1}^{p} \nabla_{ij} E_k\left(\tilde{\boldsymbol{\theta}}(t)\right) \frac{\nabla_i E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)\frac{\mathrm{d}}{\mathrm{d}t}R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)}{\left(R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^2}.
\end{aligned}$$

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

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

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

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

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

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

To bound Term4, use (45), giving

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

The proof of (40) is finished.

The inequality (41) is already proven in (46).

To prove (42), note that the bound (45) gives

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}\left\{\frac{1}{R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)}\right\}\right| = \frac{\left|\frac{\mathrm{d}}{\mathrm{d}t}R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right|}{R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)^2} \leq \frac{C_{11}}{R^2}, \tag{47}$$

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}\left\{\frac{1}{R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon}\right\}\right| = \frac{\left|\frac{\mathrm{d}}{\mathrm{d}t}R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right|}{\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^2} \leq \frac{C_{11}}{(R+\varepsilon)^2}, \tag{48}$$

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}\left\{\frac{1}{\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^2}\right\}\right| = \frac{2\left|\frac{\mathrm{d}}{\mathrm{d}t}R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)\right|}{\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon\right)^3} \leq \frac{2C_{11}}{(R+\varepsilon)^3}. \tag{49}$$

Combining two bounds above, we have

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

We are ready to conclude

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

It is left to prove (43). Since

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

and, as we have already seen in the argument for (40),

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

we are ready to bound

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

The proof of Lemma F.3 is concluded. □

Lemma F.4. *Suppose Assumption B.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 (13), $h \le T$ and that the first derivatives of all three terms in (13) are bounded by Lemma F.3. □

Finally, we need to argue that the *third* derivative of $t \mapsto \tilde{\theta}_j(t)$ is bounded in absolute value. To achieve this, we need to bound the second derivatives of the terms on the right-hand side of (13).

Lemma F.5. *Suppose Assumption B.2 holds. Then for all $n, k \in \{0, 1, \dots, \lfloor T/h \rfloor\}$, $j \in \{1, \dots, p\}$*

$$\sup_{t \in [0,T]} \left| \left(\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t)\right) \right)^{\cdot\cdot} \right| \leq C_{19}, \tag{50}$$

$$\sup_{t \in [0,T]} \left| \left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) \right)^{\cdot\cdot} \right| \leq C_{20}, \tag{51}$$

$$\sup_{t \in [0,T]} \left| \left(\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \varepsilon \right)^{-2} \right)^{\cdot\cdot} \right| \leq C_{21}, \tag{52}$$

$$\sup_{t \in [0,T]} \left| \left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)^{-1} \right)^{\cdot\cdot} \right| \leq C_{22}, \tag{53}$$

$$\sup_{t \in [0,T]} \left| \left(\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \varepsilon \right)^{-2} R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)^{-1} \right)^{\cdot\cdot} \right| \leq C_{23}, \tag{54}$$

$$\sup_{t \in [0,T]} \left| \left(\sum_{i=1}^{p} \nabla_{ij} E_k\left(\tilde{\boldsymbol{\theta}}(t)\right) \sum_{l=k}^{n-1} \frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t)\right) + \varepsilon} \right)^{\cdot\cdot} \right| \leq (n-k) C_{24} \quad \textit{for } k < n, \tag{55}$$

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

$$\begin{aligned}
C_{19} &:= p^2 M_3 D_1^2 + p M_2 D_2, \\
C_{20} &:= \frac{C_{11}}{R^2} p M_1 M_2 D_1 + \frac{1}{R} p^2 M_2^2 D_1^2 + \frac{1}{R} p^2 M_1 M_3 D_1^2 + \frac{1}{R} p M_1 M_2 D_2, \\
C_{21} &:= \frac{6 C_{11}^2}{(R+\varepsilon)^4} + \frac{2 C_{20}}{(R+\varepsilon)^3}, \\
C_{22} &:= \frac{2 C_{11}^2}{R^3} + \frac{C_{20}}{R^2}, \\
C_{23} &:= \frac{C_{21}}{R} + \frac{4 C_{11}^2}{R^2 (R+\varepsilon)^3} + \frac{C_{22}}{(R+\varepsilon)^2}, \\
C_{24} &:= p \Bigg[\frac{2 C_{11} \left(D_1 M_2^2 p + D_1 M_1 M_3 p \right)}{(R+\varepsilon)^2} + M_1 M_2 \left(\frac{2 C_{11}^2}{(R+\varepsilon)^3} + \frac{C_{20}}{(R+\varepsilon)^2} \right) \\
&\qquad + \frac{2 D_1^2 M_2 M_3 p^2 + M_2 \left(D_1^2 M_3 p^2 + D_2 M_2 p \right) + M_1 \left(D_1^2 M_4 p^2 + D_2 M_3 p \right)}{R + \varepsilon} \Bigg].
\end{aligned}$$

Proof of Lemma F.5. We prove the inequalities one by one.

The proof of (50) is straightforward:

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

To prove (51), note that

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

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

giving by (47)

$$
\begin{aligned}
\left|\Big(R_j^{(n)}\Big(\tilde{\boldsymbol{\theta}}(t)\Big)\Big)^{\cdot\cdot}\right| &\leq \frac{C_{11}}{R^2} p M_1 M_2 D_1 \sum_{k=0}^{n} \rho^{n-k}(1-\rho) + \frac{1}{R} p^2 M_2^2 D_1^2 \sum_{k=0}^{n} \rho^{n-k}(1-\rho) \\
&+ \frac{1}{R} p^2 M_1 M_3 D_1^2 \sum_{k=0}^{n} \rho^{n-k}(1-\rho) + \frac{1}{R} p M_1 M_2 D_2 \sum_{k=0} \rho^{n-k}(1-\rho) \\
&\leq C_{20}.
\end{aligned}
$$

To prove (52), note that

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

giving by (45) and (51)

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

The bound (53) follows from (45), (51) and

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

To justify (54), put temporarily $a := \Big(R_j^{(n)}\Big(\tilde{\boldsymbol{\theta}}(t)\Big) + \varepsilon\Big)^{-2}$, $b := R_j^{(n)}\Big(\tilde{\boldsymbol{\theta}}(t)\Big)^{-1}$ and use

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

combined with

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

To justify (55), put temporarily

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

and use

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

from which (55) follows.

The proof of Lemma F.5 is concluded. □

Lemma F.6. *Suppose Assumption B.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 (38), (46) and (55)

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

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 (52) 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 (13). The boundedness of the second derivatives of the other two terms is shown analogously. By (13) and since $h \le T$, this means

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

for some positive constant D_3. □

G. Proof of Theorem B.3

Our next objective is proving and identifying the constant in the equality

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

$$= \frac{1}{R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right) + \varepsilon} - h\frac{P_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)}{\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right) + \varepsilon\right)^2 R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)} + O(h^2).$$

We will make some preparations and achieve this objective in Lemma G.5. Then we will conclude the proof of Theorem B.3.

Lemma G.1. *Suppose Assumption B.2 holds. Then for all* $n \in \{0, 1, \dots, \lfloor T/h \rfloor\}$, $k \in \{0, 1, \dots, n-1\}$, $j \in \{1, \dots, p\}$ *we have*

$$\left|\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_k)\right) - \nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_n)\right)\right| \le C_7(n-k)h \tag{56}$$

Proof. (56) follows from the mean value theorem applied $n-k$ times. □

Lemma G.2. *In the setting of Lemma G.1, for any* $l \in \{k, k+1, \dots, n-1\}$ *we have*

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

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

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

Combining this with (37) gives

$$\begin{aligned}&\left|\nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_l)\right) - \nabla_j E_k\left(\tilde{\boldsymbol{\theta}}(t_{l+1})\right) - h\sum_{i=1}^p \nabla_{ij} E_k\left(\tilde{\boldsymbol{\theta}}(t_{l+1})\right)\frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t_{l+1})\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t_{l+1})\right) + \varepsilon}\right| \\ &\quad \le (C_{19}/2 + C_8)h^2.\end{aligned} \tag{57}$$

Now applying the mean-value theorem $n-l-1$ times, we have by (46)

$$\begin{aligned}&\left|\sum_{i=1}^p \nabla_{ij} E_k\left(\tilde{\boldsymbol{\theta}}(t_{l+1})\right)\frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t_{l+1})\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t_{l+1})\right) + \varepsilon} - \sum_{i=1}^p \nabla_{ij} E_k\left(\tilde{\boldsymbol{\theta}}(t_{l+2})\right)\frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t_{l+2})\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t_{l+2})\right) + \varepsilon}\right| \le C_{13}h, \\ &\dots \\ &\left|\sum_{i=1}^p \nabla_{ij} E_l\left(\tilde{\boldsymbol{\theta}}(t_{n-1})\right)\frac{\nabla_i E_k\left(\tilde{\boldsymbol{\theta}}(t_{n-1})\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t_{n-1})\right) + \varepsilon} - \sum_{i=1}^p \nabla_{ij} E_k\left(\tilde{\boldsymbol{\theta}}(t_n)\right)\frac{\nabla_i E_l\left(\tilde{\boldsymbol{\theta}}(t_n)\right)}{R_i^{(l)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right) + \varepsilon}\right| \le C_{13}h,\end{aligned}$$

and in particular

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

Combining this with (57), we conclude the proof of Lemma G.2. □

Lemma G.3. *In the setting of Lemma G.1,*

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

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

Note that

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

where (a) is by Lemma G.2. □

Lemma G.4. *Suppose Assumption B.2 holds. Then for all $n \in \{0, 1, \dots, \lfloor T/h \rfloor\}$, $j \in \{1, \dots, p\}$*

$$\left|\sum_{k=0}^{n} \rho^{n-k}(1-\rho)\Big(\nabla_j E_k\Big(\tilde{\boldsymbol{\theta}}(t_k)\Big)\Big)^2 - R_j^{(n)}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big)^2\right| \leq C_{25} h \tag{58}$$

and

$$\left|\sum_{k=0}^{n} \rho^{n-k}(1-\rho)\Big(\nabla_j E_k\Big(\tilde{\boldsymbol{\theta}}(t_k)\Big)\Big)^2 - R_j^{(n)}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big)^2 - 2hP_j^{(n)}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big)\right| \leq C_{26} h^2 \tag{59}$$

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

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

Proof. Note that

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

where (a) is by (56). Using the triangle inequality, we can conclude

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

(58) is proven.

We continue by showing

$$\begin{aligned}&\Bigg|\Big(\nabla_j E_k\big(\tilde{\boldsymbol{\theta}}(t_k)\big)\Big)^2 - \Big(\nabla_j E_k\big(\tilde{\boldsymbol{\theta}}(t_n)\big)\Big)^2\\&\qquad -2\nabla_j E_k\big(\tilde{\boldsymbol{\theta}}(t_n)\big)h\sum_{i=1}^{p}\nabla_{ij}E_k\big(\tilde{\boldsymbol{\theta}}(t_n)\big)\sum_{l=k}^{n-1}\frac{\nabla_i E_l\big(\tilde{\boldsymbol{\theta}}(t_n)\big)}{R_i^{(l)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)+\varepsilon}\Bigg|\\&\quad\le 2M_1\bigg((n-k)(C_{19}/2+C_8)+\frac{(n-k)(n-k-1)}{2}C_{13}\bigg)h^2\\&\qquad+2(n-k)C_9\bigg((n-k)(C_{19}/2+C_8)+\frac{(n-k)(n-k-1)}{2}C_{13}\bigg)h^3\\&\qquad+\bigg((n-k)(C_{19}/2+C_8)+\frac{(n-k)(n-k-1)}{2}C_{13}\bigg)^2h^4.\end{aligned}\tag{60}$$

To prove this, use

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

with

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

and bounding

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

where (a) is by Lemma G.3. (60) is proven.

We turn to the proof of (59). By (60) and the triangle inequality

$$\begin{aligned}&\left|\sum_{k=0}^{n}\rho^{n-k}(1-\rho)\Big(\nabla_j E_k\big(\tilde{\boldsymbol{\theta}}(t_k)\big)\Big)^2 - R_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)^2 - 2hP_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)\right|\\&\quad\le(1-\rho)\sum_{k=0}^{n}\rho^{n-k}\big(\mathrm{Poly}_1(n-k)h^2+\mathrm{Poly}_2(n-k)h^3+\mathrm{Poly}_3(n-k)h^4\big)\\&\quad=(1-\rho)\sum_{k=0}^{n}\rho^{k}\big(\mathrm{Poly}_1(k)h^2+\mathrm{Poly}_2(k)h^3+\mathrm{Poly}_3(k)h^4\big),\end{aligned}$$

where

$$\mathrm{Poly}_1(k):=2M_1\bigg(k(C_{19}/2+C_8)+\frac{k(k-1)}{2}C_{13}\bigg)=M_1C_{13}k^2+M_1(C_{19}+2C_8-C_{13})k,$$

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

It is left to combine this with

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

This gives

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

where in (a) we used that $h<1$. (59) is proven. □

Lemma G.5. *Suppose Assumption B.2 holds. Then*

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

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

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

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

By the triangle inequality,

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

Apply this with

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

and use bounds

$$|a-b| \le 2M_1C_7\frac{\rho}{1-\rho}h, \quad |a-b-c| \le C_{26}(\rho)h^2$$

by Lemma G.4. □

We are finally ready to prove Theorem B.3.

Proof of Theorem B.3. By (42) and (43), 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 (13), this means

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

Combining this with

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

by Taylor expansion, we get

$$\begin{aligned}&\left|\tilde{\theta}_j(t_{n+1}) - \tilde{\theta}_j(t_n) - \dot{\tilde{\theta}}_j(t_n^+)h + \frac{h^2}{2}\cdot\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon}\right)\Bigg|_{t=t_n^+}\right| \\ &\quad\leq \left(\frac{D_3}{6}+\frac{C_{27}}{2}\right)h^3.\end{aligned} \tag{61}$$

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

$$C_{28} := \frac{M_1(2C_5+C_6)}{2(R+\varepsilon)^2R} + \frac{pM_1M_2}{2(R+\varepsilon)^2}$$

by (13), and calculating the derivative, it is easy to show

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t)\right)}{R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t)\right)+\varepsilon}\right)\Bigg|_{t=t_n^+} - \mathrm{FrDer}\right| \leq C_{29}h \tag{62}$$

for a positive constant C_{29}, where

$$\begin{aligned}\mathrm{FrDer} &:= \frac{\mathrm{FrDerNum}}{\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)+\varepsilon\right)^2 R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)}\\ \mathrm{FrDerNum} &:= \nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t_n)\right)\bar{P}_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)\\ &\quad - \left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)+\varepsilon\right)R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)\sum_{i=1}^p \nabla_{ij}E_n\left(\tilde{\boldsymbol{\theta}}(t_n)\right)\frac{\nabla_i E_n\left(\tilde{\boldsymbol{\theta}}(t_n)\right)}{R_i^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)+\varepsilon},\\ C_{29} &:= \left\{\frac{pM_2}{R+\varepsilon}+\frac{M_1^2M_2p}{(R+\varepsilon)^2R}\right\}C_{28}.\end{aligned}$$

From (61) and (62), by the triangle inequality

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

which, using (13), is rewritten as

$$\begin{aligned}&\left|\tilde{\theta}_j(t_{n+1}) - \tilde{\theta}_j(t_n) + h\frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t_n)\right)}{R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)+\varepsilon} - h^2\frac{\nabla_j E_n\left(\tilde{\boldsymbol{\theta}}(t_n)\right)P_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)}{\left(R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)+\varepsilon\right)^2 R_j^{(n)}\left(\tilde{\boldsymbol{\theta}}(t_n)\right)}\right|\\ &\quad\leq \left(\frac{D_3}{6}+\frac{C_{27}+C_{29}}{2}\right)h^3.\end{aligned}$$

It is left to combine this with Lemma G.5, giving the assertion of the theorem with

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

□

H. Numerical Experiments

H.1. Models

We use small modifications of Resnet-50 and Resnet-101 implementations in the `torchvision` library for training on CIFAR-10 and CIFAR-100. 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 He et al. (2016) (see Table 1 there) except downsampling is performed by the middle convolution of each bottleneck block, as in version 1.5[3]. After `conv5` there is global average pooling and 10 or 100-way fully connected layer (for CIFAR-10 and CIFAR-100 respectively).

The MLP that we use for showing the closeness of trajectories in Figure 3 consists of two fully connected layers, each with 32 units and GeLU activation, followed by a fully-connected layer with 10 units.

In Figure 3, the curves called "first order" plot $\left\|\boldsymbol{\theta}^{(n)} - \tilde{\boldsymbol{\theta}}^{(n)}\right\|_\infty$ and the curves called "second order" plot $\left\|\boldsymbol{\theta}^{(n)} - \tilde{\tilde{\boldsymbol{\theta}}}^{(n)}\right\|_\infty$, where $\boldsymbol{\theta}^{(n)}$ is the Adam iteration defined in Definition 1.1 and

$$\begin{aligned} \tilde{\tilde{\theta}}_j^{(n+1)} &= \tilde{\tilde{\theta}}_j^{(n)} - hA_j^{(n)}\left(\tilde{\tilde{\boldsymbol{\theta}}}^{(n)}\right) + h^2 B_j^{(n)}\left(\tilde{\tilde{\boldsymbol{\theta}}}^{(n)}\right), \\ \tilde{\theta}_j^{(n+1)} &= \tilde{\theta}_j^{(n)} - hA_j^{(n)}\left(\tilde{\boldsymbol{\theta}}^{(n)}\right) \end{aligned} \tag{63}$$

for $A_j^{(n)}(\cdot)$ and $B_j^{(n)}(\cdot)$ as defined in Section 3, with the same initial point $\boldsymbol{\theta}^{(0)} = \tilde{\boldsymbol{\theta}}^{(0)} = \tilde{\tilde{\boldsymbol{\theta}}}^{(0)}$.

H.2. Data Augmentation

We subtract the per-pixel mean and divide by standard deviation, and we use the data augmentation scheme from Lee et al. (2015), following He et al. (2016), section 4.2. During each pass over the training dataset, each 32 × 32 initial image is padded evenly with zeros so that it becomes 40×40, then random crop is applied so that the picture becomes 32 × 32 again, and random (probability 0.5) horizontal (left to right) flip is used.

H.3. Experiment Details

In experiments whose results are reported in Figures 4 and 5 of the main paper, we train for a few thousand epochs and stop training when the train accuracy is near-perfect (Figure 11) and the testing accuracy does not significantly improve (Figure 12). Therefore, the maximal test accuracies are the final ones reached, and the maximal perturbed one-norms, after excluding the initial fall at the beginning of training, are at peaks of the "hills" on the norm curves (Figure 12).

Since the full dataset does not fit into GPU memory, we divide it into 100 "ghost batches" and accumulate the gradients before doing one optimization step. This means that we use ghost batch normalization (Hoffer et al., 2017) as opposed to full-dataset batch normalization, similarly to Cohen et al. (2021).

H.4. Additional Evidence

We provide evidence that the results in Figures 4 and 5 are robust to the change of architectures. In Figures 7 and 8, we show that the pictures are similar for a simple CNN created by the following code:

```
layers = [
    # First block
    nn.Conv2d(in_channels=3, out_channels=32, kernel_size=3, padding='same'),
    nn.ReLU(),
    nn.Conv2d(in_channels=32, out_channels=32, kernel_size=3, padding='same'),
    nn.ReLU(),
    nn.MaxPool2d(kernel_size=2, stride=2),

    # Second block
```

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

```
    nn.Conv2d(in_channels=32, out_channels=64, kernel_size=3, padding='same'),
    nn.ReLU(),
    nn.Conv2d(in_channels=64, out_channels=64, kernel_size=3, padding='same'),
    nn.ReLU(),
    nn.MaxPool2d(kernel_size=2, stride=2),

    # Third block
    nn.Conv2d(in_channels=64, out_channels=128, kernel_size=3, padding='same'),
    nn.ReLU(),
    nn.Conv2d(in_channels=128, out_channels=128, kernel_size=3, padding='same'),
    nn.ReLU(),
    nn.MaxPool2d(kernel_size=2, stride=2),

    # Flatten and Dense layers
    nn.Flatten(),
    nn.Linear(in_features=128 * 4 * 4, out_features=512),
    nn.ReLU(),
    nn.Linear(in_features=512, out_features=num_classes),
]
return nn.Sequential(*layers)
```

In Figures 9 and 10, we show that the same conclusions can be made for a Vision Transformer (Dosovitskiy et al., 2020; Beyer et al., 2022). In these experiments, we use the SimpleViT architecture from the vit-pytorch library with 4×4 patches, 6 transformer blocks with 16 heads, embedding size 512 and MLP dimension of 1024 (following Andriushchenko et al. (2023)).

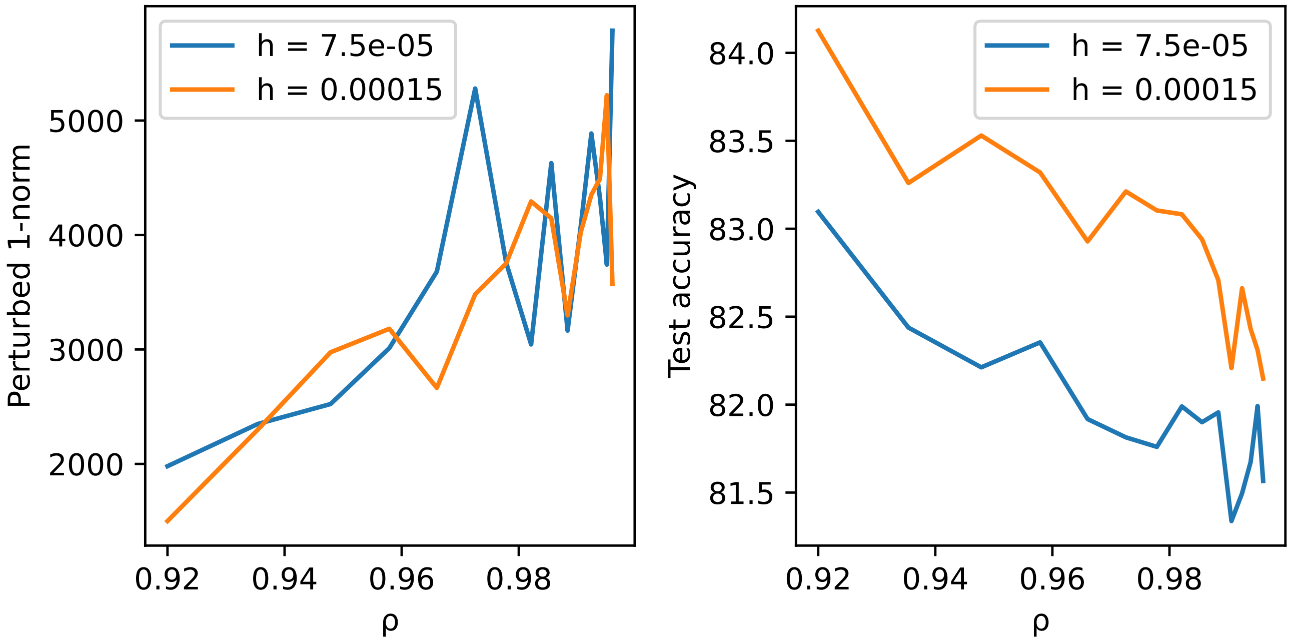

Figure 7. A simple CNN trained on CIFAR-10 with full-batch Adam, $\beta = 0.99$, $\varepsilon = 10^{-8}$. As ρ increases, the perturbed one-norm rises and the test accuracy falls. Both metrics are calculated as in Figures 4 and 5 of the main paper. All results are averaged across five runs with different initialization seeds.

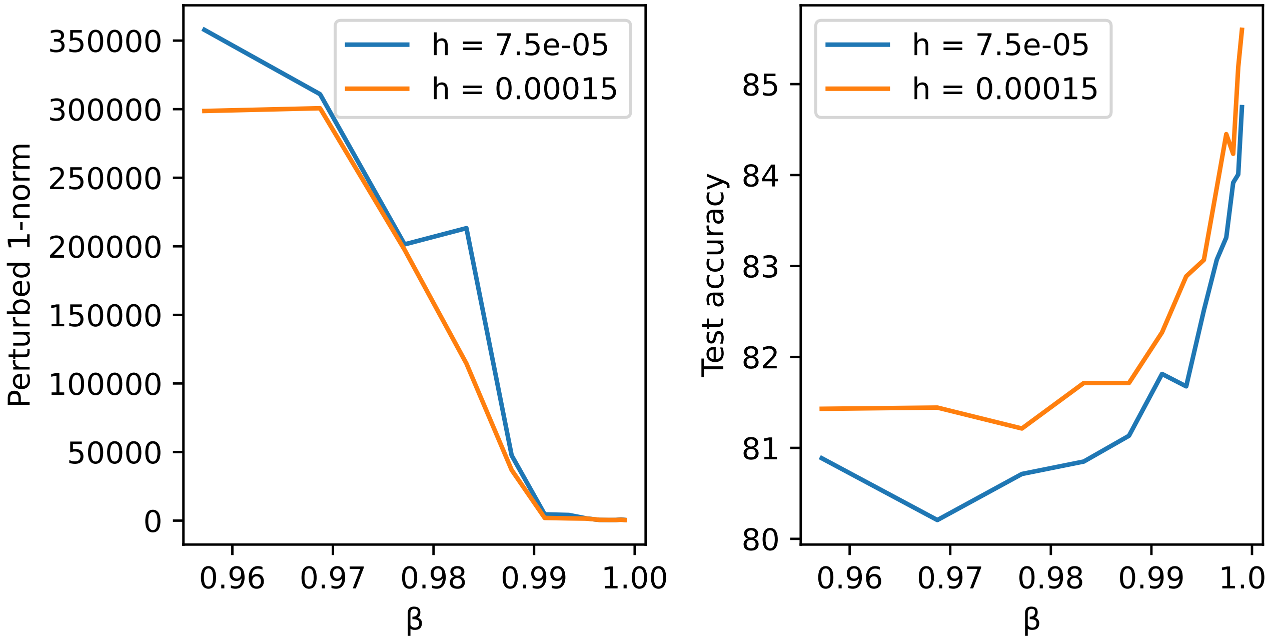

Figure 8. A simple CNN trained on CIFAR-10 with full-batch Adam, $\rho = 0.999$, $\varepsilon = 10^{-8}$. The perturbed one-norm falls as β increases, and the test accuracy rises. Both metrics are calculated as in Figures 4 and 5 of the main paper. All results are averaged across three runs with different initialization seeds.

I. Adam with ε Inside the Square Root: Informal Derivation

Our goal is to find such a trajectory $\tilde{\boldsymbol{\theta}}(t)$ that

$$\tilde{\theta}_j(t_{n+1}) = \tilde{\theta}_j(t_n) - h\frac{\sum_{k=0}^{n}\beta^{n-k}(1-\beta)\nabla_j E_k\big(\tilde{\boldsymbol{\theta}}(t_k)\big)/(1-\beta^{n+1})}{\sqrt{\sum_{k=0}^{n}\rho^{n-k}(1-\rho)\big(\nabla_j E_k\big(\tilde{\boldsymbol{\theta}}(t_k)\big)\big)^2/(1-\rho^{n+1})+\varepsilon}} + O(h^3).$$

Result I.1. *For $n \in \{0, 1, 2, \dots\}$ we have*

$$\begin{aligned}\tilde{\theta}_j(t_{n+1}) = \tilde{\theta}_j(t_n) &- h\frac{M_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)}{R_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)}\\ &+ h^2\left(\frac{M_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)P_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)}{R_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)^3} - \frac{L_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)}{R_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)}\right) + O\big(h^3\big).\end{aligned} \tag{64}$$

Derivation. We take

$$\tilde{\theta}_j(t_{n+1}) = \tilde{\theta}_j(t_n) - h\frac{M_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)}{R_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)} + O\big(h^2\big)$$

for granted. Using this and the Taylor series, we can write

$$\begin{aligned}&\nabla_j E_k\big(\tilde{\boldsymbol{\theta}}(t_{n-1})\big)\\ &\quad= \nabla_j E_k\big(\tilde{\boldsymbol{\theta}}(t_n)\big) + \sum_{i=1}^{p}\nabla_{ij}E_k\big(\tilde{\boldsymbol{\theta}}(t_n)\big)\Big\{\tilde{\theta}_i(t_{n-1}) - \tilde{\theta}_i(t_n)\Big\} + O\big(h^2\big)\end{aligned}$$

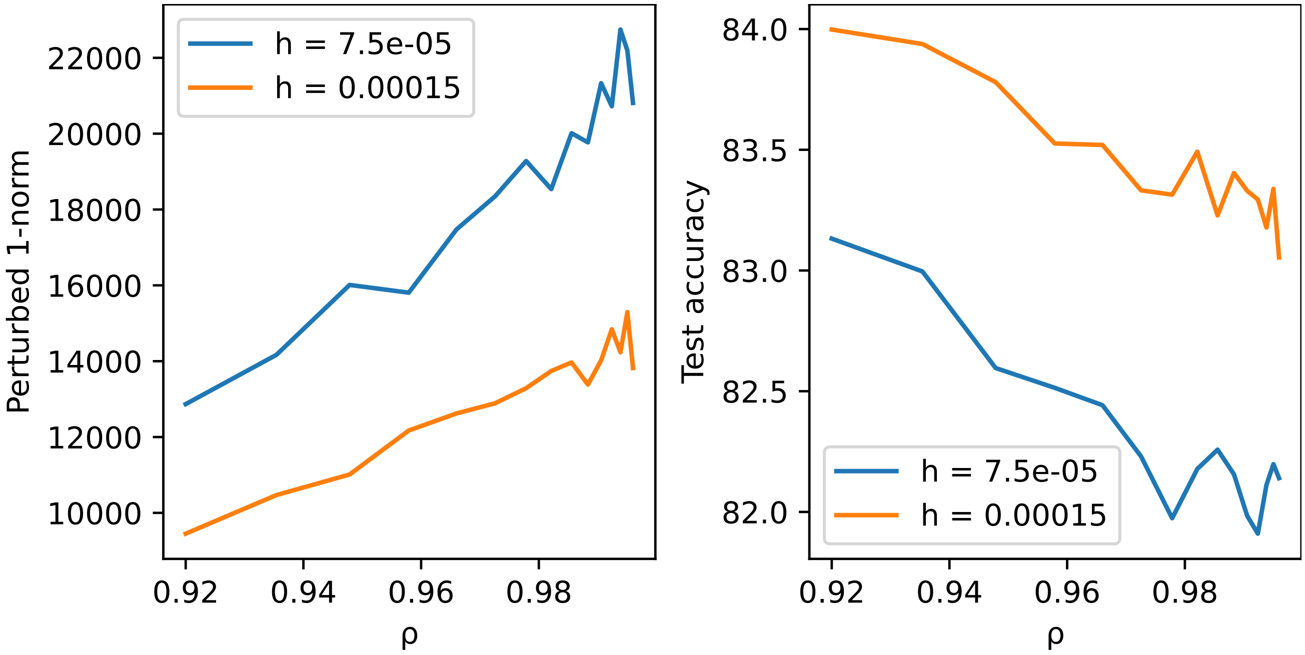

Figure 9. A vision transformer trained on CIFAR-10 with full-batch Adam. The setting and conclusions are the same as in Figure 7.

$$= \nabla_j E_k\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big) + h\sum_{i=1}^{p} \nabla_{ij} E_k\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big) \frac{M_j^{(n-1)}\Big(\tilde{\boldsymbol{\theta}}(t_{n-1})\Big)}{R_j^{(n-1)}\Big(\tilde{\boldsymbol{\theta}}(t_{n-1})\Big)} + O(h^2)$$
$$= \nabla_j E_k\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big) + h\sum_{i=1}^{p} \nabla_{ij} E_k\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big) \frac{M_j^{(n-1)}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big)}{R_j^{(n-1)}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big)} + O(h^2),$$

where in the last equality we just replaced t_{n-1} with t_n in the h-term since it only affects higher-order terms. Now doing this again for step $n-1$ instead of step n, we will have

$$\nabla_j E_k\Big(\tilde{\boldsymbol{\theta}}(t_{n-2})\Big)$$
$$= \nabla_j E_k\Big(\tilde{\boldsymbol{\theta}}(t_{n-1})\Big) + h\sum_{i=1}^{p} \nabla_{ij} E_k\Big(\tilde{\boldsymbol{\theta}}(t_{n-1})\Big) \frac{M_j^{(n-2)}\Big(\tilde{\boldsymbol{\theta}}(t_{n-1})\Big)}{R_j^{(n-2)}\Big(\tilde{\boldsymbol{\theta}}(t_{n-1})\Big)} + O(h^2)$$
$$= \nabla_j E_k\Big(\tilde{\boldsymbol{\theta}}(t_{n-1})\Big) + h\sum_{i=1}^{p} \nabla_{ij} E_k\Big(\tilde{\boldsymbol{\theta}}(t_{n-1})\Big) \frac{M_j^{(n-2)}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big)}{R_j^{(n-2)}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big)} + O(h^2),$$

where in the last equality we again replaced t_{n-1} with t_n since it only affects higher-order terms. Proceeding like this and adding the resulting equations, we have for $n \in \{0, 1, \dots\}$, $k \in \{0, \dots, n-1\}$ that

$$\nabla_j E_k\Big(\tilde{\boldsymbol{\theta}}(t_k)\Big)$$
$$= \nabla_j E_k\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big) + h\sum_{i=1}^{p} \nabla_{ij} E_k\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big) \sum_{l=k}^{n-1} \frac{M_i^{(l)}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big)}{R_i^{(l)}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big)} + O(h^2),$$

where we ignored the fact that $n-k$ is not bounded (we will get away with this because of exponential averaging). Hence, taking the square of this formal power series,

$$\rho^{n-k}(1-\rho)\Big(\nabla_j E_k\Big(\tilde{\boldsymbol{\theta}}(t_k)\Big)\Big)^2 = \rho^{n-k}(1-\rho)\Big(\nabla_j E_k\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big)\Big)^2$$

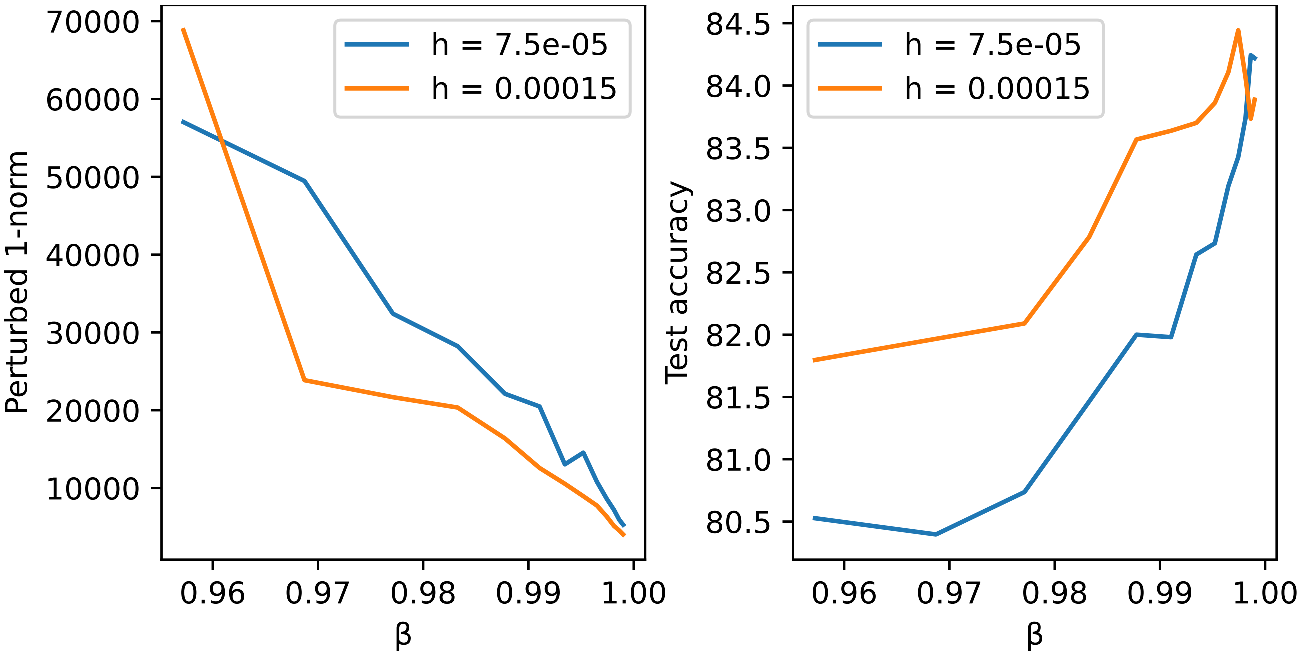

Figure 10. A vision transformer trained on CIFAR-10 with full-batch Adam. The setting and conclusions are the same as in Figure 8.

$$+ h \cdot 2\rho^{n-k}(1-\rho)\nabla_j E_k\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big) \sum_{i=1}^{p} \nabla_{ij} E_k\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big) \sum_{l=k}^{n-1} \frac{M_i^{(l)}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big)}{R_i^{(l)}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big)} + O\big(h^2\big).$$

Summing up over k, we have

$$\frac{1}{1-\rho^{n+1}} \sum_{k=0}^{n} \rho^{n-k}(1-\rho)\Big(\nabla_j E_k\Big(\tilde{\boldsymbol{\theta}}(t_k)\Big)\Big)^2 + \varepsilon = R_j^{(n)}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big)^2 + 2hP_j^{(n)}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big) + O\big(h^2\big),$$

which, using the expression for the inverse square root $\big(\sum_{r=0}^{\infty} a_r h^r\big)^{-1/2}$ of a formal power series $\sum_{r=0}^{\infty} a_r h^r$, gives us

$$\begin{aligned}&\left(\sqrt{\frac{1}{1-\rho^{n+1}} \sum_{k=0}^{n} \rho^{n-k}(1-\rho)\Big(\nabla_j E_k\Big(\tilde{\boldsymbol{\theta}}(t_k)\Big)\Big)^2 + \varepsilon}\right)^{-1} \\ &\quad = \frac{1}{R_j^{(n)}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big)} - h\frac{P_j^{(n)}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big)}{R_j^{(n)}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big)^3} + O\big(h^2\big).\end{aligned}$$

Similarly,

$$\begin{aligned}\frac{1}{1-\beta^{n+1}} \sum_{k=0}^{n} (1-\beta)\beta^{n-k} \nabla_j E_k\Big(\tilde{\boldsymbol{\theta}}(t_k)\Big) &= \frac{1}{1-\beta^{n+1}} \sum_{k=0}^{n} (1-\beta)\beta^{n-k} \nabla_j E_k\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big) \\ &+ \frac{h}{1-\beta^{n+1}} \sum_{k=0}^{n} (1-\beta)\beta^{n-k} \sum_{i=1}^{p} \nabla_{ij} E_k\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big) \sum_{l=k}^{n-1} \frac{M_i^{(l)}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big)}{R_i^{(l)}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big)} + O\big(h^2\big) \\ &= M_j^{(n)}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big) + hL_j^{(n)}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big) + O\big(h^2\big).\end{aligned}$$

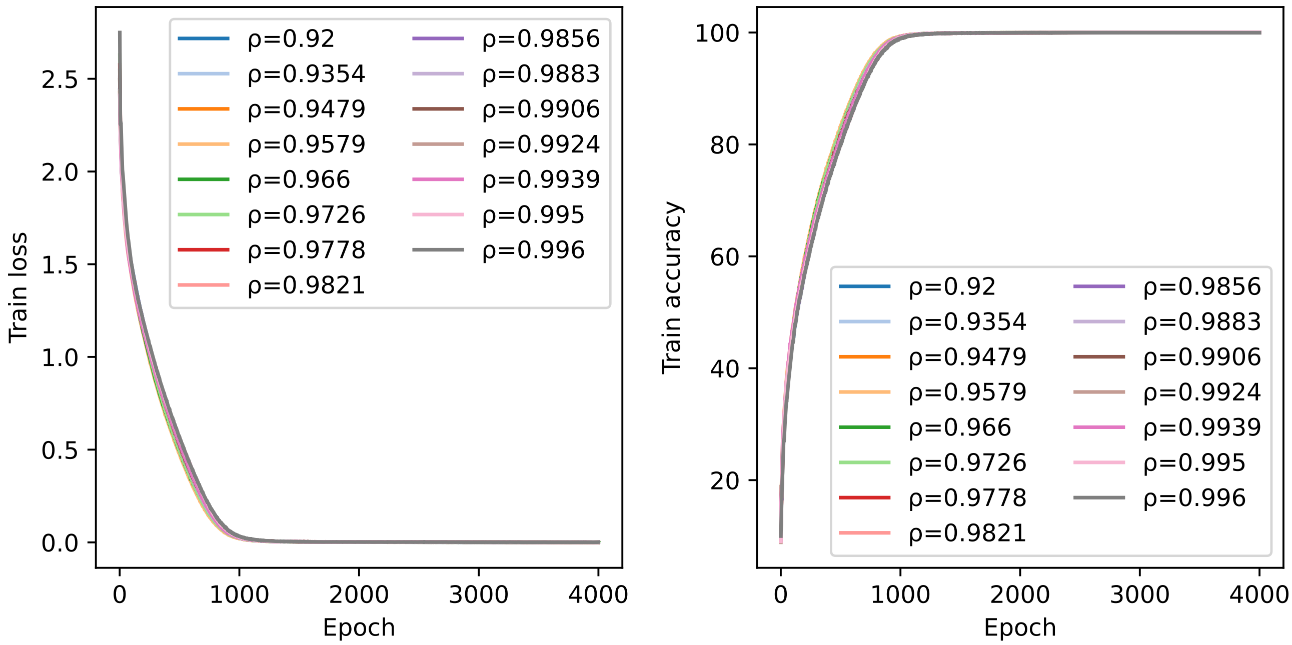

Figure 11. Train loss and train accuracy curves for full-batch Adam, ResNet-50 on CIFAR-10, $\beta = 0.99$, $\varepsilon = 10^{-8}$, $h = 10^{-4}$.

We conclude

$$\begin{aligned}\tilde{\theta}_j(t_{n+1}) &= \tilde{\theta}_j(t_n) - h\Big(M_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big) + hL_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big) + O\big(h^2\big)\Big) \\ &\quad \times \left(\frac{1}{R_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)} - h\frac{P_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)}{R_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)^3} + O\big(h^2\big)\right) + O\big(h^3\big) \\ &= \tilde{\theta}_j(t_n) - h\frac{M_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)}{R_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)} \\ &\quad + h^2\left(\frac{M_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)P_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)}{R_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)^3} - \frac{L_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)}{R_j^{(n)}\big(\tilde{\boldsymbol{\theta}}(t_n)\big)}\right) + O\big(h^3\big).\end{aligned}$$

□

Result I.2. *For $t_n \leq t < t_{n+1}$, the modified equation is* (32)*.*

Derivation. Assume that the modified flow for $t_n \leq t < t_{n+1}$ satisfies $\dot{\tilde{\boldsymbol{\theta}}} = \tilde{\mathbf{f}}\big(\tilde{\boldsymbol{\theta}}(t)\big)$ where

$$\tilde{\mathbf{f}}(\boldsymbol{\theta}) = \mathbf{f}(\boldsymbol{\theta}) + h\mathbf{f}_1(\boldsymbol{\theta}) + O\big(h^2\big).$$

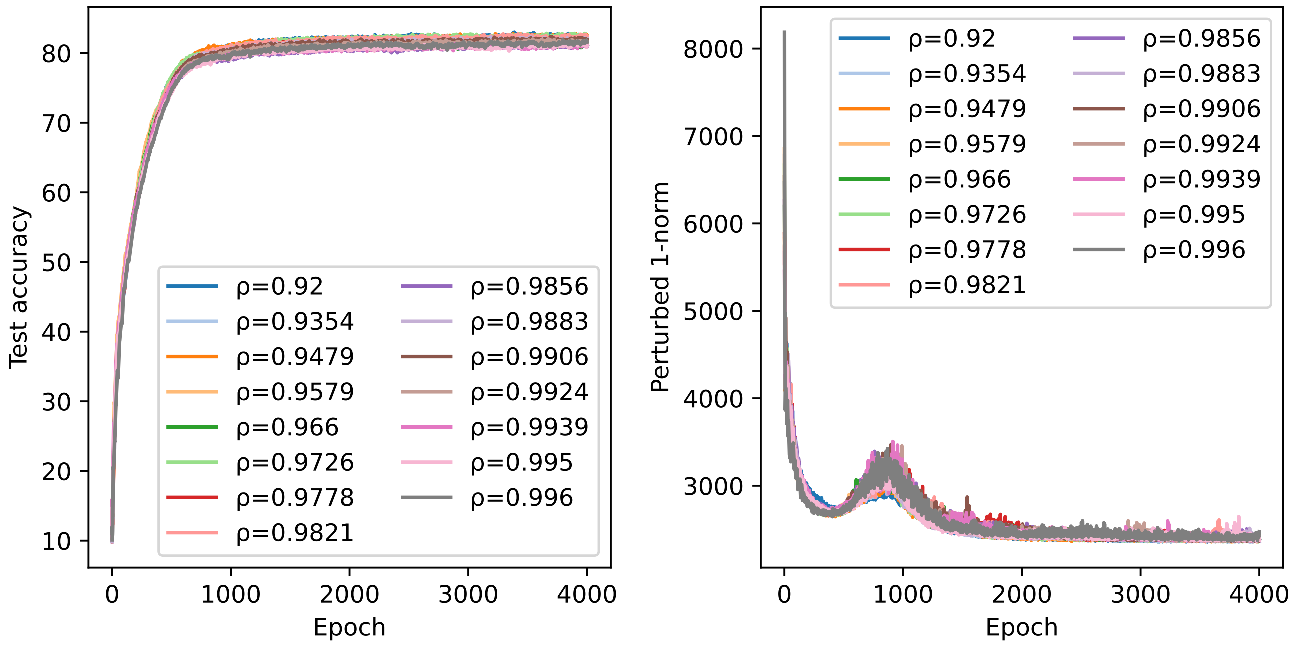

Figure 12. Test accuracy and $\|\nabla E\|_{1,\varepsilon}$ after each epoch. The setting is the same as in Figure 11.

By Taylor expansion, we have

$$\begin{aligned}\tilde{\boldsymbol{\theta}}(t_{n+1}) &= \tilde{\boldsymbol{\theta}}(t_n) + h\dot{\tilde{\boldsymbol{\theta}}}(t_n^+) + \frac{h^2}{2}\ddot{\tilde{\boldsymbol{\theta}}}(t_n^+) + O(h^3)\\ &= \tilde{\boldsymbol{\theta}}(t_n) + h\Big[\mathbf{f}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big) + h\mathbf{f}_1\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big) + O(h^2)\Big]\\ &\quad + \frac{h^2}{2}\Big[\nabla\mathbf{f}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big)\mathbf{f}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big) + O(h)\Big] + O(h^3)\\ &= \tilde{\boldsymbol{\theta}}(t_n) + h\mathbf{f}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big) + h^2\left[\mathbf{f}_1\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big) + \frac{\nabla\mathbf{f}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big)\mathbf{f}\Big(\tilde{\boldsymbol{\theta}}(t_n)\Big)}{2}\right] + O(h^3).\end{aligned} \tag{65}$$

Using Lemma I.1 and equating the terms before the corresponding powers of h in (64) and (65), we obtain

$$\begin{aligned}f_j(\boldsymbol{\theta}) &= -\frac{M_j^{(n)}(\boldsymbol{\theta})}{R_j^{(n)}(\boldsymbol{\theta})},\\ f_{1,j}(\boldsymbol{\theta}) &= -\frac{1}{2}\sum_{i=1}^{p}\nabla_i f_j(\boldsymbol{\theta}) f_i(\boldsymbol{\theta}) + \frac{M_j^{(n)}(\boldsymbol{\theta})P_j^{(n)}(\boldsymbol{\theta})}{R_j^{(n)}(\boldsymbol{\theta})^3} - \frac{L_j^{(n)}(\boldsymbol{\theta})}{R_j^{(n)}(\boldsymbol{\theta})}.\end{aligned} \tag{66}$$

It is left to find $\nabla_i f_j(\boldsymbol{\theta})$. Using

$$\begin{aligned}\nabla_i R_j^{(n)}(\boldsymbol{\theta}) &= \frac{\sum_{k=0}^{n}\rho^{n-k}(1-\rho)\nabla_{ij}E_k(\boldsymbol{\theta})\nabla_j E_k(\boldsymbol{\theta})}{(1-\rho^{n+1})R_j^{(n)}(\boldsymbol{\theta})},\\ \nabla_i M_j^{(n)}(\boldsymbol{\theta}) &= \frac{\sum_{k=0}^{n}\beta^{n-k}(1-\beta)\nabla_{ij}E_k(\boldsymbol{\theta})}{1-\beta^{n+1}}\end{aligned}$$

we have

$$\nabla_i\left(-\frac{M_j^{(n)}(\boldsymbol{\theta})}{R_j^{(n)}(\boldsymbol{\theta})}\right)$$

$$= -\frac{\frac{R_j^{(n)}(\boldsymbol{\theta})^2}{1-\beta^{n+1}} \sum_{k=0}^{n} \beta^{n-k}(1-\beta)\nabla_{ij} E_k(\boldsymbol{\theta}) - \frac{M_j^{(n)}(\boldsymbol{\theta})}{1-\rho^{n+1}} \sum_{k=0}^{n} \rho^{n-k}(1-\rho)\nabla_{ij} E_k(\boldsymbol{\theta})\nabla_j E_k(\boldsymbol{\theta})}{R_j^{(n)}(\boldsymbol{\theta})^3}$$

$$= -\frac{\sum_{k=0}^{n} \beta^{n-k}(1-\beta)\nabla_{ij} E_k(\boldsymbol{\theta})}{(1-\beta^{n+1}) R_j^{(n)}(\boldsymbol{\theta})} + \frac{M_j^{(n)}(\boldsymbol{\theta}) \sum_{k=0}^{n} \rho^{n-k}(1-\rho)\nabla_{ij} E_k(\boldsymbol{\theta})\nabla_j E_k(\boldsymbol{\theta})}{(1-\rho^{n+1}) R_j^{(n)}(\boldsymbol{\theta})^3}$$

Inserting this into (66) concludes the proof. □