Modified Loss of Momentum Gradient Descent: Fine-Grained Analysis

Matias D. Cattaneo*
Princeton University
cattaneo@princeton.edu

Boris Shigida*
Princeton University
bs1624@princeton.edu

September 9, 2025

Abstract

We analyze gradient descent with Polyak [38] heavy-ball momentum (HB) whose fixed momentum parameter $\beta \in (0, 1)$ provides exponential decay of memory. Building on Kovachki and Stuart [34], we prove that on an exponentially attractive invariant manifold the algorithm is exactly plain gradient descent with a modified loss, provided that the step size h is small enough. Although the modified loss does not admit a closed-form expression, we describe it with arbitrary precision and prove global (finite "time" horizon) approximation bounds $O(h^{\mathcal{R}})$ for any finite order $\mathcal{R} \geqslant 2$. We then conduct a fine-grained analysis of the combinatorics underlying the memoryless approximations of HB, in particular, finding a rich family of polynomials in β hidden inside which contains Eulerian and Narayana polynomials. We derive continuous modified equations of arbitrary approximation order (with rigorous bounds) and the principal flow that approximates the HB dynamics, generalizing Rosca et al. [41]. Approximation theorems cover both full-batch and mini-batch HB. Our theoretical results shed new light on the main features of gradient descent with heavy-ball momentum, and outline a road-map for similar analysis of other optimization algorithms.

*Authors are in alphabetic order. We thank Boris Hanin for insighful comments and discussions. Cattaneo gratefully acknowledges financial support from the National Science Foundation through grants SES-2019432 and SES-2241575.

Contents

1 Introduction

Gradient descent with Polyak [38] heavy-ball momentum (HB) is a well-known optimization algorithm used in practice. Given a loss function $L(\boldsymbol{\theta}) : \mathbb{R}^d \mapsto \mathbb{R}$, and initial conditions $\boldsymbol{\theta}^{(0)} \in \mathbb{R}^d$ and $\boldsymbol{v}^{(0)} \in \mathbb{R}^d$, its full-batch iteration is

$$\begin{pmatrix} \boldsymbol{\theta}^{(n+1)} \\ \boldsymbol{v}^{(n+1)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\theta}^{(n)} - h\nabla L(\boldsymbol{\theta}^{(n)}) + h\beta \boldsymbol{v}^{(n)} \\ \beta \boldsymbol{v}^{(n)} - \nabla L(\boldsymbol{\theta}^{(n)}) \end{pmatrix}, \quad n \in \mathbb{Z}_{\geqslant 0}, \tag{1}$$

or, more compactly,

$$\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)} - h \sum_{k=0}^{n} \beta^k \nabla L(\boldsymbol{\theta}^{(n-k)}) + h\beta^{n+1} \boldsymbol{v}^{(0)}, \tag{2}$$

where $\beta \in (0, 1)$ is a momentum (tuning) parameter, which we treat as a fixed constant throughout this article. The next iterate $\boldsymbol{\theta}^{(n+1)}$ depends on the whole history $\{\boldsymbol{\theta}^{(s)}\}_{s=0}^{n}$, rather than just the current iterate $\boldsymbol{\theta}^{(n)}$, which we interpret as having "memory". This algorithm or its variants, often referred to as gradient descent (GD) with momentum, are widely used in modern machine learning [23, 39, 35, 46, 28, 50, 29]. A mini-batch version of this algorithm is often used when training large-scale deep learning models.

This article studies the theoretical properties of HB, considering both full-batch and mini-batch implementations, and establishes three main results. First, we prove that on an exponentially attractive invariant manifold the algorithm is exactly plain gradient descent with a modified loss, provided that the step size h is small enough. Second, we describe the modified loss with arbitrary precision and prove global (finite "time" horizon) approximation bounds $O(h^{\mathcal{R}})$ for any finite order $\mathcal{R} \geqslant 2$. Finally, we derive continuous modified equations of arbitrary approximation order (with rigorous bounds) and the principal flow that approximates the optimization dynamics of the algorithm. Our theoretical results not only shed new light on the main properties of HB, but also outline a road-map for similar analysis of other popular optimization algorithms.

1.1 HB is GD on an Invariant Manifold

Kovachki and Stuart [34] proved that there exists a function $\boldsymbol{g}_h(\boldsymbol{\theta})$ such that $\{(\boldsymbol{\theta}, \boldsymbol{v}) : \boldsymbol{v} = -(1-\beta)^{-1}\nabla L(\boldsymbol{\theta}) + h\boldsymbol{g}_h(\boldsymbol{\theta})\}$ is an exponentially attractive invariant manifold. This perspective draws on the theory of attractive invariant manifolds [20, 30, 48]. By definition, on this manifold

$$\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)} - h\nabla L(\boldsymbol{\theta}^{(n)}) + h\beta\left(-\frac{\nabla L(\boldsymbol{\theta}^{(n)})}{1-\beta} + h\boldsymbol{g}_h(\boldsymbol{\theta}^{(n)})\right),$$

and therefore we obtain an algorithm representation without memory

$$\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)} - h\frac{1}{1-\beta}\nabla L(\boldsymbol{\theta}^{(n)}) + h^2\beta\boldsymbol{g}_h(\boldsymbol{\theta}^{(n)}). \tag{3}$$

Leveraging this insight, we further show that $\boldsymbol{g}_h(\cdot)$ has a particular structure that makes (3) plain gradient descent.

Contribution 1. Theorem 2.1 establishes that $\boldsymbol{g}_h$ is a gradient, given by $\boldsymbol{g}_h := \nabla G_h$ with G_h implicitly defined as (the anti-derivative of) the solution of a fixed point equation. This implies that on an exponentially attractive invariant manifold, HB is plain gradient descent:

$$\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)} - \frac{h}{1-\beta}\nabla\{\underbrace{L - h\beta(1-\beta)G_h}_{\text{modified loss}}\}(\boldsymbol{\theta}^{(n)}).$$

We then conduct a fine-grained analysis of how the loss is modified by memory. From the definition of $\boldsymbol{g}_h$, it is possible to write a fixed point equation that gives a formal power series expansion for $\boldsymbol{g}_h$ in h (Section 4.2). Furthermore, there are two wishes that we would like to fulfill. First, we want to have precise approximation guarantees rather than just a formal series expansion. Second, it is practically useful to cover mini-batch HB, where the loss function can be different at each iteration, rather than just full-batch. To do that, we use a different approach.

1.2 Finite-Order Memoryless Approximation

Consider the mini-batch version of (2) with the typical setting $\boldsymbol{v}^{(0)} = \mathbf{0}$:

$$\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)} - h \sum_{k=0}^{n} \beta^k \nabla L^{(n-k)}(\boldsymbol{\theta}^{(n-k)}), \tag{4}$$

where $\{L^{(s)}\}_{s \in \mathbb{Z}_{\geqslant 0}}$ are mini-batch losses. In practice, $L^{(s)}$ is the loss function obtained by taking the sth mini-batch of samples. (For now, we are agnostic to how exactly the samples are batched.) Mini-batch training usually achieves higher test accuracies [32, 36, 43], and is therefore widely used in practice [6, 39]. Cattaneo and Shigida [9] introduced a technique for converting a numerical optimization method with decaying memory into a memoryless one up to $O(h^2)$ error, that is, a second-order approximation. This paper generalizes their technique and proves an approximation bound with any desired order.

Contribution 2. For any approximation order $\mathcal{R} \in \mathbb{Z}_{\geqslant 2}$, Theorem 3.1 establishes a memoryless approximation

$$\tilde{\boldsymbol{\theta}}^{(n+1)} = \tilde{\boldsymbol{\theta}}^{(n)} + \sum_{j=1}^{\mathcal{R}} h^j \boldsymbol{d}_j^{(n)}(\tilde{\boldsymbol{\theta}}^{(n)}) \tag{5}$$

with the global guarantee

$$\sup_{n \in [0:\lfloor T/h \rfloor]} \|\boldsymbol{\theta}^{(n)} - \tilde{\boldsymbol{\theta}}^{(n)}\| = O(h^{\mathcal{R}}), \tag{6}$$

where T is any "time" horizon. Furthermore, Theorem 4.1 establishes a tractable form of the memoryless iteration coefficients $\boldsymbol{d}_j^{(n)}(\boldsymbol{\theta})$ involving a sum over unlabeled rooted trees with j vertices.

This contribution provides a higher-order, more detailed understanding of HB and its memoryless representation, which we leverage to obtain new insights on the implicit regularization and dynamics of the algorithm. For example,

$$\boldsymbol{d}_1^{(n)}(\boldsymbol{\theta}) = -\frac{1}{1-\beta} \nabla L^{(n)}(\boldsymbol{\theta}), \quad \boldsymbol{d}_2^{(n)}(\boldsymbol{\theta}) = -\beta \sum_{b=0}^{n-1} \beta^b \sum_{l'=1}^{b+1} \sum_{b'=0}^{n-l'} \beta^{b'} \nabla^2 L^{(n-1-b)}(\boldsymbol{\theta}) \nabla L^{(n-l'-b')}(\boldsymbol{\theta}).$$

In the mini-batch case, such an expression can be insightful after averaging over permutations of samples [44, 3, 9]. We will illustrate it as follows. Assume that there are $n+1$ batches in an epoch, with each batch consisting of B samples, and the kth mini-batch loss is given by

$$L^{(k)}(\boldsymbol{\theta}) = \frac{1}{B} \sum_{r=kB+1}^{kB+B} \ell_{\pi(r)}(\boldsymbol{\theta}), \quad k \in [0:n],$$

where $\{\ell_s\}_{s=1}^{(n+1)B}$ are per-sample losses and π is a random permutation of $[1:(n+1)B]$, distributed uniformly over all $((n+1)B)!$ such permutations. This corresponds to sampling without

replacement as common in practice. Denote $L(\boldsymbol{\theta}) = [(n+1)B]^{-1} \sum_{s=1}^{(n+1)B} \ell_s(\boldsymbol{\theta})$ and let

$$\Sigma_{ij} := \frac{1}{(n+1)B} \sum_{p=1}^{(n+1)B} \partial_i(\ell_p - L)\partial_j(\ell_p - L) \tag{7}$$

be the elements of the empirical covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ of per-sample gradients. Then

$$\begin{aligned}&\mathbb{E}_\pi\big(\boldsymbol{d}_1^{(n)}(\boldsymbol{\theta}) + h\boldsymbol{d}_2^{(n)}(\boldsymbol{\theta})\big)\\ &\quad= -\frac{1}{1-\beta}\nabla\Bigg(L(\boldsymbol{\theta}) + \underbrace{h\frac{\beta+o_n(1)}{2(1-\beta)^2}\|\nabla L(\boldsymbol{\theta})\|^2}_{\text{regularization by memory}} + \underbrace{h\frac{\beta+o_n(1)}{2(1-\beta)(1+\beta)}\frac{\operatorname{tr}\boldsymbol{\Sigma}(\boldsymbol{\theta})}{B}}_{\text{regularization by stochasticity}}\Bigg),\end{aligned}$$

where $\mathbb{E}_\pi$ denotes the expectation over π, $o_n(1)$ are terms that go to zero as $n \to \infty$ (for fixed β) regardless of $B \in [1\,{:}\,n+1]$ (Lemma E.2).

In the full-batch case $L^{(s)} \equiv L$, for example,

$$\begin{aligned}\boldsymbol{d}_1^{(n)}(\boldsymbol{\theta}) &= -\frac{1}{1-\beta}\nabla L(\boldsymbol{\theta}), \quad \boldsymbol{d}_2^{(n)}(\boldsymbol{\theta}) = -\frac{\beta+o_n(1)}{(1-\beta)^3}\nabla^2 L(\boldsymbol{\theta})\nabla L(\boldsymbol{\theta}),\\ \boldsymbol{d}_3^{(n)}(\boldsymbol{\theta}) &= -\frac{\beta(1+\beta)+o_n(1)}{(1-\beta)^5}\nabla^2 L(\boldsymbol{\theta})\nabla^2 L(\boldsymbol{\theta})\nabla L(\boldsymbol{\theta})\\ &\quad - \frac{\beta(1+\beta)+o_n(1)}{2(1-\beta)^5}\nabla^3 L(\boldsymbol{\theta})\big[\nabla L(\boldsymbol{\theta}), \nabla L(\boldsymbol{\theta})\big],\end{aligned}$$

so the memoryless approximation of order $\mathscr{R} = 3$ is approximately

$$\tilde{\boldsymbol{\theta}}^{(n+1)} = \tilde{\boldsymbol{\theta}}^{(n)} - \frac{h}{1-\beta}\nabla\underbrace{\left(L + \frac{h\beta}{2(1-\beta)^2}\|\nabla L\|^2 + \frac{h^2\beta(1+\beta)}{4(1-\beta)^4}\big(\nabla\|\nabla L\|^2 \cdot \nabla L\big)\right)}_{\text{modified loss up to } O(h^3)}(\tilde{\boldsymbol{\theta}}^{(n)}),$$

where the dot denotes the scalar product. We see two "implicit regularization" terms added to the loss by memory: the rescaled squared gradient norm and the rescaled directional derivative of $\|\nabla L\|^2$ along ∇L.

In both full-batch and mini-batch cases, analyzing the combinatorics of $\boldsymbol{d}_j^{(n)}(\boldsymbol{\theta})$ leads to interesting findings. First, in the limit $n \to \infty$ and after multiplying by a power of $1-\beta$, we investigate the form of the coefficients accompanying the high-order loss derivatives in $\boldsymbol{d}_j^{(n)}(\boldsymbol{\theta})$ and uncover a rich family of polynomials in β: in particular, we prove that it contains Eulerian and Narayana polynomials (Section 5.3). Second, using the natural heuristic that h multiplied by a high-order derivative (higher than two) of the loss is small, but h times the Hessian ($h\nabla^2 L$) does not have to be small mid-training [13], we can ask what the "principal" part of (5) looks like, that is, after neglecting the derivatives of the loss of order higher than two (always multiplied by some positive power of h). This leads to what we can call the *principal iteration*

$$\tilde{\boldsymbol{\theta}}^{(n+1)} = \tilde{\boldsymbol{\theta}}^{(n)} - h\sigma_\beta\big(h\nabla^2 L\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)\big)\nabla L\big(\boldsymbol{\theta}^{(n)}\big) + \mathrm{NPT},$$

after taking $\mathscr{R}$ formally to infinity, where $\sigma_\beta(\cdot)$ is a power series expansion (in powers of z) of

$$\sigma_\beta(z) = \frac{2}{1-\beta+z+\sqrt{(1-\beta-z)^2 - 4\beta z}}$$

(Corollary 5.3), and NPT means "non-principal terms" (with $\nabla^3 L$ etc.). The term "principal iteration" comes from the analogous term *principal flow* coined in Rosca et al. [41] for continuous

modified equations of plain GD. Third, we combine our framework with continuous approximations (discussed next) to derive principal flow for HB (Corollary 5.8). See [41] for a discussion of the importance of such (complex) flows: they capture oscillatory behavior and divergence, in contrast to standard continuous approximations. We provide an illustration in Section 5.4.

In the case of a quadratic loss, the principal iteration and principal flow are both exact (if HB is initialized on the manifold introduced above; see Section 5.4). To illustrate the implicit regularization effect further, consider the example of least-squares regression. Let $\boldsymbol{X} \in \mathbb{R}^{N\times d}$, $\boldsymbol{y} \in \mathbb{R}^N$ and

$$L(\boldsymbol{\theta}) = \frac{1}{2N}\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\|^2.$$

Additionally, let $\boldsymbol{\theta}^\star$ satisfy the normal equations $\boldsymbol{X}^\intercal\boldsymbol{X}\boldsymbol{\theta}^\star = \boldsymbol{X}^\intercal\boldsymbol{y}$; in particular, $L(\boldsymbol{\theta}) = \frac{1}{2N}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\intercal\boldsymbol{X}^\intercal\boldsymbol{X}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star) + C$ (where C does not depend on $\boldsymbol{\theta}$). Then, the memoryless iteration is the gradient descent

$$\tilde{\boldsymbol{\theta}}^{(n+1)} = \tilde{\boldsymbol{\theta}}^{(n)} - h\nabla\tilde{L}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big) \quad \text{with} \quad \tilde{L}(\boldsymbol{\theta}) = \frac{1}{2N}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\intercal\sigma_\beta\Big(\frac{h}{N}\boldsymbol{X}^\intercal\boldsymbol{X}\Big)\boldsymbol{X}^\intercal\boldsymbol{X}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star).$$

Introducing $\sigma_\beta\big(\frac{h}{N}\boldsymbol{X}^\intercal\boldsymbol{X}\big)$ rescales the learning rate (by $(1-\beta)^{-1}$) and applies a spectral filter to the Hessian.

1.3 Modified Equations

The discrete iteration without memory (5) has the following advantage over HB: it only has one evolving variable $\boldsymbol{\theta}^{(n)} \in \mathbb{R}^d$, and it can be approximated by a continuous ODE solution without sacrificing anything in terms of the approximation guarantee. (Enlarging the phase space by treating HB as an evolution of $(\boldsymbol{\theta}^{(n)}, \boldsymbol{v}^{(n)}) \in \mathbb{R}^{2d}$ or, equivalently, considering higher-order ODEs does not lead to an improvement in the approximation guarantee [34].) Then, finding continuous trajectories closely tracking discrete numerical iterations allows to analyze qualitatively the finite-h behavior of the iteration. This approach, widely used in numerical analysis and machine learning, is called the method of modified equations, backward error analysis (BEA), or high-resolution ODE approximation: it has been employed and studied for many decades, e. g. in works [49, 47, 24, 19, 5, 15, 26, 8, 40, 11, 42, 17, 34] and many others; textbook references include [45, 16]. Recently, this method was used in machine learning for finding implicit biases of gradient-based algorithms [2, 44, 37, 22, 10, 41]. For us, it is just a corollary of the approximation (5).

Contribution 3. For any approximation order $\mathcal{R} \in \mathbb{Z}_{\geqslant 2}$, Corollary 5.1 finds a set of functions $\{\boldsymbol{f}_j^{(n)}(\boldsymbol{\theta}) : j \in [1:\mathcal{R}], n \in \mathbb{Z}_{\geqslant 0}\}$ such that the (unique) continuous solution to the piecewise ODE

$$\dot{\boldsymbol{\theta}}(t) = \sum_{j=1}^{\mathcal{R}} h^{j-1}\boldsymbol{f}_j^{(n)}(\boldsymbol{\theta}(t)), \quad t \in [nh, (n+1)h],$$

with an appropriate initial condition, has a global approximation guarantee

$$\sup_{n\in[0:\lfloor T/h\rfloor]} \|\boldsymbol{\theta}^{(n)} - \boldsymbol{\theta}(nh)\| = O(h^{\mathcal{R}}), \tag{8}$$

where T is any time horizon.

The ODE is defined piecewise because of the dependence of $\boldsymbol{d}_j^{(n)}(\cdot)$ on n, and the solution will be continuous but not necessarily smooth. Nonetheless, in the full-batch case, the ODEs corresponding to neighboring pieces are the same up to an exponentially decaying error: we

do not lose any information compared to having an ODE only on the attractive manifold but globally defined, as done for $\mathcal{R} = 2$ in Kovachki and Stuart [34]. The idea of using piecewise ODEs in this setting is due to Ghosh et al. [22], who derived the modified equations for $\mathcal{R} = 2$.

In contrast to the discrete case, the right-hand side of the modified equation does *not* become a gradient for large n, even for full-batch HB. Therefore, HB does *not* approximately follow a smooth gradient flow trajectory with a modified loss (with approximation order higher than h^2), as we make explicit in Remark 5.2.

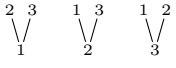
1.4 Organization

The paper continues as follows. In an attempt to make the paper self-contained, we introduce the relevant mathematical concepts and notation (e. g., labeled and unlabeled rooted trees, their symmetry coefficients, marking vertices to split the tree into a forest) in Section 1.5. Section 2 states the theorem and briefly outlines the argument corresponding to Contribution 1. Section 3 is devoted to the main approximation theorem corresponding to Contribution 2, removing memory from HB and controlling the error while doing so. In particular, we introduce a somewhat complicated-looking recursive definition for the memoryless iteration coefficients $\boldsymbol{d}_j^{(n)}(\boldsymbol{\theta})$. In Section 4, we analyze the form of these coefficients, and prove that a much simpler characterization holds which involves the sum over rooted trees. We use the notation introduced to solve a fixed-point equation for $\boldsymbol{g}_h$ as a series over rooted trees. Section 5 reports corollaries and implications of our main theorems, including an approximation theorem for the modified equation corresponding to Contribution 3, a study of the polynomials arising in the form of $\boldsymbol{d}_j^{(n)}(\boldsymbol{\theta})$, a formalization of the principal iteration, and a discussion of principal flow. All proof strategies are explained in the main text, and omitted technical details are provided in the Appendix.

1.5 Technical Background and Notation

We start with well-known definitions of labeled and unlabeled rooted trees. See, for example, [18, 31] for details. For a fixed non-empty finite set V and a point $r \in V$, a *labeled rooted tree* τ with root r is a pair (G, r), where $G = (V, E)$ is a connected acyclic graph with V as the vertex set (where E is the edge set; $|E| = |V| - 1$). Taking a different point of view, τ is a function $\tau\colon V \setminus \{r\} \to V$ with no non-empty invariant subset (mapping a vertex to its parent). Let $\mathcal{A}[V]$ denote all labeled trees with vertex set V. We will write $|\tau| := |V|$ for the number of vertices. It will also be convenient to use the notation $\mathcal{A} := \bigcup_{m=1}^{\infty} \mathcal{A}[1\colon m]$ and $\mathcal{A}_\varnothing := \{\varnothing\} \cup \mathcal{A}$.

Two labeled rooted trees $\tau\colon V \setminus \{r\} \to V$ and $\tau'\colon V' \setminus \{r'\} \to V'$ are called isomorphic if there is a bijection $\mathcal{b}\colon V \to V'$ such that $\mathcal{b}(r) = r'$ and $\tau' = \mathcal{b} \circ \tau \circ \mathcal{b}^{-1}\big|_{V' \setminus \{r'\}}$. An unlabeled rooted tree with m vertices is an object that corresponds to a class of isomorphic labeled rooted trees with m vertices. Specifically, we can fix a canonical set of m elements $[1\colon m]$ and see an *unlabeled rooted tree* as an orbit of the permutation group $S[1\colon m]$ acting on $\mathcal{A}[1\colon m]$: the action of $\pi \in S[1\colon m]$ on $\tau \in \mathcal{A}[1\colon m]$ is $\pi\tau = \pi \circ \tau \circ \pi^{-1}\big|_{[1:m]\setminus\{\pi(r)\}}$. The set of unlabeled rooted trees with m vertices will be denoted $\tilde{\mathcal{A}}[m]$, and the set of unlabeled rooted trees with any number of vertices by $\tilde{\mathcal{A}} := \bigcup_{m=1}^{\infty} \tilde{\mathcal{A}}[m]$; in addition, put $\tilde{\mathcal{A}}_\varnothing := \{\varnothing\} \cup \tilde{\mathcal{A}}$. For example, the following orbit (consisting of three labeled rooted trees) is the unlabeled rooted tree $\curlyvee$:



(9)

For any such orbit, the order of the stabilizer group of each element is the same. (The stabilizer group is the subgroup of permutations leaving a labeled rooted tree intact. For example, the first tree in (9) has stabilizer group $\{\mathrm{id}, 2 \leftrightarrow 3\}$.) Hence, for an unlabeled rooted tree τ we can define the *symmetry coefficient* $\sigma(\tau)$ as the order of the stabilizer group of each element. By the

well-known theorem, the length of the orbit is the order of the group $m!$ divided by the order of each stabilizer group $\sigma(\tau)$; so, there are $m!/\sigma(\tau)$ labeled rooted trees (on a fixed vertex set of size m) corresponding to an unlabeled rooted tree τ.

Whenever there is no chance of confusion, we will use the same terms and symbols when working with labeled and unlabeled rooted trees, for example, we will write $\sigma(\tau)$ regardless of whether $\tau \in \mathcal{A}$ or $\tau \in \tilde{\mathcal{A}}$ (because each labeled rooted tree has a corresponding unlabeled rooted tree); we will say an unlabeled rooted tree $\tau \in \tilde{\mathcal{A}}[m]$ "has $|\tau| = m$ vertices" even though there is no such thing as a vertex of an unlabeled rooted tree, etc.

Since the ordering of subtrees does not matter, there is a one-to-one correspondence between $\tau \in \tilde{\mathcal{A}}[m]$ and a multiset $[\tau_1, \dots, \tau_\ell]$ of unlabeled rooted trees whose sum of vertices is $m-1$ (the subtrees rooted at the children of τ's root). We will sometimes write simply $\tau = [\tau_1, \dots, \tau_\ell]$ to reflect this fact. Fixing some canonical ordering in each set $\tilde{\mathcal{A}}[s]$, we will denote by $\big\{\mu_1^s(\tau), \dots, \mu_{|\tilde{\mathcal{A}}[s]|}^s(\tau)\big\}_{s=1}^{m-1}$ the *multiplicities* of such subtrees: this means that the first unlabeled subtree with s vertices appears $\mu_1^s(\tau)$ times, the second one appears μ_2^s times, and so on. In particular, $\sum_{s=1}^{m-1}\big(\mu_1^s(\tau) + \ldots + \mu_{|\tilde{\mathcal{A}}[s]|}^s(\tau)\big) = \ell$. It is a standard fact that

$$\sigma(\tau) = \sigma(\tau_1)\dots\sigma(\tau_\ell)\prod_{s=1}^{m-1}\prod_{t=1}^{|\tilde{\mathcal{A}}[s]|}\mu_t^s(\tau)!. \tag{10}$$

In addition, it will be convenient to call $c_m \in \tilde{\mathcal{A}}[m]$ the *chain* with m vertices if either $m = 1$ or it is the (unique) unlabeled rooted tree corresponding to any element of $\mathcal{A}[1\!:\!m]$ whose root has degree 1 and all other vertex degrees do not exceed 2.

Marking Vertices (via Admissible Cuts). For a labeled rooted tree $\tau \in \mathcal{A}[V]$, define a *marking* m of that tree as a subset of non-root vertices of τ (interpreted as *marked*) such that if $v \in V$ is marked, then no other vertices in the subtree of v are marked. In other words, the marked vertices are the upper (farther from the root) vertices of an admissible cut (e. g. [1]), or the roots of the forest obtained after removing a subtree with the same root [18, 12]. Hence, marking $|m|$ vertices is the same as selecting $|m|$ disjoint subtrees $\tau_1, \dots, \tau_{|m|}$ (rooted at those vertices), and it splits the tree τ into $|m|+1$ labeled rooted trees $\tau_0^m = \tau \setminus (\tau_1^m \cup \ldots \cup \tau_{|m|}^m)$, $\tau_1^m, \dots, \tau_{|m|}^m$. So, we will sometimes write $m = (\tau_0^m, \{\tau_1^m, \dots, \tau_{|m|}^m\})$ to reflect this fact. Denote $\mathcal{M}_{\tau,i}$ the set of all markings of τ with $|m| = i$ marked vertices, and $\mathcal{M}_\tau := \bigcup_{i=0}^{|V|-1}\mathcal{M}_{\tau,i}$ the set of all markings of τ.

Other Notation. We denote by $L(\boldsymbol{\theta})$ the full-batch loss and by $L^{(s)}(\boldsymbol{\theta})$ the sth mini-batch loss, where $\boldsymbol{\theta} \in \mathbb{R}^d$ is the evolving parameter of fixed dimension d. We denote by ∇^k the kth derivative tensor, for example, $\nabla^3 L(\boldsymbol{\theta})\big[\nabla L(\boldsymbol{\theta}), \nabla L(\boldsymbol{\theta})\big]$ is a vector whose jth component is $\sum_{i,l=1}^d \partial_{jil}L(\boldsymbol{\theta})\partial_i L(\boldsymbol{\theta})\partial_l L(\boldsymbol{\theta})$. The notation for the norm $\|\cdot\|$ without indices will always mean the Euclidean (operator) norm. The set of integers no smaller than some k will be denoted by $\mathbb{Z}_{\geqslant k} := [k, +\infty) \cap \mathbb{Z}$, and the set of integers between a and b inclusive by $[a\!:\!b] := [a, b] \cap \mathbb{Z}$. A sum over an empty set is be definition 0, and a product is 1. Section 3 defines the set $\mathcal{K}_{i,l}$, the memoryless iteration coefficients $\boldsymbol{d}_j^{(n)}(\boldsymbol{\theta})$, the history coefficients $\widetilde{\boldsymbol{d}}_m^{(n,a)}(\boldsymbol{\theta})$; Section 5.1 defines the backward error analysis coefficients $\boldsymbol{f}_j^{(n)}(\boldsymbol{\theta})$. For any operator $\boldsymbol{a}_{n,h}$ such that $\|\boldsymbol{a}_{n,h}\|$ is defined, we write $\boldsymbol{a}_{n,h} = O(g)$ if $\sup_{n,h}\|\boldsymbol{a}_{n,h}\| \leqslant Cg$ with some constant C, where the supremum is over the set of admissible n and h which is clear from context.

2 Full-Batch HB is (Almost) GD with a Modified Loss

Kovachki and Stuart [34] proved that for full-batch HB (1) there exists a function $\boldsymbol{g}_h(\boldsymbol{\theta})$ such that $\{(\boldsymbol{\theta}, \boldsymbol{v}) : \boldsymbol{v} = -(1-\beta)^{-1}\nabla L(\boldsymbol{\theta}) + h\boldsymbol{g}_h(\boldsymbol{\theta})\}$ is an invariant manifold, which means

$$\boldsymbol{v}^{(n)} = -\frac{\nabla L(\boldsymbol{\theta}^{(n)})}{1-\beta} + h\boldsymbol{g}_h(\boldsymbol{\theta}^{(n)}) \quad \Rightarrow \quad \boldsymbol{v}^{(n+1)} = -\frac{\nabla L(\boldsymbol{\theta}^{(n+1)})}{1-\beta} + h\boldsymbol{g}_h(\boldsymbol{\theta}^{(n+1)}).$$

On this invariant manifold, on the one hand,

$$\begin{aligned}\boldsymbol{v}^{(n+1)} &= -(1-\beta)^{-1}\nabla L(\boldsymbol{\theta}^{(n+1)}) + h\boldsymbol{g}_h(\boldsymbol{\theta}^{(n+1)}) \\ &= -\frac{1}{1-\beta}\nabla L\left(\boldsymbol{\theta}^{(n)} - \frac{h}{1-\beta}\nabla L(\boldsymbol{\theta}^{(n)}) + h^2\beta\boldsymbol{g}_h(\boldsymbol{\theta}^{(n)})\right) \\ &\quad + h\boldsymbol{g}_h\left(\boldsymbol{\theta}^{(n)} - \frac{h}{1-\beta}\nabla L(\boldsymbol{\theta}^{(n)}) + h^2\beta\boldsymbol{g}_h(\boldsymbol{\theta}^{(n)})\right),\end{aligned}$$

while, on the other hand,

$$\begin{aligned}\boldsymbol{v}^{(n+1)} &= \beta\boldsymbol{v}^{(n)} - \nabla L(\boldsymbol{\theta}^{(n)}) = -\frac{\beta}{1-\beta}\nabla L(\boldsymbol{\theta}^{(n)}) + h\beta\boldsymbol{g}_h(\boldsymbol{\theta}^{(n)}) - \nabla L(\boldsymbol{\theta}^{(n)}) \\ &= -\frac{1}{1-\beta}\nabla L(\boldsymbol{\theta}^{(n)}) + h\beta\boldsymbol{g}_h(\boldsymbol{\theta}^{(n)}).\end{aligned}$$

This must hold for any $\boldsymbol{\theta}^{(n)}$, so $\boldsymbol{g}_h(\cdot)$ must satisfy

$$\begin{aligned}&-\frac{1}{1-\beta}\nabla L\left(\boldsymbol{\theta} - \frac{h}{1-\beta}\nabla L(\boldsymbol{\theta}) + h^2\beta\boldsymbol{g}_h(\boldsymbol{\theta})\right) + h\boldsymbol{g}_h\left(\boldsymbol{\theta} - \frac{h}{1-\beta}\nabla L(\boldsymbol{\theta}) + h^2\beta\boldsymbol{g}_h(\boldsymbol{\theta})\right) \\ &\quad = -\frac{1}{1-\beta}\nabla L(\boldsymbol{\theta}) + h\beta\boldsymbol{g}_h(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \mathbb{R}^d.\end{aligned} \tag{11}$$

To prove that such a function exists, they define a mapping $T\colon \Gamma \to \Gamma$, where Γ is an appropriately chosen closed subset of $C(\mathbb{R}^d, \mathbb{R}^d)$ with the usual sup-norm, by

$$T\boldsymbol{g}(\boldsymbol{\zeta}) = \frac{1}{h(1-\beta)}\big\{\nabla L(\boldsymbol{\zeta}) - \nabla L(\boldsymbol{\theta})\big\} + \beta\boldsymbol{g}(\boldsymbol{\theta}), \tag{12}$$

where $\boldsymbol{\theta} \leftrightarrow \boldsymbol{\zeta} = \boldsymbol{\theta} - h(1-\beta)^{-1}\nabla L(\boldsymbol{\theta}) + h^2\beta\boldsymbol{g}(\boldsymbol{\theta})$ is a bijection between $\mathbb{R}^d$ and itself for any $\boldsymbol{g} \in \Gamma$ (this fact is Lemma A.1). A fixed point of such a mapping T would satisfy (11). They prove that T is a contraction on Γ, allowing to apply the contracting mapping principle.

We follow the same conceptual approach, but apply some technical tweaks (different metric, different Γ), to prove that the additional requirement $\boldsymbol{g}_h \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ with symmetric $\nabla\boldsymbol{g}_h$ is satisfiable as well, which implies that such $\boldsymbol{g}_h$ is a gradient. Specifically, define the set

$$\Gamma := \big\{\boldsymbol{g} \in C^1(\mathbb{R}^d, \mathbb{R}^d) : \sup_{\boldsymbol{\theta}\in\mathbb{R}^d}\|\boldsymbol{g}\| \leqslant \gamma,\ \sup_{\boldsymbol{\theta}\in\mathbb{R}^d}\|\nabla\boldsymbol{g}\| \leqslant \delta,\ \nabla\boldsymbol{g} \text{ is symmetric and } \lambda\text{-Lipschitz}\big\},$$

with the norm

$$\|\boldsymbol{g}\|_\Gamma := \sup_{\boldsymbol{\theta}\in\mathbb{R}^d}\|\boldsymbol{g}(\boldsymbol{\theta})\| + \sup_{\boldsymbol{\theta}\in\mathbb{R}^d}\|\nabla\boldsymbol{g}(\boldsymbol{\theta})\|, \tag{13}$$

where both the vector norm and the matrix norm in the right-hand side are Euclidean. The space $\big\{\boldsymbol{g} \in C^1(\mathbb{R}^d, \mathbb{R}^d) : \|\boldsymbol{g}\|_\Gamma < \infty\big\}$ with norm (13) is a Banach space, and Γ is a closed subset of it; therefore Γ can be seen as a complete metric space with the metric induced by the norm $\|\cdot\|_\Gamma$. The constants γ, δ, λ are chosen in Appendix A.

The assumptions of the theorem are essentially the same as in [34], except we (naturally) need one more derivative of the loss to be Lipschitz.

Theorem 2.1. *Assume $L(\cdot) \in C^3(\mathbb{R}^d, \mathbb{R})$ with constants D_1, D_2, D_3, D_4 such that*

$$\sup_{\boldsymbol{\theta}} \|\nabla^j L(\boldsymbol{\theta})\| \leqslant D_j \text{ for } j \in [1:3], \quad \sup_{\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2} \frac{\|\nabla^3 L(\boldsymbol{\theta}_1) - \nabla^3 L(\boldsymbol{\theta}_2)\|}{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|} \leqslant D_4,$$

where the norms are Euclidean operator norms. Then, if h is small enough, there exists a unique $\boldsymbol{g}_h(\cdot) \in \Gamma$ satisfying (11), and the following exponential attractivity property holds:

$$\left\| \boldsymbol{v}^{(n)} + \frac{1}{1-\beta} \nabla L(\boldsymbol{\theta}^{(n)}) - h \boldsymbol{g}_h(\boldsymbol{\theta}^{(n)}) \right\| \leqslant (\beta + h^2 \beta \delta)^n \left\| \boldsymbol{v}^{(0)} + \frac{1}{1-\beta} \nabla L(\boldsymbol{\theta}^{(0)}) - h \boldsymbol{g}_h(\boldsymbol{\theta}^{(0)}) \right\|.$$

The proof is a bounding argument very heavy on long equations, so it is moved to Appendix A. Lemma A.2 proves that indeed T maps Γ to itself, and Lemma A.3 proves that T is a contraction on Γ. Since Γ is a complete metric space with the metric $\|\boldsymbol{g}_1 - \boldsymbol{g}_2\|_\Gamma$, the contraction mapping principle implies that there is a unique fixed point $\boldsymbol{g}_h \in \Gamma$ of the operator T, i.e., $T\boldsymbol{g}_h = \boldsymbol{g}_h$. Exponential attractivity is tackled in Lemma A.4.

3 Mini-Batch HB: Approximation Theorem

Cattaneo and Shigida [9] introduced a method for removing memory in a class of numerical optimization algorithms with decaying memory, which can be applied to HB up to $O(h^2)$. Following their idea, we prove the approximation of HB by a memoryless iteration with the global error bound $O(h^{\mathcal{R}})$ where $\mathcal{R} \in \mathbb{Z}_{\geqslant 2}$. Considering higher-order approximations requires additional notation and technical work. Denoting

$$\mathcal{K}_{i,l} = \left\{ (k_0, \ldots, k_l) \in \mathbb{Z}_{\geqslant 0}^{l+1} : k_0 + \ldots + k_l = i, k_1 + \ldots + k_l = l \right\},$$

define the *memoryless iteration coefficients*

$$\begin{aligned} \boldsymbol{d}_1^{(n)}(\boldsymbol{\theta}) &= -\sum_{k=0}^{n} \beta^k \nabla L^{(n-k)}(\boldsymbol{\theta}), \\ \boldsymbol{d}_m^{(n)}(\boldsymbol{\theta}) &= -\sum_{k=1}^{n} \beta^k \sum_{\substack{i,l \geqslant 0 \\ i+l=m-1}} \sum_{(i_0,\ldots,i_l) \in \mathcal{K}_{i,l}} \frac{1}{i_0! \ldots i_l!} \\ &\quad \times \nabla^{i+1} L^{(n-k)}(\boldsymbol{\theta}) \Big(\underbrace{\widetilde{\boldsymbol{d}}_1^{(n,k)}(\boldsymbol{\theta})}_{i_0 \text{ times}}, \ldots, \underbrace{\widetilde{\boldsymbol{d}}_{l+1}^{(n,k)}(\boldsymbol{\theta})}_{i_l \text{ times}} \Big), \quad m \geqslant 2, \end{aligned} \tag{14}$$

where the *history terms* satisfy the iteration

$$\widetilde{\boldsymbol{d}}_m^{(n,a)}(\boldsymbol{\theta}) = -\sum_{s=1}^{a} \sum_{\substack{j \geqslant 1, i,l \geqslant 0 \\ i+j+l=m}} \sum_{(k_0,\ldots,k_l) \in \mathcal{K}_{i,l}} \frac{1}{k_0! \ldots k_l!} \nabla^i \boldsymbol{d}_j^{(n-s)}(\boldsymbol{\theta}) \Big(\underbrace{\widetilde{\boldsymbol{d}}_1^{(n,s)}(\boldsymbol{\theta})}_{k_0 \text{ times}}, \ldots, \underbrace{\widetilde{\boldsymbol{d}}_{l+1}^{(n,s)}(\boldsymbol{\theta})}_{k_l \text{ times}} \Big), \tag{15}$$

with $a \in [1:n]$, and $m \in \mathbb{Z}_{\geqslant 1}$.

Although this recursion may look complex, we will use it as the main definition for $\{\boldsymbol{d}_m^{(n)}(\boldsymbol{\theta})\}$ because it is convenient for proving Theorem 3.1. It involves taking high-order derivatives of recursively defined quantities, specifically $\nabla^i \boldsymbol{d}_j^{(n-s)}$ in (15), and for this reason it is hard to analyze. We provide a different, more natural, form in Section 4, where only the loss function is differentiated.

Theorem 3.1 (Approximation by a memoryless iteration)**.** *Let a family of loss functions $\{L^{(n)}\}_{n\geqslant 0}$ be defined on an open convex domain $\mathcal{D}$ in $\mathbb{R}^d$, and assume each loss function $L^{(n)}\colon \mathcal{D}\to\mathbb{R}$ is $2\mathcal{R}$-times continuously differentiable with bounded uniformly in n derivatives up to order $2\mathcal{R}$, where $\mathcal{R}\in\mathbb{Z}_{\geqslant 2}$. Let $\{\boldsymbol{\theta}^{(n)}\}_{n=0}^{\infty}\subset\mathcal{D}$ be the HB iteration (4) with initial condition $\boldsymbol{\theta}^{(0)}\in\mathcal{D}$, and $\{\tilde{\boldsymbol{\theta}}^{(n)}\}_{n=0}^{\infty}\subset\mathcal{D}$ the iteration (5) with the same initial condition $\tilde{\boldsymbol{\theta}}^{(0)}=\boldsymbol{\theta}^{(0)}$. Let $T>0$ be a fixed "time" horizon. Then, for each $h\in(0,1/2)$,*

$$\sup_{n\in[0:\lfloor T/h\rfloor]}\left\|\tilde{\boldsymbol{\theta}}^{(n+1)}-\tilde{\boldsymbol{\theta}}^{(n)}+h\sum_{k=0}^{n}\beta^k\nabla L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n-k)}\big)\right\|\leqslant C_1h^{\mathcal{R}+1}\tag{16}$$

and, as a consequence,

$$\sup_{n\in[0:\lfloor T/h\rfloor]}\left\|\boldsymbol{\theta}^{(n)}-\tilde{\boldsymbol{\theta}}^{(n)}\right\|\leqslant C_2h^{\mathcal{R}},\tag{17}$$

where C_1 and C_2 are some constants depending on T.

To build intuition we offer a sketch of out proof next; the full argument can be found in Appendix B.

3.1 Proof Sketch

By definition, (16) will follow from the following bound on the error introduced by removing memory:

$$-\sum_{k=0}^{n}\beta^k\nabla L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n-k)}\big)\overset{?}{=}\sum_{m=0}^{\mathcal{R}-1}h^m\boldsymbol{d}_{m+1}^{(n)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)+O(h^{\mathcal{R}}),\tag{18}$$

that is, our task is to rewrite the left-hand side in such a way that instead of a function of all previous $\tilde{\boldsymbol{\theta}}^{(n-k)}$ it becomes just a function of $\tilde{\boldsymbol{\theta}}^{(n)}$.

It is shown by induction (Lemma B.1) that $\boldsymbol{d}_j^{(n)}(\boldsymbol{\theta})=O(1)$ which implies $\tilde{\boldsymbol{\theta}}^{(n-k)}-\tilde{\boldsymbol{\theta}}^{(n)}=O(kh)$ (Lemma B.2). This allows to claim that the remainder is $\mathrm{Rem}^{(n-k)}=O(k^{\mathcal{R}}h^{\mathcal{R}})$ in the Taylor expansion

$$\begin{aligned}\nabla L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n-k)}\big)=&\sum_{i=0}^{\mathcal{R}-1}\frac{1}{i!}\nabla^{i+1}L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)\big(\underbrace{\tilde{\boldsymbol{\theta}}^{(n-k)}-\tilde{\boldsymbol{\theta}}^{(n)},\dots,\tilde{\boldsymbol{\theta}}^{(n-k)}-\tilde{\boldsymbol{\theta}}^{(n)}}_{i\text{ times}}\big)\\&+\mathrm{Rem}^{(n-k)},\end{aligned}\tag{19}$$

where

$$\begin{aligned}\mathrm{Rem}^{(n-k)}=&\frac{1}{(\mathcal{R}-1)!}\int_0^1(1-t)^{\mathcal{R}-1}\\&\times\nabla^{\mathcal{R}+1}L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}+t\big(\tilde{\boldsymbol{\theta}}^{(n-k)}-\tilde{\boldsymbol{\theta}}^{(n)}\big)\big)\big(\underbrace{\tilde{\boldsymbol{\theta}}^{(n-k)}-\tilde{\boldsymbol{\theta}}^{(n)},\dots,\tilde{\boldsymbol{\theta}}^{(n-k)}-\tilde{\boldsymbol{\theta}}^{(n)}}_{\mathcal{R}\text{ times}}\big)\,\mathrm{d}t.\end{aligned}$$

This is partial progress: we rewrote $\nabla L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n-k)}\big)$ in such a way that it depends only multilinearly on (copies of) $\tilde{\boldsymbol{\theta}}^{(n-k)}-\tilde{\boldsymbol{\theta}}^{(n)}$, whereas all loss derivatives are evaluated at the current iterate $\tilde{\boldsymbol{\theta}}^{(n)}$. Now we need to express $\tilde{\boldsymbol{\theta}}^{(n-k)}$ through $\tilde{\boldsymbol{\theta}}^{(n)}$. This is done in the following lemma, which clarifies the meaning of the history terms $\widetilde{\boldsymbol{d}}_m^{(n,k)}(\boldsymbol{\theta})$:

Lemma 3.2. *We have for $r\in[1:\mathcal{R}]$, $k\in[1:n]$*

$$\tilde{\boldsymbol{\theta}}^{(n-k)}=\tilde{\boldsymbol{\theta}}^{(n)}+\sum_{m=1}^{r-1}h^m\widetilde{\boldsymbol{d}}_m^{(n,k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)+O(k^rh^r).\tag{20}$$

The proof is by induction and given in Appendix B. Inserting this history expansion (20) into the main term in the right-hand side of (19) and carefully keeping track of the error gives

$$\begin{aligned}&\nabla L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n-k)}\big)\\&\quad=\sum_{m=0}^{\mathcal{R}-1}h^m\sum_{\substack{i,l\geqslant 0\\ i+l=m}}\sum_{(i_0,\dots,i_l)\in\mathcal{K}_{i,l}}\frac{1}{i_0!\dots i_l!}\nabla^{i+1}L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)\Big(\underbrace{\widetilde{\boldsymbol{d}}_1^{(n,k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)}_{i_0\text{ times}},\dots,\underbrace{\widetilde{\boldsymbol{d}}_{l+1}^{(n,k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)}_{i_l\text{ times}}\Big)\\&\qquad+O(k^{\mathcal{R}}h^{\mathcal{R}}).\end{aligned}$$

It is left to sum this over k with an exponentially decaying weight β^k. The error $O(k^{\mathcal{R}}h^{\mathcal{R}})$ is polynomial in k but it is not a problem because it will turn into $\sum_{k=0}^{n}\beta^k O(k^{\mathcal{R}}h^{\mathcal{R}})=O(h^{\mathcal{R}})$ by exponential summation. The coefficients $\boldsymbol{d}_m^{(n)}\big(\boldsymbol{\theta}\big)$ are defined in such a way that the result is exactly (18) that we need, proving the local error bound (16). A standard argument (Lemma B.3) for converting a local error bound to a global error bound gives (17).

4 Analyzing the Form of Memoryless Iteration Coefficients

Let us write the first few terms $\boldsymbol{d}_m^{(n)}(\boldsymbol{\theta})$ from (14) and describe the pattern that can be used to generate them. To declutter notation, we will omit the argument $\boldsymbol{\theta}$ since it will be fixed. By definition, the first memoryless iteration coefficient is just the exponential average of past gradients:

$$\boldsymbol{d}_1^{(n)}=-\sum_{b=0}^{n}\beta^b\nabla L^{(n-b)}.$$

Using the definition, we can also write the following expression for $\boldsymbol{d}_2^{(n)}$:

$$\boldsymbol{d}_2^{(n)}=-\beta\sum_{b=0}^{n-1}\beta^b\nabla^2L^{(n-1-b)}\sum_{l'=1}^{b+1}\sum_{b'=0}^{n-l'}\beta^{b'}\nabla L^{(n-l'-b')}.$$

This triple sum on the right can be generated as follows. Write the rooted tree consisting of two nodes $(1,2)$ with corresponding parents $(\varnothing,1)$. Let us introduce a variable l with value 1, which we will call the memory distance variable. Corresponding to the root 1, write the symbolic "sum"

$$\sum_{b=0}^{n-l}\beta^b\nabla^2L^{(n-l-b)}\sum_{l'=1}^{b+1}\square=\sum_{b=0}^{n-1}\beta^b\nabla^2L^{(n-1-b)}\sum_{l'=1}^{b+1}\square\tag{21}$$

and call l' the new memory distance variable (it is now being summed over, so it does not have a fixed value). We write the second derivative tensor ∇^2L (matrix in this case) because the number of children is 2; later, the order of the derivative will be $\ell+1$ where ℓ is the number of children. Let us go down the tree and consider node 2. Replace $\square$ in (21) with the corresponding expression:

$$\sum_{b'=0}^{n-l'}\beta^{b'}\nabla L^{(n-l'-b')}.$$

Again, the order of the derivative ∇L is equal to the number of children (zero) plus one. The upper limit of the sum is n minus the current memory distance variable. We do not write a trailing sum at the end like in (21) because there are no children.

We will further illustrate this process by looking at $\boldsymbol{d}_3^{(n)}$. Using (14) again and after some algebra, we can write

$$\boldsymbol{d}_3^{(n)} = -\beta \sum_{b=0}^{n-1} \beta^b \nabla^2 L^{(n-1-b)} \sum_{l'=1}^{b+1} \sum_{b'=0}^{n-l'} \beta^{b'} \nabla^2 L^{(n-l'-b')} \sum_{l''=1}^{l'+b'} \sum_{b''=0}^{n-l''} \beta^{b''} \nabla L^{(n-l''-b'')} \tag{22}$$

$$-\frac{\beta}{2} \sum_{b=0}^{n-1} \beta^b \nabla^3 L^{(n-b-1)} \Big[\sum_{l'=1}^{b+1} \sum_{b'=0}^{n-l'} \beta^{b'} \nabla L^{(n-l'-b')}, \sum_{l'=1}^{b+1} \sum_{b'=0}^{n-l'} \beta^{b'} \nabla L^{(n-l'-b')} \Big]. \tag{23}$$

There are two non-isomorphic rooted trees with 3 vertices. The first one is a "chain": the nodes $(1, 2, 3)$ have corresponding parents $(\varnothing, 1, 2)$. Let us now describe a sum that corresponds to this tree. As previously, the current memory distance is $l = 1$, and the root 1 (having one child) generates a symbolic expression

$$\sum_{b=0}^{n-1} \beta^b \nabla^2 L^{(n-l-b)} \sum_{l'=1}^{b+1} \square = \sum_{b=0}^{n-1} \beta^b \nabla^2 L^{(n-1-b)} \sum_{l'=1}^{b+1} \square.$$

The current memory distance variable is now l' (with no fixed value). Next comes node 2 with one child, generating a symbolic expression

$$\sum_{b'=0}^{n-l'} \beta^{b'} \nabla^2 L^{(n-l'-b')} \sum_{l''=1}^{l'+b'} \square.$$

The current memory distance is l'' (with no fixed value). Finally, node 3 has no children, so it closes the sum with the expression

$$\sum_{b''=0}^{n-l''} \beta^{b''} \nabla L^{(n-l''-b'')}.$$

Up to coefficient $-\beta$ in front, we have obtained (22).

The second rooted tree with 3 vertices is the tree consisting of nodes $(1, 2, 3)$ with corresponding parents $(\varnothing, 1, 1)$ (root with two children). The initial memory distance variable is denoted as l and has value 1. The root has two children, so the order of the derivative corresponding to it will be 3, and we will write two $\square$ signs corresponding to two subtrees:

$$\sum_{b=0}^{n-l} \beta^b \nabla^3 L^{(n-b-1)} \Big[\sum_{l'=1}^{b+1} \square, \sum_{l'=1}^{b+1} \square \Big].$$

Then we replace the first $\square$ with the expression corresponding to node 2 (with initial memory variable l'):

$$\sum_{b'=0}^{n-l'} \beta^{b'} \nabla L^{(n-l'-b')},$$

and the second $\square$ with the same expression corresponding to node 3. Up to coefficient $-\beta/2$ in front, we have obtained (23). The reason for the division by 2 is that this tree has a symmetry coefficient 2.

This consideration of special cases highlights a pattern in how the memoryless iteration coefficients are structured. The following result is a formalization of this pattern.

Theorem 4.1 (The form of memoryless iteration coefficients)**.** *For $m \geqslant 2$, the memoryless iteration coefficient $\boldsymbol{d}_m^{(n)}$ is equal (up to a coefficient $-\beta$) to a sum over the set $\tilde{\mathcal{A}}[m]$ of unlabeled rooted trees with m vertices:*

$$\boldsymbol{d}_m^{(n)} = -\beta \sum_{\tau \in \tilde{\mathcal{A}}[m]} \frac{1}{\sigma(\tau)} \boldsymbol{E}_{\tau,1}^{(n)},$$

where $\sigma(\tau)$ is the symmetry coefficient of the tree (background in Section 1.5). The expression $\boldsymbol{E}_{\tau,l}^{(n)} := \boldsymbol{E}_{\tau,l}^{(n)}(\beta)$, depending on the iteration number n and memory distance variable l, is defined recursively by

$$\boldsymbol{E}_{\tau,l}^{(n)} = \sum_{b=0}^{n-l} \beta^b \nabla^{\ell+1} L^{(n-l-b)} \left[\sum_{l_1=1}^{l+b} \boldsymbol{E}_{\tau_1,l_1}^{(n)}, \ldots, \sum_{l_\ell=1}^{l+b} \boldsymbol{E}_{\tau_\ell,l_\ell}^{(n)} \right], \quad l \in [1\!:\!n], \tag{24}$$

where $(\tau_1, \ldots, \tau_\ell)$ are the subtrees rooted at the children of the root of τ (with $\ell \in \mathbb{Z}_{\geqslant 0}$). In particular, $\boldsymbol{E}_{\bullet,l}^{(n)} = \sum_{b=0}^{n-l} \beta^b \nabla L^{(n-l-b)}$.

The $\boldsymbol{E}_{\tau,l}^{(n)}$ expression is an exact translation of our informal observations above into the mathematical language. For example,

$$\boldsymbol{E}_{\bullet,l}^{(n)} = \sum_{b=0}^{n-l} \beta^b \nabla^2 L^{(n-l-b)} \sum_{l_1=1}^{l+b} \sum_{b_1=0}^{n-l_1} \beta^{b_1} \nabla L^{(n-l_1-b_1)}.$$

We outline the main ideas underlying the proof next, but the formal details are deferred to Appendix C.

4.1 Proof Sketch

Before sketching the argument for establishing Theorem 4.1, we introduce some preliminary results. We employ the notation and concepts from Section 1.5.

4.1.1 Auxiliary Quantity Related to (24)

Fix a labeled rooted tree τ in $\mathcal{A}[1\!:\!m]$, and choose a marking $m \in \mathcal{M}_\tau$, where the marked vertices are $v_1, \ldots, v_{|m|}$ with corresponding subtrees $\tau_1^m, \ldots, \tau_{|m|}^m$ and the remaining subtree τ_0^m. Let $p_1, \ldots, p_{|m|}$ be the parents of the marked vertices (not necessarily distinct). Consider the derivative of the loss corresponding to p_1 in the symbolic expression for $\boldsymbol{E}_{\tau_0^m,a}^{(n-l)}$. Add one to the order of the derivative and add $\sum_{l'=1}^{l} \boldsymbol{E}_{\tau_1^m,l'}^{(n)}$ as an argument. Continue this process (possibly increasing the order of the same derivative more than once) until all marked vertices are processed. We will denote the resulting expression by

$$\boldsymbol{E}_{\tau,m,a}^{(n-l\to n)}.$$

For example, consider the tree τ consisting of the root and two leaves, with one of the leaves marked:

Consider the loss derivative corresponding to vertex $p_1 = 1$ in the symbolic expression

$$\boldsymbol{E}_{\bullet,a}^{(n-l)} = \sum_{b=0}^{n-l-a} \beta^b \underbrace{\nabla^2 L^{(n-l-b-a)}} \sum_{l_1=1}^{a+b} \sum_{b_1=0}^{n-l-l_1} \beta^{b_1} \nabla L^{(n-l-l_1-b_1)}.$$

Increase the order of the derivative and insert $\sum_{l'=1}^{l} \boldsymbol{E}_{\bullet,l'}^{(n)}$, giving

$$\sum_{b=0}^{n-l-a} \beta^b \nabla^3 L^{(n-l-b-a)} \Big[\sum_{l_1=1}^{a+b} \sum_{b_1=0}^{n-l-l_1} \beta^{b_1} \nabla L^{(n-l-l_1-b_1)}, \sum_{l'=1}^{l} \boldsymbol{E}_{\bullet,l'}^{(n)} \Big].$$

4.1.2 Useful Properties of $\boldsymbol{E}_{\tau,l}^{(n)}$

We give two lemmas about $\boldsymbol{E}_{\tau,l}^{(n)}$ that we will use in the argument. Both lemmas are proven in Appendix C. The following important fact is the reason why the induction step in the main argument goes through.

Lemma 4.2. *Let $m \geqslant 2$. For any $\tau \in \mathscr{A}[1:m]$ we have*

$$\boldsymbol{E}_{\tau,l+a}^{(n)} = \sum_{m \in \mathscr{M}_\tau} \boldsymbol{E}_{\tau,m,a}^{(n-l \to n)}. \tag{25}$$

The reason why we invoke marked trees is that differentiation naturally creates such trees. The following lemma establishes a connection between the sum over marked trees in (25) and the high-order derivative tensor which arises in the main argument for Theorem 4.1.

Lemma 4.3. *We have for $m \geqslant 2$*

$$\sum_{i=0}^{m-1} \sum_{j=1}^{m-i} \sum_{(k_0,\ldots,k_{m-j-i}) \in \mathcal{K}_{i,m-j-i}} \frac{1}{k_0! \ldots k_{m-j-i}!} \sum_{\tau_0 \in \tilde{\mathscr{A}}[j]} \frac{1}{\sigma(\tau_0)}$$
$$\times \nabla^i \boldsymbol{E}_{\tau_0,a}^{(n-l)} \Big[\underbrace{\sum_{\tau \in \tilde{\mathscr{A}}[1]} \frac{1}{\sigma(\tau)} \sum_{l'=1}^{l} \boldsymbol{E}_{\tau,l'}^{(n)}, \ldots,}_{k_0 \text{ times}} \underbrace{\sum_{\tau \in \tilde{\mathscr{A}}[m-j-i+1]} \frac{1}{\sigma(\tau)} \sum_{l'=1}^{l} \boldsymbol{E}_{\tau,l'}^{(n)}}_{k_{m-j-i} \text{ times}} \Big] = \sum_{\tau \in \tilde{\mathscr{A}}[m]} \frac{1}{\sigma(\tau)} \boldsymbol{E}_{\tau,l+a}^{(n)}.$$

Remark 4.4. Note that from the definition of $\boldsymbol{E}_{\tau,l}^{(n)}$,

$$\nabla^{\ell+1} L^{(n-l)} \Big[\sum_{l'=1}^{l} \boldsymbol{E}_{\tau_1,l'}^{(n)}, \ldots, \sum_{l'=1}^{l} \boldsymbol{E}_{\tau_\ell,l'}^{(n)} \Big] + \beta \boldsymbol{E}_{\tau,l+1}^{(n)} = \boldsymbol{E}_{\tau,l}^{(n)}. \tag{26}$$

Replacing $\boldsymbol{E}_{\tau,l+1}^{(n)}$ with $\sum_{m \in \mathscr{M}_\tau} \boldsymbol{E}_{\tau,m,1}^{(n-l \to n)}$ by Lemma 4.2 and setting $l = 1$, we get the following alternative recursion:

$$\boldsymbol{E}_{\tau,1}^{(n)} = \nabla^{\ell+1} L^{(n-1)} \big[\boldsymbol{E}_{\tau_1,1}^{(n)}, \ldots, \boldsymbol{E}_{\tau_\ell,1}^{(n)} \big] + \beta \sum_{m \in \mathscr{M}_\tau} \boldsymbol{E}_{\tau,m,1}^{(n-1 \to n)}.$$

The advantage of this form is that the memory distance variable is always 1, but the disadvantage is that the right-hand side contains a sum over all markings of τ.

4.1.3 Proof Sketch of Theorem 4.1

The strategy is to prove the following two statements simultaneously by induction over $m \geqslant 2$:

$$\boldsymbol{d}_m^{(n)} = -\beta \sum_{\tau \in \tilde{\mathscr{A}}[m]} \frac{1}{\sigma(\tau)} \boldsymbol{E}_{\tau,1}^{(n)}, \quad \text{and} \tag{27}$$

$$\widetilde{\boldsymbol{d}}_m^{(n,k)} = \sum_{l=1}^{k} \sum_{\tau \in \tilde{\mathcal{A}}[m]} \frac{1}{\sigma(\tau)} \boldsymbol{E}_{\tau,l}^{(n)}. \tag{28}$$

For $m = 2$, they are already verified above. Note also that the second statement holds for $m = 1$ as well:

$$\widetilde{\boldsymbol{d}}_1^{(n,k)} = \sum_{l=1}^{k} \boldsymbol{E}_{\bullet,l}^{(n)}.$$

By definition,

$$\boldsymbol{d}_m^{(n)} = -\beta \sum_{b=0}^{n-1} \beta^b \sum_{\ell=0}^{m-1} \sum_{(i_0,\dots,i_{m-1-\ell}) \in \mathcal{K}_{\ell,m-1-\ell}} \frac{1}{i_0! \dots i_{m-1-\ell}!} \times$$
$$\times \nabla^{\ell+1} L^{(n-1-b)} \Big(\underbrace{\widetilde{\boldsymbol{d}}_1^{(n,b+1)}}_{i_0 \text{ times}}, \dots, \underbrace{\widetilde{\boldsymbol{d}}_{m-\ell}^{(n,b+1)}}_{i_{m-1-\ell} \text{ times}} \Big)$$

Insert the induction hypothesis (recall that (28) holds for $m = 1$ too):

$$\boldsymbol{d}_m^{(n)} = -\beta \sum_{b=0}^{n-1} \beta^b \sum_{\ell=0}^{m-1} \sum_{(i_0,\dots,i_{m-1-\ell}) \in \mathcal{K}_{\ell,m-1-\ell}} \frac{1}{i_0! \dots i_{m-1-\ell}!}$$
$$\times \nabla^{\ell+1} L^{(n-1-b)} \Big(\underbrace{\sum_{l=1}^{b+1} \sum_{\tau \in \tilde{\mathcal{A}}[1]} \frac{1}{\sigma(\tau)} \boldsymbol{E}_{\tau,l}^{(n)}}_{i_0 \text{ times}}, \dots, \underbrace{\sum_{l=1}^{b+1} \sum_{\tau \in \tilde{\mathcal{A}}[m-\ell]} \frac{1}{\sigma(\tau)} \boldsymbol{E}_{\tau,l}^{(n)}}_{i_{m-1-\ell} \text{ times}} \Big)$$

Careful rearrangement is used to simplify this to

$$\boldsymbol{d}_m^{(n)} = -\beta \sum_{b=0}^{n-1} \beta^b \sum_{\ell=0}^{m-1} \sum_{\tau=[\tau_1,\dots,\tau_\ell] \in \tilde{\mathcal{A}}[m]} \frac{1}{\sigma(\tau)} \nabla^{\ell+1} L^{(n-1-b)} \Big(\sum_{l=1}^{b+1} \boldsymbol{E}_{\tau_1,l}^{(n)}, \dots, \sum_{l=1}^{b+1} \boldsymbol{E}_{\tau_\ell,l}^{(n)} \Big),$$

and that is, by definition, equal to $-\beta \sum_{\tau \in \tilde{\mathcal{A}}[m]} \frac{1}{\sigma(\tau)} \boldsymbol{E}_{\tau,1}^{(n)}$. Hence, under the induction hypothesis for smaller m, (27) holds.

By definition of the history terms in (15) and the induction hypothesis, to prove (28) and complete the induction step, it is enough to show that

$$-\sum_{j=1}^{m} \sum_{i=0}^{m-j} \sum_{(k_0,\dots,k_{m-i-j}) \in \mathcal{K}_{i,m-i-j}} \frac{1}{k_0! \dots k_{m-i-j}!}$$
$$\times \nabla^i \boldsymbol{d}_j^{(n-l)} \Big[\underbrace{\sum_{l'=1}^{l} \sum_{\tau \in \tilde{\mathcal{A}}[1]} \frac{1}{\sigma(\tau)} \boldsymbol{E}_{\tau,l'}^{(n)}}_{k_0 \text{ times}}, \dots, \underbrace{\sum_{l'=1}^{l} \sum_{\tau \in \tilde{\mathcal{A}}[m-i-j+1]} \frac{1}{\sigma(\tau)} \boldsymbol{E}_{\tau,l'}^{(n)}}_{k_{m-i-j} \text{ times}} \Big] \stackrel{?}{=} \sum_{\tau \in \tilde{\mathcal{A}}[m]} \frac{1}{\sigma(\tau)} \boldsymbol{E}_{\tau,l}^{(n)}. \tag{29}$$

By the induction hypothesis and (27) already proven, we can replace $\nabla^i \boldsymbol{d}_j^{(n-l)}$ in the left-hand side of (29) by $-\beta \nabla^i \sum_{\tau_0 \in \tilde{\mathcal{A}}[j]} \frac{1}{\sigma(\tau_0)} \boldsymbol{E}_{\tau_0,1}^{(n-l)}$. The result will involve precisely the big sum that we saw in Lemma 4.3, applying which we will simplify the left-hand side of (29) to

$$\sum_{i=0}^{m-1} \sum_{\substack{\tau \in \tilde{\mathcal{A}}[m] \\ \tau=[\tau_1,\dots,\tau_i]}} \frac{1}{\sigma(\tau)} \nabla^{i+1} L^{(n-l)} \Big[\sum_{l'=1}^{l} \boldsymbol{E}_{\tau_1,l'}^{(n)}, \dots, \sum_{l'=1}^{l} \boldsymbol{E}_{\tau_i,l'}^{(n)} \Big] + \beta \sum_{\tau \in \tilde{\mathcal{A}}[m]} \frac{1}{\sigma(\tau)} \boldsymbol{E}_{\tau,l+1}^{(n)}.$$

Combining this with (26), we see that the left-hand side of (29) is equal to the right-hand side of (29). This completes the induction step. Omitted technical details are given in Appendix C.

4.2 Connection with the Solution to the Fixed-Point Equation

The result in Theorem 4.1 can be connected to the invariant manifold perspective discussed in Sections 1.1 and 2. Recall that HB on that manifold can be rewritten as (3) where $\boldsymbol{g}_h$ satisfies the fixed point equation (11). Then, write $\boldsymbol{g}_h(\boldsymbol{\theta})$ as a formal power series [7, 27], sometimes called B-series, to obtain

$$h^2\boldsymbol{g}_h(\boldsymbol{\theta}) = \sum_{\tau\in\mathcal{A}_\varnothing} \frac{h^{|\tau|}}{|\tau|!} g(\tau)\nabla^\tau L(\boldsymbol{\theta}),$$

where $\nabla^\tau L$ is the elementary differential defined recursively by $\nabla^\varnothing L(\boldsymbol{\theta}) = \boldsymbol{\theta}$, $\nabla^\bullet L = \nabla L$ and $\nabla^\tau L = \nabla^{\ell+1} L\big[\nabla^{\tau_1} L, \dots, \nabla^{\tau_\ell} L\big]$ for $\tau = [\tau_1, \dots, \tau_\ell]$, $g\colon \tilde{\mathcal{A}}_\varnothing \to \mathbb{R}$ is the coefficient mapping (with the induced mapping $g\colon \mathcal{A}_\varnothing \to \mathbb{R}$ denoted by the same symbol), $g(\varnothing) = g(\bullet) = 0$.

In addition, define a mapping $a\colon \tilde{\mathcal{A}}_\varnothing \to \mathbb{R}$ by putting $a(\varnothing) := 1$, $a(\bullet) = -(1-\beta)^{-1}$, and $a(\tau) = \beta g(\tau)$ for $|\tau| \geqslant 2$. Then, by the composition rule (e. g. [18]), we have

$$h^2\boldsymbol{g}_h\Big(\boldsymbol{\theta} - \frac{h}{1-\beta}\nabla L(\boldsymbol{\theta}) + h^2\beta\boldsymbol{g}_h(\boldsymbol{\theta})\Big) = \sum_{\tau\in\mathcal{A}_\varnothing} \frac{h^{|\tau|}}{|\tau|!}(a*g)(\tau)\nabla^\tau L(\boldsymbol{\theta}), \tag{30}$$

where $a*g$ is the subtree convolution, that is, $(a*g)(\varnothing) = g(\varnothing) = 0$ and

$$(a*g)(\tau) = g(\varnothing)a(\tau) + \sum_{m\in M_\tau} g(\tau_0^m)a(\tau_1^m)\dots a(\tau_{|m|}^m) = \sum_{m\in M_\tau} g(\tau_0^m)a(\tau_1^m)\dots a(\tau_{|m|}^m).$$

Similarly, by the composition rule,

$$-\frac{h}{1-\beta}\nabla L\Big(\boldsymbol{\theta} - \frac{h}{1-\beta}\nabla L(\boldsymbol{\theta}) + h^2\beta\boldsymbol{g}_h(\boldsymbol{\theta})\Big) = \sum_{\tau\in\mathcal{A}_\varnothing} \frac{h^{|\tau|}}{|\tau|!}(a*l)(\tau)\nabla^\tau L(\boldsymbol{\theta}), \tag{31}$$

where $l\colon \tilde{\mathcal{A}}_\varnothing \to \mathbb{R}$ is defined by $l(\varnothing) = 0$, $l(\bullet) = -(1-\beta)^{-1}$, $l(\tau) = 0$ for $|\tau| \geqslant 2$, which means

$$(a*l)(\tau) = \sum_{m\in M_\tau} l(\tau_0^m)a(\tau_1^m)\dots a(\tau_{|m|}^m) = -\frac{1}{1-\beta}a(\tau_1)\dots a(\tau_\ell)$$

for $\tau = [\tau_1, \dots, \tau_\ell]$; in particular, $(a*l)(\bullet) = -(1-\beta)^{-1}$. Combining (30) and (31) gives

$$\begin{aligned}
&-\frac{h}{1-\beta}\nabla L\Big(\boldsymbol{\theta} - \frac{h}{1-\beta}\nabla L(\boldsymbol{\theta}) + h^2\beta\boldsymbol{g}_h(\boldsymbol{\theta})\Big) + h^2\boldsymbol{g}_h\Big(\boldsymbol{\theta} - \frac{h}{1-\beta}\nabla L(\boldsymbol{\theta}) + h^2\beta\boldsymbol{g}_h(\boldsymbol{\theta})\Big)\\
&\quad= \sum_{\tau\in\mathcal{A}_\varnothing} \frac{h^{|\tau|}}{|\tau|!}\big\{(a*l)(\tau) + (a*g)(\tau)\big\}\nabla^\tau L(\boldsymbol{\theta}).
\end{aligned}$$

By (11), this should be equal to

$$-\frac{h}{1-\beta}\nabla L(\boldsymbol{\theta}) + h^2\beta\boldsymbol{g}_h(\boldsymbol{\theta}) = -\frac{h}{1-\beta}\nabla L(\boldsymbol{\theta}) + \sum_{\tau\in\mathcal{A}_\varnothing} \frac{h^{|\tau|}}{|\tau|!}\beta g(\tau)\nabla^\tau L(\boldsymbol{\theta}).$$

Matching the coefficients before equal powers of h gives for $\tau = [\tau_1, \dots, \tau_\ell]$ with $|\tau| \geqslant 2$

$$(a*l)(\tau) + (a*g)(\tau) = \beta g(\tau),$$

that is,

$$-\frac{1}{1-\beta}a(\tau_1)\dots a(\tau_\ell)+\sum_{m\in M_\tau\setminus\{\varnothing\}} g(\tau_0^m)a(\tau_1^m)\dots a(\tau_{|m|}^m)+g(\tau)=\beta g(\tau).$$

Hence, the coefficients $g(\tau)$ satisfy the recursion

$$g(\tau)=\frac{1}{(1-\beta)^2}a(\tau_1)\dots a(\tau_\ell)-\frac{1}{1-\beta}\sum_{m\in M_\tau\setminus\{\varnothing\}} g(\tau_0^m)a(\tau_1^m)\dots a(\tau_{|m|}^m)$$

with $g(\varnothing)=g(\bullet)=0$; the same is rewritten in terms of only the a mapping as

$$a(\tau)=\frac{\beta}{(1-\beta)^2}a(\tau_1)\dots a(\tau_\ell)-\frac{1}{1-\beta}\sum_{\substack{m\in M_\tau\setminus\{\varnothing\}\\ |\tau_0^m|\geqslant 2}} a(\tau_0^m)a(\tau_1^m)\dots a(\tau_{|m|}^m)$$

with $a(\varnothing)=1$, $a(\bullet)=-\frac{1}{1-\beta}$. For example,

$$a(\begin{smallmatrix}\bullet\\ \bullet\end{smallmatrix})=-\frac{\beta}{(1-\beta)^3},\quad a(\begin{smallmatrix}\bullet\\ \bullet\\ \bullet\end{smallmatrix})=a(\begin{smallmatrix}\bullet\\ \bullet\\ \bullet\end{smallmatrix})=-\frac{\beta(1+\beta)}{(1-\beta)^5}.$$

This is a similar-looking although not quite the same characterization of $\boldsymbol{g}_h$ as we would obtain by taking $n\to\infty$ in (26) (when all losses are equal $L^{(s)}\equiv L$). Of course, they lead to the same results (despite the different recursions), for example,

$$\begin{aligned}h^2\boldsymbol{g}_h(\boldsymbol{\theta})=&-\frac{h^2}{(1-\beta)^3}\nabla^2 L(\boldsymbol{\theta})\nabla L(\boldsymbol{\theta})\\ &-\frac{h^3}{2}\frac{1+\beta}{(1-\beta)^5}\nabla^3 L(\boldsymbol{\theta})\big[\nabla L(\boldsymbol{\theta}),\nabla L(\boldsymbol{\theta})\big]-h^3\frac{1+\beta}{(1-\beta)^5}\nabla^2 L(\boldsymbol{\theta})\nabla^2 L(\boldsymbol{\theta})\nabla L(\boldsymbol{\theta})+O(h^4),\end{aligned}$$

giving the memoryless update (3)

$$\begin{aligned}\boldsymbol{\theta}^{(n+1)}=\boldsymbol{\theta}^{(n)}-\frac{h}{1-\beta}\Big\{&\nabla L(\boldsymbol{\theta}^{(n)})+\frac{h\beta}{(1-\beta)^2}\nabla^2 L(\boldsymbol{\theta}^{(n)})\nabla L(\boldsymbol{\theta}^{(n)})\\ &+\frac{h^2\beta(1+\beta)}{2(1-\beta)^4}\nabla^3 L(\boldsymbol{\theta}^{(n)})\big[\nabla L(\boldsymbol{\theta}^{(n)}),\nabla L(\boldsymbol{\theta}^{(n)})\big]\\ &+\frac{h^2\beta(1+\beta)}{(1-\beta)^4}\nabla^2 L(\boldsymbol{\theta}^{(n)})\nabla^2 L(\boldsymbol{\theta}^{(n)})\nabla L(\boldsymbol{\theta}^{(n)})+O(h^3)\Big\}.\end{aligned}$$

Here, in contrast to approximation theorems, by $O(h^3)$ we just mean terms of order h^3 and higher in the formal infinite sum.

5 Corollaries and Implications

Our main theoretical results (Theorem 3.1 and Theorem 4.1) can be used to obtain useful additional results for the analysis of HB and variants thereof. This section focuses on deriving continuous modified equations of arbitrary approximation order (with rigorous bounds), as well as principal iteration and principal flow approximations capturing the HB dynamics. Our results generalize the work in Rosca et al. [41]. Due to their practical importance, we consider both full-batch and mini-batch implementations.

5.1 Modified Equation

The global approximation by a memoryless iteration (Theorem 3.1) allows to prove the existence of a modified equation or, in other words, an approximation by a continuous flow. Define the *BEA coefficients* $\left\{\boldsymbol{f}_j^{(n)}(\boldsymbol{\theta})\right\}_{j=1}^{\infty}$ by

$$\boldsymbol{f}_j^{(n)}(\boldsymbol{\theta}) = \boldsymbol{d}_j^{(n)}(\boldsymbol{\theta}) - \sum_{i=2}^{j} \frac{1}{i!} \sum_{\substack{k_1,\dots,k_i \geqslant 1 \\ k_1+\dots+k_i=j}} (D_{k_1}^{(n)} \dots D_{k_{i-1}}^{(n)} \boldsymbol{f}_{k_i}^{(n)})(\boldsymbol{\theta}), \tag{32}$$

with the ith Lie derivative $D_i^{(n)} := \sum_l f_i^{(n)l}(\boldsymbol{\theta})\partial_l$. This formula is standard in the literature on backward error analysis applied to numerical methods [16].

We now state the continuous approximation result.

Corollary 5.1 (Modified equation)**.** *Assume the conditions of Theorem 3.1. Let $\boldsymbol{\theta}(t) \equiv \boldsymbol{\theta}(\mathcal{R}; t)$ be the unique continuous solution to the piecewise ODE in $\mathcal{D}$*

$$\dot{\boldsymbol{\theta}}(t) = \sum_{i=0}^{\mathcal{R}-1} h^i f_{i+1}^{(n)}(\boldsymbol{\theta}(t)) \tag{33}$$

on $t \in [t_n, t_{n+1}]$ with the initial condition $\boldsymbol{\theta}(0) = \boldsymbol{\theta}^{(0)}$, assumed to exist, where we use the shortcut $t_n := nh$. Then for each $h \in (0, 1/2)$

$$\sup_{n \in [0:\lfloor T/h \rfloor]} \|\boldsymbol{\theta}^{(n)} - \boldsymbol{\theta}(t_n)\| \leqslant C_3 h^{\mathcal{R}},$$

where C_3 is some constant depending on T.

The proof is by Taylor-expanding $\boldsymbol{\theta}(t)$ around t_n at t_{n+1} and rearranging; this allows to get the local error bound, which is then converted into the global error bound. Full details are provided in Appendix D.1.

For example, the second-order approximation is

$$\dot{\boldsymbol{\theta}}(t) = -\sum_{b=0}^{n} \beta^b \nabla L^{(n-b)}(\boldsymbol{\theta}(t)) + h\boldsymbol{f}_2^{(n)}(\boldsymbol{\theta}(t)) \tag{34}$$

with

$$\begin{aligned}\boldsymbol{f}_2^{(n)}(\boldsymbol{\theta}) = &-\sum_{b=1}^{n} \beta^b \nabla^2 L^{(n-b)}(\boldsymbol{\theta}) \sum_{l=1}^{b} \sum_{b'=0}^{n-l} \beta^{b'} \nabla L^{(n-l-b')}(\boldsymbol{\theta}) \\ &- \frac{1}{2}\nabla \sum_{b=0}^{n} \beta^b \nabla^2 L^{(n-b)} \sum_{b'=0}^{n} \beta^{b'} \nabla L^{(n-b')}(\boldsymbol{\theta}).\end{aligned}$$

In the simpler full-batch case (where all $L^{(s)} \equiv L$) (34) is rewritten as

$$\begin{aligned}\dot{\boldsymbol{\theta}}(t) = &-\frac{1-\beta^{n+1}}{1-\beta}\nabla L(\boldsymbol{\theta}(t)) \\ &- \frac{1+\beta-4(n+1)(\beta^{n+1}-\beta^{n+2})-\beta^{2n+2}-\beta^{2n+3}}{2(1-\beta)^3}\nabla^2 L(\boldsymbol{\theta}(t))\nabla L(\boldsymbol{\theta}(t)).\end{aligned}$$

on the segment $t \in [nh, (n+1)h]$. As n becomes large, this trajectory turns into the smooth solution to

$$\dot{\boldsymbol{\theta}}(t) = -\frac{1}{1-\beta}\nabla L(\boldsymbol{\theta}(t)) - \frac{1+\beta}{2(1-\beta)^3}\nabla^2 L(\boldsymbol{\theta}(t))\nabla L(\boldsymbol{\theta}(t)). \tag{35}$$

The equation found in Kovachki and Stuart [34] (after fixing a small typo) is the same ODE after neglecting β^n, but for (35) the guarantee (8) (with $\mathcal{R} = 2$) is only true if the initialization happened exactly on the attractive manifold. Importantly, we did not lose any information but gained a guarantee regardless of initialization.

Remark 5.2. Note that (35) can be rewritten as

$$\dot{\boldsymbol{\theta}}(t) = -\frac{1}{1-\beta}\nabla\left\{L + \frac{1+\beta}{4(1-\beta)^2}\|\nabla L\|^2\right\}(\boldsymbol{\theta}(t)).$$

This can be seen as gradient flow with a modified loss [17, 34, 22]. However, for $\mathcal{R} = 3$ this is already not true: $\boldsymbol{f}_3^{(n)}(\boldsymbol{\theta})$ becomes close to

$$-\frac{1+4\beta+\beta^2}{3(1-\beta)^5}\nabla^2 L(\boldsymbol{\theta})\nabla^2(\boldsymbol{\theta})\nabla L(\boldsymbol{\theta}) - \frac{1+10\beta+\beta^2}{12(1-\beta)^5}\nabla^3 L(\boldsymbol{\theta})\big[\nabla L(\boldsymbol{\theta}), \nabla L(\boldsymbol{\theta})\big],$$

which is in general not a gradient.

5.2 Principal Iteration

Consider full-batch HB (2). Let us formally take $\mathcal{R}$ to infinity in (5) and write a formal series $\sum_{m=1}^{\infty} h^m \boldsymbol{d}_m^{(n)}(\boldsymbol{\theta})$. It is an infinite sum of terms that are of two types: terms containing only derivatives of order no higher than two of the loss, which we will call *principal terms*, and the remaining terms (containing derivatives of order at least three), which we will call *non-principal* ones. For example, the term in (22) is principal and the term in (23) is non-principal.

Write

$$\sum_{m=2}^{\infty} h^m \boldsymbol{d}_m^{(n)}(\boldsymbol{\theta}) = -\beta \sum_{m=2}^{\infty} v_m^{(n)} h^m \{\nabla^2 L(\boldsymbol{\theta})\}^{m-1}\nabla L(\boldsymbol{\theta}) + \mathrm{NPT},$$

where the notation NPT means "non-principal terms", $\{v_m^{(n)}\}$ are coefficients (not depending on h). Theorem 4.1 and its proof give an easy way to write down a recursion for $v_m^{(n)}$.

Corollary 5.3 (Principal iteration)**.** *Define $\{v_m^{(n)}\}_{m=1}^{\infty}$ by putting $v_1^{(n)} = \sum_{b=0}^{n-1}\beta^b$ and so that the principal part of $\boldsymbol{d}_m^{(n)}(\boldsymbol{\theta})$ is $-\beta v_m^{(n)}\{\nabla^2 L(\boldsymbol{\theta})\}^{m-1}\nabla L(\boldsymbol{\theta})$ for $m \geqslant 2$. Then the coefficients $v_m^{(n)}$ satisfy*

$$v_m^{(n)} = v_{m-1}^{(n)} + \beta\sum_{j=1}^{m-1} v_j^{(n-1)} v_{m-j}^{(n)} + \beta v_m^{(n-1)}, \quad m \geqslant 2 \tag{36}$$

and there are limits $v_m^{(\infty)} := \lim_{n\to\infty} v_m^{(n)}$, which satisfy $v_1^{(\infty)} = (1-\beta)^{-1}$ and

$$v_m^{(\infty)} = \frac{v_{m-1}^{(\infty)}}{1-\beta} + \frac{\beta}{1-\beta}\sum_{j=1}^{m-1} v_j^{(\infty)} v_{m-j}^{(\infty)}, \quad m \geqslant 2. \tag{37}$$

The generating function $g_\beta(x) := \sum_{m=0}^{\infty} v_{m+1}^{(\infty)} x^m$ is given by

$$g_\beta(x) = \frac{1-\beta-x-\sqrt{(1-\beta-x)^2 - 4\beta x}}{2\beta x}. \tag{38}$$

In particular, for $m \geqslant 1$

$$v_{m+1}^{(\infty)} = \frac{N_m(\beta)}{(1-\beta)^{2m+1}},$$

where

$$N_m(\beta) := \sum_{k=1}^{m} \frac{1}{m}\binom{m}{k}\binom{m}{k-1}\beta^{m-k}, \quad m \geqslant 1$$

are the Narayana polynomials.

The proof is given in Appendix D.2.

Informally, Corollary 5.3 means

$$\begin{aligned}\sum_{m=1}^{\infty} h^m \boldsymbol{d}_m^{(n)}(\boldsymbol{\theta}) &\approx -\frac{h}{1-\beta}\nabla L(\boldsymbol{\theta}) - \beta \sum_{m=1}^{\infty} v_{m+1}^{(\infty)} h^{m+1}\{\nabla^2 L(\boldsymbol{\theta})\}^m \nabla L(\boldsymbol{\theta}) + \text{NPT} \\ &= -h\nabla L(\boldsymbol{\theta}) - \beta h \sum_{m=0}^{\infty} v_{m+1}^{(\infty)} h^m \{\nabla^2 L(\boldsymbol{\theta})\}^m \nabla L(\boldsymbol{\theta}) + \text{NPT} \\ &= -h\nabla L(\boldsymbol{\theta}) - \beta h g_\beta(h\nabla^2 L(\boldsymbol{\theta}))\nabla L(\boldsymbol{\theta}) + \text{NPT},\end{aligned}$$

where $\approx$ hides the fact that n is taken to infinity. Therefore, the "full-order" memoryless iteration is approximately

$$\tilde{\boldsymbol{\theta}}^{(n+1)} = \tilde{\boldsymbol{\theta}}^{(n)} - h\nabla L\big(\tilde{\boldsymbol{\theta}}^{(n)}\big) - \beta h g_\beta\big(h\nabla^2 L\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)\big)\nabla L\big(\tilde{\boldsymbol{\theta}}^{(n)}\big) + \text{NPT} \tag{39}$$

for large n. The series $g_\beta\big(h\nabla^2 L\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)\big)$ converges in Euclidean operator norm when $\big\|\nabla^2 L\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)\big\| < R_\beta/h$, where $R_\beta = (1-\sqrt{\beta})^2$ is the convergence radius.

Remark 5.4. We thank Boris Hanin for the following interesting observation: $g_\beta(x)$ is the Stieltjes transform of the standard Marchenko-Pastur law with parameter β, which we can write as

$$g_\beta(x) = \mathbb{E}_{\xi\sim\mathrm{MP}(\beta)}[(\xi - x)^{-1}].$$

Hence, (39) becomes

$$\tilde{\boldsymbol{\theta}}^{(n+1)} = \tilde{\boldsymbol{\theta}}^{(n)} - h\mathbb{E}_{\xi\sim\mathrm{MP}(\beta)}\big[\boldsymbol{I} + \beta\big(\xi\boldsymbol{I} - h\nabla^2 L\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)\big)^{-1}\big]\nabla L\big(\tilde{\boldsymbol{\theta}}^{(n)}\big) + \text{NPT}$$

as long as $\|\nabla^2 L\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)\| < (1-\sqrt{\beta})^2/h$ as above.

5.3 Comments on Combinatorics

We see from Corollary 5.3 that the coefficients corresponding to $\boldsymbol{E}_{\tau,1}^{(n)}$, where τ is a chain with m vertices (see Section 1.5 for the definition of a chain), are the rescaled Narayana polynomials.

Remark 5.5. Corollary 5.3 and (69) show in particular that the Narayana polynomials can be defined as $N_m(\beta) \equiv N_{m,1}(\beta)$ where $\{N_{m,l}(\beta)\}$ satisfy the recursion

$$N_{m,l}(\beta) = (1-\beta)^2 \sum_{b=0}^{\infty} \beta^b \sum_{l_1=1}^{l+b} N_{m-1,l_1}(\beta), \quad m \geqslant 2, l \geqslant 1$$

with initial condition $N_{1,l}(\beta) = \beta + (1-\beta)l$ for $l \geqslant 1$. We are not aware of this characterization in the literature.

Let us now write the coefficient before another type of trees, namely, the trees consisting only of the root and a number of leaves.

Corollary 5.6. *Define $q^{(n)}_{m,l}$ as such coefficients that*

$$\boldsymbol{E}^{(n)}_{\mathfrak{r}_m,l} = q^{(n)}_{m,l}\nabla^m L(\boldsymbol{\theta})\big[\underbrace{\nabla L(\boldsymbol{\theta}),\dots,\nabla L(\boldsymbol{\theta})}_{m-1 \text{ times}}\big]$$

in (24) in the full-batch case, where $\mathfrak{r}_m$ consists of a root and $m-1$ leaves. Then the limit $q^{(\infty)}_{m,1} = \lim_{n\to\infty} q^{(n)}_{m,1}$ satisfies

$$q^{(\infty)}_{m+1,1} = \frac{1}{(1-\beta)^{2m+1}} A_m(\beta), \quad m \geqslant 1,$$

where

$$A_m(\beta) = (1-\beta)^{m+1}\sum_{j=1}^{\infty} j^m \beta^{j-1}$$

are the Eulerian polynomials.

The proof is given in Appendix D.3.

Corollaries 5.3 and 5.6 motivate the following definition.

Definition 5.7. For $\tau = [\tau_1,\dots,\tau_\ell] \in \tilde{\mathscr{A}}[m]$, let $e_{\tau,l}$ be the coefficient before $\boldsymbol{E}^{(n)}_{\tau,l}$ in Theorem 4.1 after taking $n\to\infty$ and multiplying by $(1-\beta)^{2m-1}$, that is,

$$e_{\tau,l} \equiv e_{\tau,l}(\beta) = (1-\beta)^{\ell+1}\sum_{b=0}^{\infty}\beta^b \sum_{l_1=1}^{l+b} e_{\tau_1,l_1}\dots\sum_{l_\ell=1}^{l+b} e_{\tau_\ell,l_\ell}, \quad l\in\mathbb{Z}_{\geqslant 1}, \quad \ell\in\mathbb{Z}_{\geqslant 0}.$$

In particular, $e_{\bullet,l} = 1$.

By induction, $e_{\tau,l}$ is a polynomial of degree no more than $m-1$ in the variable $l(1-\beta)$ with coefficients that are themselves polynomials *only* of β (not depending on l). In particular, $e_{\tau,l}$ is a polynomial in β (as opposed to just a rational function).

By Corollaries 5.3 and 5.6, $(1-\beta)^{2m-1}v^{(\infty)}_m = e_{c_m,1}$ where c_m is the chain with m vertices; $(1-\beta)^{2m-1}q^{(\infty)}_{m,l} = e_{\mathfrak{r}_m,l}$ where $\mathfrak{r}_m$ is the tree consisting of a root and $m-1$ leaves, whereas

$$e_{c_m,1} = N_{m-1}(\beta), \quad e_{\mathfrak{r}_m,1} = A_{m-1}(\beta).$$

To conclude, $e_{\tau,1}\equiv e_{\tau,1}(\beta)$ in Definition 5.7 form a rich $\tilde{\mathscr{A}}[m]$-parametrized ($m\geqslant 1$) family of polynomials of β, containing both the Narayana polynomials and the Eulerian polynomials, and many other polynomials "in-between". This combinatorial digression may be of independent interest.

5.4 Principal Flow

Using the same framework as in Section 5.2, we can derive a "full-order" modified equation up to non-principal terms.

Corollary 5.8 (Principal flow)**.** *Define $\{z^{(n)}_m\}$ as such coefficients that the principal part of $\boldsymbol{f}^{(n)}_m(\boldsymbol{\theta})$ is $z^{(n)}_m\{\nabla^2 L(\boldsymbol{\theta})\}^{m-1}\nabla L(\boldsymbol{\theta})$. Then,*

$$z^{(n)}_m := \sum_{l=1}^{m}\frac{(-1)^{l+1}}{l}\sum_{\substack{k_1,\dots,k_l\geqslant 1\\ k_1+\dots+k_l=m}} p^{(n)}_{k_1}\dots p^{(n)}_{k_l}, \tag{40}$$

where

$$p_k^{(n)} := \begin{cases} -v_k^{(n+1)}, & \text{if } k = 1, \\ -\beta v_k^{(n)}, & \text{if } k \geqslant 2. \end{cases}$$

Moreover, the limiting sequence $z_m^{(\infty)} := \lim_{n\to\infty} z_m^{(n)}$ *admits a generating function* $\bar{g}_\beta(x) := \sum_{k=0}^{\infty} z_{k+1}^{(\infty)} x^k$ *given by*

$$\bar{g}_\beta(x) = \frac{1}{x} \ln\left(\frac{1+\beta-x+\sqrt{(1-\beta-x)^2-4\beta x}}{2} \right).$$

The proof is given in Appendix D.4.

Informally, the "full-order" modified equation (called "principal flow" by Rosca et al. [41]) is approximately

$$\begin{aligned} \dot{\boldsymbol{\theta}}(t) &= \sum_{k=0}^{\infty} z_{k+1}^{(n)} h^k \{\nabla^2 L(\boldsymbol{\theta}(t))\}^k \nabla L(\boldsymbol{\theta}(t)) + \text{NPT} \\ &= \bar{g}_\beta\big(h\nabla^2 L(\boldsymbol{\theta}(t))\big) \nabla L(\boldsymbol{\theta}(t)) + \text{NPT} \end{aligned}$$

for large n. Taking $\beta = 0$, we recover the result of [41] (for GD).

For example, consider the one-dimensional case and $L(\theta) = \theta^2/2$. The HB iteration (1) is solved by

$$\begin{pmatrix} \theta^{(n)} \\ v^{(n)} \end{pmatrix} = \begin{pmatrix} 1-h & h\beta \\ -1 & \beta \end{pmatrix}^n \begin{pmatrix} \theta^{(0)} \\ v^{(0)}. \end{pmatrix}$$

The general solution is

$$\begin{aligned} \theta^{(n)} &= \frac{((1-h-\lambda_-)\theta^{(0)} + h\beta v^{(0)})\lambda_+^n + ((\lambda_+ + h - 1)\theta^{(0)} - h\beta v^{(0)})\lambda_-^n}{\lambda_+ - \lambda_-}, \\ v^{(n)} &= \frac{\big((\beta-\lambda_-)v^{(0)} - \theta^{(0)}\big)\lambda_+^n + \big(\theta^{(0)} + (\lambda_+ - \beta)v^{(0)}\big)\lambda_-^n}{\lambda_+ - \lambda_-} \end{aligned}$$

with

$$\lambda_\pm = \frac{1+\beta-h \pm \sqrt{(1-\beta-h)^2 - 4\beta h}}{2},$$

taking for simplicity the case $\lambda_+ \neq \lambda_-$ and $\lambda_\pm \in \mathbb{R}$. If the initialization happened on the invariant manifold we discussed above (attractive if h is small enough), that is, $v^{(0)} = \theta^{(0)}(\lambda_+ + h - 1)/(h\beta)$, then

$$\theta^{(n)} = \lambda_+^n \theta^{(0)}. \tag{41}$$

The solution of the principal flow $\dot{\theta}(t) = \bar{g}_\beta\big(h\big)\theta(t)$ in this case is

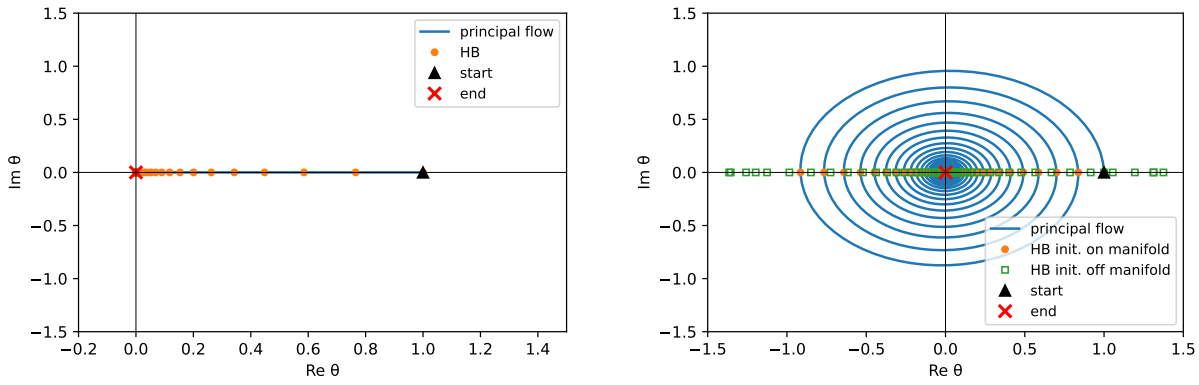
$$\theta(t) = \theta(0) \exp\left\{ \frac{t}{h} \ln\left(\frac{1+\beta-h+\sqrt{(1-\beta-h)^2-4\beta h}}{2} \right) \right\}, \tag{42}$$

coinciding with (41) at points $t = nh$.

We note in passing that for large step sizes $(1+\sqrt{\beta})^2 < h < 2+2\beta$ there is another attractive manifold of importance $v/\theta = (h + \lambda_- - 1)/(h\beta)$, where the right-hand side blows up as $h \to 0$. If the initialization happened there, the solution will be

$$\theta^{(n)} = \lambda_-^n \theta^{(0)}.$$

This also coincides with principal flow (42) by choosing appropriate values of the complex square root and logarithm. We illustrate both situations in Fig. 1.



(a) Small $h = 0.02$

(b) Large $h = 3.38$

Figure 1: Principal flow ($\beta = 0.7$, $\theta^{(0)} = 1$) corresponding to the quadratic loss $L(\theta) = \theta^2/2$ on the complex plane. It is purely real for small enough step size h. For high step sizes, it is not purely real, and the real values $\theta(nh)$ capture the oscillatory behavior of HB (with HB exactly matching the flow if initialized on the attractive invariant manifold).

6 Concluding Remarks

We studied the theoretical properties of gradient descent with Polyak [38] heavy-ball momentum, the simplest and one of the most commonly used algorithm with memory in optimization. We established an approximation of the algorithm by a memoryless iteration with arbitrary precision. In the full-batch case, this memoryless iteration is (roughly speaking) plain gradient descent with a modified loss. This loss modification can be seen as implicit regularization by memory, and can sometimes explain good empirical performance [2, 44, 22]. In the stochastic (mini-batch) case, additional implicit regularization by mini-batch noise can be identified. These insights can be of practical importance in machine learning, not just for understanding existing algorithms, but also for proposing new ones, e. g. by introducing (or strengthening) similar regularization explicitly [21, 51]. In addition, analyzing the terms of the memoryless iteration revealed rich combinatorics hidden inside the algorithm, which may be of independent theoretical interest.

Bernstein and Newhouse [4] notes that "the precise role of EMA [exponential moving averages] is perhaps still an open problem" in optimization. Even though we tailored the presentation to be specifically about HB, one of the strengths of our techniques is that they can be used to study other algorithms with decaying memory (defined by smooth enough functions of the parameter $\boldsymbol{\theta}$). Thus, our paper not only makes a step in the large task of theoretically understanding memory and its effects for a specific optimization algorithm, but also outlines a more general framework that can be used to analyze other algorithms. For example, our techniques could be used to study other (ubiquitous) numerical methods with memory such as Adam [33] or Shampoo [25].

A Proof of Theorem 2.1

A.1 Constants

It will be convenient to define

$$K_1 := D_1/(1-\beta) + h\beta\gamma, \tag{43}$$
$$K_2 := D_2/(1-\beta) + h\beta\delta. \tag{44}$$

The constants γ, δ and λ are chosen as follows: γ is as in [34, Lemma 10], that is, $\gamma \in [\tau_1, \infty)$ with

$$\tau_1 := \frac{(1-\beta)^{-2} D_1 D_2}{1-\beta - hD_2\beta/(1-\beta)},$$

δ is chosen at the end of the proof of Lemma A.2: namely, it needs to be positive and satisfy (48), and λ is chosen to lie in $[\tau_2(\delta), \infty)$ with

$$\begin{aligned}\tau_2(\delta) :=& \Bigg(\frac{D_3(1-\beta)^{-1}[(1-\beta)^{-1}D_2 + h\beta\delta]}{1 - h\{(1-\beta)^{-1}D_2 + h\beta\delta\}} \\ &+ \frac{(1-\beta)^{-1}[D_4K_1 + D_3\{(1-\beta)^{-1}D_2 + h\beta\delta\}]}{\big(1 - h\{(1-\beta)^{-1}D_2 + h\beta\delta\}\big)^2} \\ &+ \frac{\big((1-\beta)^{-1}D_2 + h\beta\delta\big)\big[(1-\beta)^{-1}D_3\big]}{\big(1 - h\{(1-\beta)^{-1}D_2 + h\beta\delta\}\big)^3} \Bigg) \\ &\times \left(1 - \frac{\beta}{\big(1 - h\{(1-\beta)^{-1}D_2 + h\beta\delta\}\big)^2} - \frac{h\beta[(1-\beta)^{-1}D_2 + h\beta\delta]}{\big(1 - h\{(1-\beta)^{-1}D_2 + h\beta\delta\}\big)^3} \right)^{-1}.\end{aligned}$$

Notice that $\tau_2(\delta) > 0$ for small enough h.

A.2 Omitted Lemmas

Lemma A.1 ($\boldsymbol{\theta} \mapsto \boldsymbol{\zeta}$ is bijective)**.** *For small enough h and for any $\boldsymbol{g} \in \Gamma$, the function $\boldsymbol{\theta} \mapsto \boldsymbol{\zeta} = \boldsymbol{\theta} - h(1-\beta)^{-1}\nabla L(\boldsymbol{\theta}) + h^2\beta\boldsymbol{g}(\boldsymbol{\theta})$ is a bijection $\mathbb{R}^d \to \mathbb{R}^d$.*

Proof. This is proven in [34]. □

Lemma A.2 (T maps from Γ to Γ)**.** *For small enough h, the operator T defined in (12) maps from Γ to Γ.*

Proof. Take $\boldsymbol{g} \in \Gamma$. By Lemma A.1, the function $\boldsymbol{\theta} \mapsto \boldsymbol{\zeta} = \boldsymbol{\theta} - h(1-\beta)^{-1}\nabla L(\boldsymbol{\theta}) + h^2\beta\boldsymbol{g}(\boldsymbol{\theta})$ is a bijection. This function is continuously differentiable:

$$\nabla_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\boldsymbol{\theta}) = I - \frac{h}{1-\beta}\nabla^2 L(\boldsymbol{\theta}) + h^2\beta\nabla\boldsymbol{g}(\boldsymbol{\theta}).$$

For h small enough, this matrix is invertible. By the inverse mapping theorem, we see that the inverse mapping is continuously differentiable with Jacobian

$$\nabla_{\boldsymbol{\zeta}}\boldsymbol{\theta}(\boldsymbol{\zeta}) = \left(I - \frac{h}{1-\beta}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta})) + h^2\beta\nabla\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta})) \right)^{-1}. \tag{45}$$

Therefore, the function $T\boldsymbol{g}$ is also continuously differentiable with derivative

$$\nabla_{\boldsymbol{\zeta}} T\boldsymbol{g}(\boldsymbol{\zeta}) = \frac{1}{h(1-\beta)}\nabla_{\boldsymbol{\zeta}}\big\{\nabla L(\boldsymbol{\zeta}) - \nabla L(\boldsymbol{\theta}(\boldsymbol{\zeta}))\big\} + \beta\nabla_{\boldsymbol{\zeta}}\big\{\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\big\}$$

$$
\begin{aligned}
&= \frac{1}{h(1-\beta)}\left\{\nabla^2 L(\boldsymbol{\zeta}) - \nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}))\nabla_{\boldsymbol{\zeta}}\boldsymbol{\theta}(\boldsymbol{\zeta})\right\} + \beta\nabla \boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\nabla_{\boldsymbol{\zeta}}\boldsymbol{\theta}(\boldsymbol{\zeta}) \\
&= \frac{1}{h(1-\beta)}\left\{\nabla^2 L(\boldsymbol{\zeta}) - \nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}))\left(I - \frac{h}{1-\beta}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta})) + h^2\beta\nabla \boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\right)^{-1}\right\} \\
&\quad + \beta\nabla \boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\left(I - \frac{h}{1-\beta}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta})) + h^2\beta\nabla \boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\right)^{-1} \\
&= \frac{1}{h(1-\beta)}\nabla^2 L(\boldsymbol{\zeta}) \\
&\quad + \frac{1}{h^2}\left(-\frac{h}{1-\beta}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta})) + h^2\beta\nabla \boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\right)\left(I - \frac{h}{1-\beta}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta})) + h^2\beta\nabla \boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\right)^{-1} \\
&= \frac{\nabla^2 L(\boldsymbol{\zeta}) - \nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}))}{h(1-\beta)} + \beta\nabla \boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta})) + \frac{1}{h^2}\left(\frac{h}{1-\beta}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta})) - h^2\beta\nabla \boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\right) \\
&\quad - \frac{1}{h^2}\left(\frac{h}{1-\beta}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta})) - h^2\beta\nabla \boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\right)\left(I - \frac{h}{1-\beta}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta})) + h^2\beta\nabla \boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\right)^{-1} \\
&= \frac{\nabla^2 L(\boldsymbol{\zeta}) - \nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}))}{h(1-\beta)} + \beta\nabla \boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta})) \\
&\quad + \frac{1}{h^2}\left(\frac{h}{1-\beta}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta})) - h^2\beta\nabla \boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\right) \\
&\qquad\qquad \times\left\{I - \left(I - \frac{h}{1-\beta}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta})) + h^2\beta\nabla \boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\right)^{-1}\right\}. \qquad (46)
\end{aligned}
$$

This matrix is symmetric for any $\boldsymbol{g} \in \Gamma$ because any square matrix A commutes with $(I-A)^{-1}$ (provided the latter exists). We have proven that $T\boldsymbol{g}$ is a continuously differentiable function and its Jacobian is symmetric.

Bounding $T\boldsymbol{g}$ We need to show that the norm of the function $T\boldsymbol{g}$ does not exceed γ for any $g \in \Gamma$. We have for any $\boldsymbol{\zeta} \in \mathbb{R}^d$

$$
\|T\boldsymbol{g}(\boldsymbol{\zeta})\| \leqslant \frac{\sup_{\boldsymbol{\theta}}\|\nabla^2 L(\boldsymbol{\theta})\|}{h(1-\beta)}\|\boldsymbol{\zeta} - \boldsymbol{\theta}(\boldsymbol{\zeta})\| + \beta\|\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\|
$$

(where we used the definition of T in Eq. (12))

$$
\begin{aligned}
&= \frac{\sup_{\boldsymbol{\theta}}\|\nabla^2 L(\boldsymbol{\theta})\|}{h(1-\beta)}\left\|-\frac{h}{1-\beta}\nabla L(\boldsymbol{\theta}(\boldsymbol{\zeta})) + h^2\beta\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\right\| + \beta\|\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\| \\
&\leqslant \frac{\sup_{\boldsymbol{\theta}}\|\nabla^2 L(\boldsymbol{\theta})\|\sup_{\boldsymbol{\theta}}\|\nabla L(\boldsymbol{\theta})\|}{(1-\beta)^2} + \left(\beta + h\frac{\beta}{1-\beta}\sup_{\boldsymbol{\theta}}\|\nabla^2 L(\boldsymbol{\theta})\|\right)\gamma \\
&= \frac{D_1 D_2}{(1-\beta)^2} + \left(\beta + h\frac{\beta}{1-\beta}D_2\right)\gamma \\
&\leqslant \gamma, \qquad (47)
\end{aligned}
$$

where the last inequality follows from the definition of τ_1 and since γ was taken to be at least τ_1.

Bounding the derivative of $T\boldsymbol{g}$ Consider the expression ending in (46). We will use

$$
\|\nabla_{\boldsymbol{\zeta}} T\boldsymbol{g}(\boldsymbol{\zeta})\| \leqslant \left\|\frac{\nabla^2 L(\boldsymbol{\zeta}) - \nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}))}{h(1-\beta)}\right\|
$$

$$+\left\|\frac{1}{h(1-\beta)}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}))\left\{I-\left(I-\frac{h}{1-\beta}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}))+h^2\beta\nabla\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\right)^{-1}\right\}\right\|$$
$$+\left\|\beta\nabla\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\left(I-\frac{h}{1-\beta}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}))+h^2\beta\nabla\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\right)^{-1}\right\|$$

and bound each of the three terms in the right-hand side separately.

The first term is easy to bound:

$$\begin{aligned}\left\|\frac{\nabla^2 L(\boldsymbol{\zeta})-\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}))}{h(1-\beta)}\right\| &\leqslant \frac{D_3}{h(1-\beta)}\|\boldsymbol{\zeta}-\boldsymbol{\theta}(\boldsymbol{\zeta})\| \\ &= \frac{D_3}{h(1-\beta)}\left\|-\frac{h}{1-\beta}\nabla L(\boldsymbol{\theta}(\boldsymbol{\zeta}))+h^2\beta\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\right\| \\ &\leqslant \frac{D_3}{h(1-\beta)}\left\{\frac{hD_1}{1-\beta}+h^2\beta\gamma\right\}=\frac{D_3K_1}{1-\beta}.\end{aligned}$$

Now we bound

$$\begin{aligned}&\left\|\frac{1}{h(1-\beta)}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}))\left\{I-\left(I-\frac{h}{1-\beta}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}))+h^2\beta\nabla\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\right)^{-1}\right\}\right\| \\ &\quad\leqslant \frac{D_2}{h(1-\beta)}\left\|I-\left(I-\frac{h}{1-\beta}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}))+h^2\beta\nabla\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\right)^{-1}\right\|.\end{aligned}$$

Notice that the eigenvalues of the matrix

$$I-\left(I-\frac{h}{1-\beta}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}))+h^2\beta\nabla\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\right)^{-1}$$

are of the form $-h\lambda_i/(1-h\lambda_i)$ where $\{\lambda_i\}$ are eigenvalues of the matrix $(1-\beta)^{-1}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}))-h\beta\nabla\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))$. For small enough h, we have $h\max_i|\lambda_i|<1$, so we can bound

$$\begin{aligned}\frac{|\lambda_i|}{|1-h\lambda_i|} &\leqslant \frac{|\lambda_i|}{1-h|\lambda_i|} \leqslant \frac{\max_i|\lambda_i|}{1-h\max_i|\lambda_i|}=\frac{\|(1-\beta)^{-1}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}))+h\beta\nabla\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\|}{1-h\|(1-\beta)^{-1}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}))+h\beta\nabla\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\|} \\ &\leqslant \frac{(1-\beta)^{-1}D_2+h\beta\delta}{1-h\{(1-\beta)^{-1}D_2+h\beta\delta\}}.\end{aligned}$$

Similarly,

$$\left\|\beta\nabla\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\left(I-\frac{h}{1-\beta}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}))+h^2\beta\nabla\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\right)^{-1}\right\| \leqslant \frac{\beta\delta}{1-h\{(1-\beta)^{-1}D_2+h\beta\delta\}}.$$

We conclude that

$$\begin{aligned}\|\nabla_{\boldsymbol{\zeta}}T\boldsymbol{g}(\boldsymbol{\zeta})\| &\leqslant \frac{D_3K_1}{1-\beta}+\frac{D_2}{1-\beta}\frac{(1-\beta)^{-1}D_2+h\beta\delta}{1-h\{(1-\beta)^{-1}D_2+h\beta\delta\}}+\frac{\beta\delta}{1-h\{(1-\beta)^{-1}D_2+h\beta\delta\}} \\ &= \frac{D_3K_1\big(1-h\big\{\frac{D_2}{1-\beta}+h\beta\delta\big\}\big)+D_2\big(\frac{D_2}{1-\beta}+h\beta\delta\big)+\beta\delta(1-\beta)}{(1-\beta)[1-h\{(1-\beta)^{-1}D_2+h\beta\delta\}]} \\ &= \frac{(D_2-hD_3K_1)\big(\frac{D_2}{1-\beta}+h\beta\delta\big)+D_3K_1+\beta\delta(1-\beta)}{(1-\beta)[1-h\{(1-\beta)^{-1}D_2+h\beta\delta\}]}.\end{aligned}$$

The inequality

$$\frac{(D_2-hD_3K_1)\big(\frac{D_2}{1-\beta}+h\beta\delta\big)+D_3K_1+\beta\delta(1-\beta)}{(1-\beta)[1-h\{(1-\beta)^{-1}D_2+h\beta\delta\}]} \leqslant \delta \tag{48}$$

can be rewritten as

$$h^2\beta\delta^2 + \left(\beta - 1 + h\frac{1+\beta}{1-\beta}D_2 - \frac{h^2\beta}{1-\beta}K_1D_3\right)\delta + \frac{D_2(D_2 - hD_3K_1)}{(1-\beta)^2} + \frac{D_3K_1}{1-\beta} \leqslant 0.$$

For small enough h, we can ensure that the quadratic equation in the left-hand side has a positive root, which means we can take any $\delta > 0$ slightly less than this root, and the inequality will hold.

Bounding the Lipschitz constant of the derivative of Tg Consider the expression ending in (46). To declutter notation, denote $A_i := (1-\beta)^{-1}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}_i)) - h\beta\nabla\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}_i))$ for $i \in \{1, 2\}$. First, decompose

$$\begin{aligned}
\nabla_{\boldsymbol{\zeta}} T\boldsymbol{g}(\boldsymbol{\zeta}_1) - \nabla_{\boldsymbol{\zeta}} T\boldsymbol{g}(\boldsymbol{\zeta}_2) &= \frac{\nabla^2 L(\boldsymbol{\zeta}_1)}{h(1-\beta)} - \left(\frac{\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}_1))}{h(1-\beta)} - \beta\nabla\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}_1))\right)(I - hA_1)^{-1} \\
&\quad - \frac{\nabla^2 L(\boldsymbol{\zeta}_2)}{h(1-\beta)} + \left(\frac{\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}_2))}{h(1-\beta)} - \beta\nabla\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}_2))\right)(I - hA_2)^{-1} \\
&= \frac{\nabla^2 L(\boldsymbol{\zeta}_1) - \nabla^2 L(\boldsymbol{\zeta}_2)}{h(1-\beta)} \\
&\quad - \left(\frac{\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}_1)) - \nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}_2))}{h(1-\beta)} - \beta\nabla\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}_1)) + \beta\nabla\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}_2))\right)(I - hA_1)^{-1} \\
&\quad - \left(\frac{\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}_2))}{h(1-\beta)} - \beta\nabla\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}_2))\right)\{(I - hA_1)^{-1} - (I - hA_2)^{-1}\} \\
&= T_1 - T_2 - T_3,
\end{aligned}$$

where

$$\begin{aligned}
T_1 &:= \frac{\nabla^2 L(\boldsymbol{\zeta}_1) - \nabla^2 L(\boldsymbol{\zeta}_2)}{h(1-\beta)}\{I - (I - hA_1)^{-1}\}, \\
T_2 &:= \left(\frac{\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}_1)) - \nabla^2 L(\boldsymbol{\zeta}_1) - \nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}_2)) + \nabla^2 L(\boldsymbol{\zeta}_2)}{h(1-\beta)} - \beta\nabla\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}_1)) + \beta\nabla\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}_2))\right)\times \\
&\quad \times (I - hA_1)^{-1}, \\
T_3 &:= \left(\frac{\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}_2))}{h(1-\beta)} - \beta\nabla\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}_2))\right)\{(I - hA_1)^{-1} - (I - hA_2)^{-1}\}.
\end{aligned}$$

To bound T_1, use $\|I - (I - hA_1)^{-1}\| \leqslant h\frac{(1-\beta)^{-1}D_2 + h\beta\delta}{1 - h\{(1-\beta)^{-1}D_2 + h\beta\delta\}}$ and conclude

$$\|T_1\| \leqslant \frac{D_3(1-\beta)^{-1}[(1-\beta)^{-1}D_2 + h\beta\delta]}{1 - h\{(1-\beta)^{-1}D_2 + h\beta\delta\}}\|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2\|.$$

To bound T_2, note that

$$\begin{aligned}
\nabla_{\boldsymbol{\zeta}}\{\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta})) - \nabla^2 L(\boldsymbol{\zeta})\} &= \nabla^3 L(\boldsymbol{\theta}(\boldsymbol{\zeta}))(I - hA)^{-1} - \nabla^3 L(\boldsymbol{\zeta}) \\
&= \big(\nabla^3 L(\boldsymbol{\theta}(\boldsymbol{\zeta})) - \nabla^3 L(\boldsymbol{\zeta})\big)(I - hA)^{-1} + \nabla^3 L(\boldsymbol{\zeta})\big((I - hA)^{-1} - I\big)
\end{aligned}$$

where $A := (1-\beta)^{-1}\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta})) - h\beta\nabla\boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}))$. Therefore,

$$\begin{aligned}
&\|\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}_1)) - \nabla^2 L(\boldsymbol{\zeta}_1) - \nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}_2)) + \nabla^2 L(\boldsymbol{\zeta}_2)\| \\
&\quad \leqslant \sup_{\boldsymbol{\zeta}}\big\|\nabla_{\boldsymbol{\zeta}}\{\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta})) - \nabla^2 L(\boldsymbol{\zeta})\}\big\|\|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2\|
\end{aligned}$$

$$\leqslant \frac{D_4 \sup_{\boldsymbol{\zeta}} \|\boldsymbol{\theta}(\boldsymbol{\zeta}) - \boldsymbol{\zeta}\|}{1 - h\{(1-\beta)^{-1} D_2 + h\beta\delta\}} \|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2\| + h \frac{D_3\{(1-\beta)^{-1} D_2 + h\beta\delta\}}{1 - h\{(1-\beta)^{-1} D_2 + h\beta\delta\}} \|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2\|$$
$$\leqslant h \frac{D_4 K_1 + D_3\{(1-\beta)^{-1} D_2 + h\beta\delta\}}{1 - h\{(1-\beta)^{-1} D_2 + h\beta\delta\}} \|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2\|,$$

where the last inequality is obtained by bounding $\sup_{\boldsymbol{\zeta}} \|\boldsymbol{\theta}(\boldsymbol{\zeta}) - \boldsymbol{\zeta}\|$ like for Eq. (47). Since also

$$\beta \|\nabla \boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}_1)) - \nabla \boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}_2))\| \leqslant \beta\lambda \|\boldsymbol{\theta}(\boldsymbol{\zeta}_1) - \boldsymbol{\theta}(\boldsymbol{\zeta}_2)\| \leqslant \beta\lambda \sup_{\boldsymbol{\zeta}} \|\nabla_{\boldsymbol{\zeta}} \boldsymbol{\theta}(\boldsymbol{\zeta})\| \|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2\|$$
$$\leqslant \frac{\beta\lambda}{1 - h\{(1-\beta)^{-1} D_2 + h\beta\delta\}} \|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2\|,$$

we conclude

$$\|T_2\| \leqslant \frac{(1-\beta)^{-1}[D_4 K_1 + D_3\{(1-\beta)^{-1} D_2 + h\beta\delta\}] + \beta\lambda}{\big(1 - h\{(1-\beta)^{-1} D_2 + h\beta\delta\}\big)^2} \|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2\|.$$

To bound T_3, use

$$\|(I - hA_1)^{-1} - (I - hA_2)^{-1}\| = h \|(I - hA_1)^{-1}[A_1 - A_2](I - hA_2)^{-1}\|$$
$$\leqslant \frac{h}{\big(1 - h\{(1-\beta)^{-1} D_2 + h\beta\delta\}\big)^2}$$
$$\times \left\| \frac{\nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}_1)) - \nabla^2 L(\boldsymbol{\theta}(\boldsymbol{\zeta}_2))}{1-\beta} - h\beta\big\{\nabla \boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}_1)) - \nabla \boldsymbol{g}(\boldsymbol{\theta}(\boldsymbol{\zeta}_2))\big\} \right\|$$
$$\leqslant \frac{h}{\big(1 - h\{(1-\beta)^{-1} D_2 + h\beta\delta\}\big)^2} \left[\frac{D_3}{1-\beta} + h\beta\lambda \right] \|\boldsymbol{\theta}(\boldsymbol{\zeta}_1) - \boldsymbol{\theta}(\boldsymbol{\zeta}_2)\|$$
$$\leqslant \frac{h\big[(1-\beta)^{-1} D_3 + h\beta\lambda\big]}{\big(1 - h\{(1-\beta)^{-1} D_2 + h\beta\delta\}\big)^3} \|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2\|,$$

and conclude

$$\|T_3\| \leqslant \frac{\big((1-\beta)^{-1} D_2 + h\beta\delta\big)\big[(1-\beta)^{-1} D_3 + h\beta\lambda\big]}{\big(1 - h\{(1-\beta)^{-1} D_2 + h\beta\delta\}\big)^3} \|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2\|.$$

Combining, we get

$$\|\nabla_{\boldsymbol{\zeta}} T\boldsymbol{g}(\boldsymbol{\zeta}_1) - \nabla_{\boldsymbol{\zeta}} T\boldsymbol{g}(\boldsymbol{\zeta}_2)\|$$
$$\leqslant \Bigg(\frac{D_3(1-\beta)^{-1}[(1-\beta)^{-1} D_2 + h\beta\delta]}{1 - h\{(1-\beta)^{-1} D_2 + h\beta\delta\}}$$
$$+ \frac{(1-\beta)^{-1}[D_4 K_1 + D_3\{(1-\beta)^{-1} D_2 + h\beta\delta\}] + \beta\lambda}{\big(1 - h\{(1-\beta)^{-1} D_2 + h\beta\delta\}\big)^2}$$
$$+ \frac{\big((1-\beta)^{-1} D_2 + h\beta\delta\big)\big[(1-\beta)^{-1} D_3 + h\beta\lambda\big]}{\big(1 - h\{(1-\beta)^{-1} D_2 + h\beta\delta\}\big)^3} \Bigg) \|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2\|.$$

It is left to note that

$$\frac{D_3(1-\beta)^{-1}[(1-\beta)^{-1} D_2 + h\beta\delta]}{1 - h\{(1-\beta)^{-1} D_2 + h\beta\delta\}}$$
$$+ \frac{(1-\beta)^{-1}[D_4 K_1 + D_3\{(1-\beta)^{-1} D_2 + h\beta\delta\}] + \beta\lambda}{\big(1 - h\{(1-\beta)^{-1} D_2 + h\beta\delta\}\big)^2}$$

$$+\frac{\big((1-\beta)^{-1}D_2+h\beta\delta\big)\big[(1-\beta)^{-1}D_3+h\beta\lambda\big]}{\big(1-h\{(1-\beta)^{-1}D_2+h\beta\delta\}\big)^3}\leqslant\lambda$$

for small enough h because λ was chosen to be no less than $\tau_2(\delta)$. □

Lemma A.3 (T is a contraction)**.** *For any $\boldsymbol{g}_1,\boldsymbol{g}_2\in\Gamma$, we have*

$$\|T\boldsymbol{g}_1-T\boldsymbol{g}_2\|_\Gamma\leqslant\alpha\|\boldsymbol{g}_1-\boldsymbol{g}_2\|_\Gamma$$

with some $\alpha<1$.

Proof. It is proven in [34, Lemma 16] that for h small enough $hK_2<1$ and

$$\sup_{\boldsymbol{\zeta}\in\mathbb{R}^d}\|T\boldsymbol{g}_1(\boldsymbol{\zeta})-T\boldsymbol{g}_2(\boldsymbol{\zeta})\|\leqslant\alpha_1\sup_{\boldsymbol{\theta}\in\mathbb{R}^d}\|\boldsymbol{g}_1(\boldsymbol{\theta})-\boldsymbol{g}_2(\boldsymbol{\theta})\| \tag{49}$$

with

$$\alpha_1:=\beta+h\beta\frac{D_2+hD_3K_1}{1-\beta}+\frac{h^2\beta}{1-hK_2}\Big(\beta\delta+\frac{(D_2+hD_3K_1)(D_2/(1-\beta)+h\beta\delta)+D_3K_1}{1-\beta}\Big).$$

Next, fix $\boldsymbol{\zeta}\in\mathbb{R}^d$, $\boldsymbol{g}_1,\boldsymbol{g}_2\in\Gamma$. Let $\boldsymbol{\theta}_1$ be the preimage of $\boldsymbol{\zeta}$ under the mapping $\boldsymbol{\theta}\mapsto\boldsymbol{\zeta}_1(\boldsymbol{\theta})=\boldsymbol{\theta}-h(1-\beta)^{-1}\nabla L(\boldsymbol{\theta})+h^2\beta\boldsymbol{g}_1(\boldsymbol{\theta})$, and $\boldsymbol{\theta}_2$ the preimage of $\boldsymbol{\zeta}$ under the mapping $\boldsymbol{\theta}\mapsto\boldsymbol{\zeta}_2(\boldsymbol{\theta})=\boldsymbol{\theta}-h(1-\beta)^{-1}\nabla L(\boldsymbol{\theta})+h^2\beta\boldsymbol{g}_2(\boldsymbol{\theta})$. Recall that the form of the derivative of $T\boldsymbol{g}_i(\boldsymbol{\zeta})$ is given in (46). Decompose

$$\begin{aligned}
&\nabla_{\boldsymbol{\zeta}}T\boldsymbol{g}_1(\boldsymbol{\zeta})-\nabla_{\boldsymbol{\zeta}}T\boldsymbol{g}_2(\boldsymbol{\zeta})\\
&\quad=\left(\frac{\nabla^2L(\boldsymbol{\theta}_2)}{h(1-\beta)}-\beta\nabla\boldsymbol{g}_2(\boldsymbol{\theta}_2)\right)(I-hA_2)^{-1}-\left(\frac{\nabla^2L(\boldsymbol{\theta}_1)}{h(1-\beta)}-\beta\nabla\boldsymbol{g}_1(\boldsymbol{\theta}_1)\right)(I-hA_1)^{-1}\\
&\quad=\underbrace{\left[\frac{\nabla^2L(\boldsymbol{\theta}_2)-\nabla^2L(\boldsymbol{\theta}_1)}{h(1-\beta)}-\beta\big\{\nabla\boldsymbol{g}_2(\boldsymbol{\theta}_2)-\nabla\boldsymbol{g}_1(\boldsymbol{\theta}_1)\big\}\right](I-hA_2)^{-1}}_{T_1}\\
&\qquad+\underbrace{\left[\frac{\nabla^2L(\boldsymbol{\theta}_1)}{h(1-\beta)}-\beta\nabla\boldsymbol{g}_1(\boldsymbol{\theta}_1)\right]\big((I-hA_2)^{-1}-(I-hA_1)^{-1}\big)}_{T_2},
\end{aligned}$$

where $A_i=(1-\beta)^{-1}\nabla^2L(\boldsymbol{\theta}_i)-h\beta\nabla\boldsymbol{g}_i(\boldsymbol{\theta}_i)$, $i\in\{1,2\}$.

To bound T_1, first note

$$\left\|\frac{\nabla^2L(\boldsymbol{\theta}_1)-\nabla^2L(\boldsymbol{\theta}_2)}{h(1-\beta)}\right\|\leqslant\frac{D_3}{h(1-\beta)}\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|. \tag{50}$$

Since

$$\boldsymbol{\zeta}=\boldsymbol{\theta}_1-\frac{h}{1-\beta}\nabla L(\boldsymbol{\theta}_1)+h^2\beta\boldsymbol{g}_1(\boldsymbol{\theta}_1)=\boldsymbol{\theta}_2-\frac{h}{1-\beta}\nabla L(\boldsymbol{\theta}_2)+h^2\beta\boldsymbol{g}_2(\boldsymbol{\theta}_2),$$

we have

$$\begin{aligned}
\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|&\leqslant\frac{h}{1-\beta}\|\nabla L(\boldsymbol{\theta}_1)-\nabla L(\boldsymbol{\theta}_2)\|+h^2\beta\|\boldsymbol{g}_1(\boldsymbol{\theta}_1)-\boldsymbol{g}_2(\boldsymbol{\theta}_2)\|\\
&\leqslant\frac{hD_2}{1-\beta}\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|+h^2\beta\|\boldsymbol{g}_1(\boldsymbol{\theta}_1)-\boldsymbol{g}_1(\boldsymbol{\theta}_2)\|+h^2\beta\|\boldsymbol{g}_1(\boldsymbol{\theta}_2)-\boldsymbol{g}_2(\boldsymbol{\theta}_2)\|\\
&\leqslant\left(\frac{hD_2}{1-\beta}+h^2\beta\delta\right)\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|+h^2\beta\|\boldsymbol{g}_1(\boldsymbol{\theta}_2)-\boldsymbol{g}_2(\boldsymbol{\theta}_2)\|.
\end{aligned}$$

For h small enough, $\frac{hD_2}{1-\beta} + h^2\beta\delta < 1$, and we obtain

$$\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \leqslant \frac{h^2\beta}{1 - \frac{hD_2}{1-\beta} - h^2\beta\delta}\|\boldsymbol{g}_1(\boldsymbol{\theta}_2) - \boldsymbol{g}_2(\boldsymbol{\theta}_2)\|. \tag{51}$$

Continuing (50), we get

$$\begin{aligned}\left\|\frac{\nabla^2 L(\boldsymbol{\theta}_1) - \nabla^2 L(\boldsymbol{\theta}_2)}{h(1-\beta)}\right\| &\leqslant \frac{D_3}{h(1-\beta)}\frac{h^2\beta}{1 - \frac{hD_2}{1-\beta} - h^2\beta\delta}\sup_{\boldsymbol{\theta}}\|\boldsymbol{g}_1(\boldsymbol{\theta}) - \boldsymbol{g}_2(\boldsymbol{\theta})\| \\ &= \frac{(1-\beta)^{-1}h\beta D_3}{1 - \frac{hD_2}{1-\beta} - h^2\beta\delta}\sup_{\boldsymbol{\theta}}\|\boldsymbol{g}_1(\boldsymbol{\theta}) - \boldsymbol{g}_2(\boldsymbol{\theta})\|.\end{aligned}$$

Next, use

$$\begin{aligned}\|\nabla\boldsymbol{g}_1(\boldsymbol{\theta}_1) - \nabla\boldsymbol{g}_2(\boldsymbol{\theta}_2)\| &\leqslant \|\nabla\boldsymbol{g}_1(\boldsymbol{\theta}_1) - \nabla\boldsymbol{g}_1(\boldsymbol{\theta}_2)\| + \|\nabla\boldsymbol{g}_1(\boldsymbol{\theta}_2) - \nabla\boldsymbol{g}_2(\boldsymbol{\theta}_2)\| \\ &\leqslant \lambda\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| + \sup_{\boldsymbol{\theta}}\|\nabla\boldsymbol{g}_1(\boldsymbol{\theta}) - \nabla\boldsymbol{g}_2(\boldsymbol{\theta})\| \\ &\leqslant \frac{h^2\beta\lambda}{1 - \frac{hD_2}{1-\beta} - h^2\beta\delta}\sup_{\boldsymbol{\theta}}\|\boldsymbol{g}_1(\boldsymbol{\theta}) - \boldsymbol{g}_2(\boldsymbol{\theta})\| + \sup_{\boldsymbol{\theta}}\|\nabla\boldsymbol{g}_1(\boldsymbol{\theta}) - \nabla\boldsymbol{g}_2(\boldsymbol{\theta})\|\end{aligned}$$

with the last inequality by Eq. (51). As previously, using $\|I - hA_1\|^{-1} \leqslant [1 - h\{(1-\beta)^{-1}D_2 + h\beta\delta\}]^{-1}$, we can now conclude

$$\begin{aligned}\|T_1\| \leqslant &\left(\frac{(1-\beta)^{-1}h\beta D_3}{\left(1 - \frac{hD_2}{1-\beta} - h^2\beta\delta\right)^2} + \frac{h^2\beta^2\lambda}{\left(1 - \frac{hD_2}{1-\beta} - h^2\beta\delta\right)^2}\right)\sup_{\boldsymbol{\theta}}\|\boldsymbol{g}_1(\boldsymbol{\theta}) - \boldsymbol{g}_2(\boldsymbol{\theta})\| \\ &+ \frac{\beta}{1 - \frac{hD_2}{1-\beta} - h^2\beta\delta}\sup_{\boldsymbol{\theta}}\|\nabla\boldsymbol{g}_1(\boldsymbol{\theta}) - \nabla\boldsymbol{g}_2(\boldsymbol{\theta})\|.\end{aligned}$$

To bound T_2, use

$$\begin{aligned}&\left\|(I - hA_2)^{-1} - (I - hA_1)^{-1}\right\| = h\|(I - hA_2)^{-1}[A_2 - A_1](I - hA_1)^{-1}\| \\ &\leqslant \frac{h}{\left(1 - h\{(1-\beta)^{-1}D_2 + h\beta\delta\}\right)^2}\|A_2 - A_1\| \\ &= \frac{h}{\left(1 - h\{(1-\beta)^{-1}D_2 + h\beta\delta\}\right)^2}\left\|\frac{\nabla^2 L(\boldsymbol{\theta}_2) - \nabla^2 L(\boldsymbol{\theta}_1)}{1-\beta} - h\beta\{\nabla\boldsymbol{g}_2(\boldsymbol{\theta}_2) - \nabla\boldsymbol{g}_1(\boldsymbol{\theta}_1)\}\right\| \\ &\leqslant \frac{h}{\left(1 - h\{(1-\beta)^{-1}D_2 + h\beta\delta\}\right)^2}\left[\frac{(1-\beta)^{-1}h^2\beta D_3 + h^3\beta^2\lambda}{1 - \frac{hD_2}{1-\beta} - h^2\beta\delta}\sup_{\boldsymbol{\theta}}\|\boldsymbol{g}_1(\boldsymbol{\theta}) - \boldsymbol{g}_2(\boldsymbol{\theta})\|\right. \\ &\left. + h\beta\sup_{\boldsymbol{\theta}}\|\nabla\boldsymbol{g}_1(\boldsymbol{\theta}) - \nabla\boldsymbol{g}_2(\boldsymbol{\theta})\|\right] \\ &\leqslant \frac{(1-\beta)^{-1}h^3\beta D_3 + h^4\beta^2\lambda}{\left(1 - h\{(1-\beta)^{-1}D_2 + h\beta\delta\}\right)^3}\sup_{\boldsymbol{\theta}}\|\boldsymbol{g}_1(\boldsymbol{\theta}) - \boldsymbol{g}_2(\boldsymbol{\theta})\| \\ &+ \frac{h^2\beta}{\left(1 - h\{(1-\beta)^{-1}D_2 + h\beta\delta\}\right)^2}\sup_{\boldsymbol{\theta}}\|\nabla\boldsymbol{g}_1(\boldsymbol{\theta}) - \nabla\boldsymbol{g}_2(\boldsymbol{\theta})\|,\end{aligned}$$

and conclude

$$\|T_2\| \leqslant \frac{\left((1-\beta)^{-1}D_2 + h\beta\delta\right)\left[(1-\beta)^{-1}h^2\beta D_3 + h^3\beta^2\lambda\right]}{\left(1 - h\{(1-\beta)^{-1}D_2 + h\beta\delta\}\right)^3}\sup_{\boldsymbol{\theta}}\|\boldsymbol{g}_1(\boldsymbol{\theta}) - \boldsymbol{g}_2(\boldsymbol{\theta})\|$$

$$+\frac{\big((1-\beta)^{-1}D_2+h\beta\delta\big)h\beta}{\big(1-h\{(1-\beta)^{-1}D_2+h\beta\delta\}\big)^2}\sup_{\boldsymbol{\theta}}\|\nabla \boldsymbol{g}_1(\boldsymbol{\theta})-\nabla \boldsymbol{g}_2(\boldsymbol{\theta})\|.$$

Combining the bounds on T_1 and T_2 gives

$$\big\|\nabla_{\boldsymbol{\zeta}} T\boldsymbol{g}_1(\boldsymbol{\zeta})-\nabla_{\boldsymbol{\zeta}} T\boldsymbol{g}_2(\boldsymbol{\zeta})\big\| \leqslant \tilde{\alpha}_1 \sup_{\boldsymbol{\theta}}\|\boldsymbol{g}_1(\boldsymbol{\theta})-\boldsymbol{g}_2(\boldsymbol{\theta})\|+\tilde{\alpha}_2\sup_{\boldsymbol{\theta}}\|\nabla \boldsymbol{g}_1(\boldsymbol{\theta})-\nabla \boldsymbol{g}_2(\boldsymbol{\theta})\|,$$

where

$$\begin{aligned}
\tilde{\alpha}_1 &:= \frac{(1-\beta)^{-1}h\beta D_3}{\big(1-h\{(1-\beta)^{-1}D_2+h\beta\delta\}\big)^2}+\frac{h^2\beta^2\lambda}{\big(1-h\{(1-\beta)^{-1}D_2+h\beta\delta\}\big)^2}\\
&\quad+\frac{\big((1-\beta)^{-1}D_2+h\beta\delta\big)\big[(1-\beta)^{-1}h^2\beta D_3+h^3\beta^2\lambda\big]}{\big(1-h\{(1-\beta)^{-1}D_2+h\beta\delta\}\big)^3},\\
&= \frac{\beta h\{D_3/(1-\beta)+h\beta\lambda\}}{(1-hK_2)^3},\\
\tilde{\alpha}_2 &:= \frac{\beta}{1-h\{(1-\beta)^{-1}D_2+h\beta\delta\}}+\frac{\big((1-\beta)^{-1}D_2+h\beta\delta\big)h\beta}{\big(1-h\{(1-\beta)^{-1}D_2+h\beta\delta\}\big)^2}=\frac{\beta}{(1-hK_2)^2}.
\end{aligned}$$

Combining with (49), we get

$$\begin{aligned}
&\sup_{\boldsymbol{\zeta}}\|T\boldsymbol{g}_1(\boldsymbol{\zeta})-T\boldsymbol{g}_2(\boldsymbol{\zeta})\|+\sup_{\boldsymbol{\zeta}}\big\|\nabla_{\boldsymbol{\zeta}} T\boldsymbol{g}_1(\boldsymbol{\zeta})-\nabla_{\boldsymbol{\zeta}} T\boldsymbol{g}_2(\boldsymbol{\zeta})\big\|\\
&\quad\leqslant(\alpha_1+\tilde{\alpha}_1)\sup_{\boldsymbol{\theta}}\|\boldsymbol{g}_1(\boldsymbol{\theta})-\boldsymbol{g}_2(\boldsymbol{\theta})\|+\tilde{\alpha}_2\sup_{\boldsymbol{\theta}}\|\nabla \boldsymbol{g}_1(\boldsymbol{\theta})-\nabla \boldsymbol{g}_2(\boldsymbol{\theta})\|.
\end{aligned}$$

For h small enough $(\alpha_1+\tilde{\alpha}_1)\vee\tilde{\alpha}_2<1$. □

Lemma A.4 (Exponential attractivity)**.** *Equation* (1) *satisfies*

$$\left\|\boldsymbol{v}^{(n)}+\frac{1}{1-\beta}\nabla L(\boldsymbol{\theta}^{(n)})-h\boldsymbol{g}_h(\boldsymbol{\theta}^{(n)})\right\|\leqslant(\beta+h^2\beta\delta)^n\left\|\boldsymbol{v}^{(0)}+\frac{1}{1-\beta}\nabla L(\boldsymbol{\theta}^{(0)})-h\boldsymbol{g}_h(\boldsymbol{\theta}^{(0)})\right\|.$$

Proof. The same argument as in [34, Theorem 5] proves this result as well. □

B Proof of Theorem 3.1

In the proof of this theorem, $O(\cdot)$ will denote a term that is bounded (in absolute value or in norm) by a constant times the argument, where the constant cannot depend on n or h but can depend on other fixed values (like $\mathcal{R}$, T, β).

Lemma B.1. *For all $\boldsymbol{\theta}\in\mathcal{D}$ and all $m\in[1:\mathcal{R}]$, we have*

$$\nabla^r \boldsymbol{d}_m^{(n)}(\boldsymbol{\theta})=O(1),\quad \nabla^r\widetilde{\boldsymbol{d}}_m^{(n,a)}(\boldsymbol{\theta})=O(a^m),\quad r\in[0:2\mathcal{R}-m]. \tag{52}$$

Proof. Note that if $\widetilde{\boldsymbol{d}}_{l+1}^{(n,s)}(\boldsymbol{\theta})$ actually appears in the sum on the right of (15), then $l+1$ cannot exceed $m-1$, because $l+1\geqslant m$ would mean $i\leqslant 0$ and this term is actually absent (there are no history terms if $i=0$). Therefore, $\widetilde{\boldsymbol{d}}_m^{(n,a)}(\boldsymbol{\theta})$ is defined by $\big\{\widetilde{\boldsymbol{d}}_j^{(n,s)}(\boldsymbol{\theta})\big\}_{1\leqslant j\leqslant m-1,1\leqslant s\leqslant a}$ and $\big\{\boldsymbol{d}_j^{(n-s)}(\boldsymbol{\theta})\big\}_{1\leqslant j\leqslant m}$. By the same logic, from (14), $\boldsymbol{d}_m^{(n)}(\boldsymbol{\theta})$ is defined through $\big\{\widetilde{\boldsymbol{d}}_j^{(n,k)}(\boldsymbol{\theta})\big\}_{1\leqslant j\leqslant m-1,1\leqslant k\leqslant n}$.

So, we prove (52) by induction in m. For $m=1$ it is easy to check. To prove that the statement holds for m assuming that it holds for all $m'<m$, we notice that since there is

exponential averaging in the definition of $\boldsymbol{d}_m^{(n)}(\boldsymbol{\theta})$ and the derivatives of the mini-batch losses are bounded, we have

$$|\boldsymbol{d}_m^{(n)}(\boldsymbol{\theta})| \lesssim \sum_{k=1}^{n} \beta^k \sum_{\substack{i,l\geqslant 0\\ i+l=m-1}} \sum_{(i_0,\dots,i_l)\in\mathcal{K}_{i,l}} k^{i_0+\dots+(l+1)i_l}$$
$$\lesssim \sum_{k=1}^{n} \beta^k \sum_{\substack{i,l\geqslant 0\\ i+l=m-1}} k^{i+l} \lesssim \sum_{k=1}^{n} \beta^k k^{m-1} = O(1).$$

Adding ∇^r in front of $\boldsymbol{d}_m^{(n)}(\boldsymbol{\theta})$ influences only the number of terms in the sum (but it remains bounded) and constant coefficients in the bounds above (the induction assumption is chosen such that we can still apply it if we add derivatives to the history terms), so the statement $\nabla^r \boldsymbol{d}_m^{(n)}(\boldsymbol{\theta}) = O(1)$ also holds.

Using this and bounds on the previous history terms (and their derivatives), we can also see

$$|\widetilde{\boldsymbol{d}}_m^{(n,a)}(\boldsymbol{\theta})| \lesssim \sum_{s=1}^{a} \sum_{\substack{j\geqslant 1,i,l\geqslant 0\\ i+j+l=m}} \sum_{(k_0,\dots,k_l)\in\mathcal{K}_{i,l}} s^{k_0+\dots+(l+1)k_l}$$
$$\lesssim \sum_{s=1}^{a} \sum_{\substack{j\geqslant 1,i,l\geqslant 0\\ i+j+l=m}} s^{i+l} \lesssim \sum_{s=1}^{a} (1+\dots+s^{m-1}) \lesssim \sum_{s=1}^{a} s^{m-1} \lesssim a^m.$$

Again, adding ∇^r in front of $\widetilde{\boldsymbol{d}}_m^{(n,a)}(\boldsymbol{\theta})$ only influences the number of terms in the sum (though it remains bounded) and constant coefficients in these bounds. This proves $\nabla^r \widetilde{\boldsymbol{d}}_m^{(n,a)}(\boldsymbol{\theta}) = O(a^m)$, completing the induction step. □

Lemma B.2. *We have $\tilde{\boldsymbol{\theta}}^{(n-k)} - \tilde{\boldsymbol{\theta}}^{(n)} = O(kh)$ for any $k \in [1\,{:}\,n]$.*

Proof. This follows from the fact that the right-hand side in (5) is $O(h)$, because $\boldsymbol{d}_j^{(n)}(\boldsymbol{\theta}) = O(1)$ by Lemma B.1 and

$$\sum_{j=0}^{\mathcal{R}-1} h^j = \frac{1-h^{\mathcal{R}}}{1-h} \leqslant \frac{1}{1-h} \overset{\text{(a)}}{=} O(1),$$

where (a) is because h is bounded away from 1. □

Recall that, by definition, (16) will follow from the following bound (18) on the error introduced by removing memory. By Taylor's theorem, we have

$$\nabla L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n-k)}\big) = \sum_{i=0}^{\mathcal{R}-1} \frac{1}{i!} \nabla^{i+1} L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)\big(\underbrace{\tilde{\boldsymbol{\theta}}^{(n-k)} - \tilde{\boldsymbol{\theta}}^{(n)}, \dots, \tilde{\boldsymbol{\theta}}^{(n-k)} - \tilde{\boldsymbol{\theta}}^{(n)}}_{i \text{ times}}\big) + \mathrm{Rem}^{(n-k)},$$

where

$$\mathrm{Rem}^{(n-k)} = \frac{1}{(\mathcal{R}-1)!} \int_0^1 (1-t)^{\mathcal{R}-1} \times \nabla^{\mathcal{R}+1} L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n)} + t\big(\tilde{\boldsymbol{\theta}}^{(n-k)} - \tilde{\boldsymbol{\theta}}^{(n)}\big)\big)\big(\underbrace{\tilde{\boldsymbol{\theta}}^{(n-k)} - \tilde{\boldsymbol{\theta}}^{(n)}, \dots, \tilde{\boldsymbol{\theta}}^{(n-k)} - \tilde{\boldsymbol{\theta}}^{(n)}}_{\mathcal{R} \text{ times}}\big)\,\mathrm{d}t.$$

Using the boundedness of derivatives of $L^{(n-k)}(\cdot)$ and Lemma B.2, we can write

$$\nabla L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n-k)}\big) = \sum_{i=0}^{\mathcal{R}-1} \frac{1}{i!} \nabla^{i+1} L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)\big(\underbrace{\tilde{\boldsymbol{\theta}}^{(n-k)} - \tilde{\boldsymbol{\theta}}^{(n)}, \ldots, \tilde{\boldsymbol{\theta}}^{(n-k)} - \tilde{\boldsymbol{\theta}}^{(n)}}_{i \text{ times}}\big) + O(k^{\mathcal{R}} h^{\mathcal{R}}). \tag{53}$$

We will now express the difference $\tilde{\boldsymbol{\theta}}^{(n-k)} - \tilde{\boldsymbol{\theta}}^{(n)}$ in (53) through $\tilde{\boldsymbol{\theta}}^{(n)}$.

Proof of Lemma 3.2. We induct over r.

For $r = 1$ this is the assertion of Lemma B.2. Further, $r \geqslant 2$.

Now make the induction assumption that (20) holds with r replaced by $r - j$ for any $j \in [1 : r-1]$.

Like in the proof of Lemma B.2, using $\boldsymbol{d}_r^{(n)}(\boldsymbol{\theta}) = O(1)$ by Lemma B.1 and the fact that h is bounded away from 1, we write for $s \in [1 : k]$

$$\tilde{\boldsymbol{\theta}}^{(n-s+1)} = \tilde{\boldsymbol{\theta}}^{(n-s)} + \sum_{j=1}^{r-1} h^j \boldsymbol{d}_j^{(n-s)}\big(\tilde{\boldsymbol{\theta}}^{(n-s)}\big) + O(h^r).$$

Inserting this, we get

$$\begin{aligned}
\tilde{\boldsymbol{\theta}}^{(n-k)} - \tilde{\boldsymbol{\theta}}^{(n)} &= \sum_{s=1}^{k} \big\{\tilde{\boldsymbol{\theta}}^{(n-s)} - \tilde{\boldsymbol{\theta}}^{(n-s+1)}\big\} \\
&= -\sum_{s=1}^{k} \Big\{\sum_{j=1}^{r-1} h^j \boldsymbol{d}_j^{(n-s)}\big(\tilde{\boldsymbol{\theta}}^{(n-s)}\big) + O(h^r)\Big\} \\
&= -\sum_{s=1}^{k} \sum_{j=1}^{r-1} h^j \boldsymbol{d}_j^{(n-s)}\big(\tilde{\boldsymbol{\theta}}^{(n-s)}\big) + O(kh^r).
\end{aligned} \tag{54}$$

By Taylor's theorem,

$$\begin{aligned}
\boldsymbol{d}_j^{(n-s)}\big(\tilde{\boldsymbol{\theta}}^{(n-s)}\big) &= \sum_{i=0}^{r-1-j} \frac{1}{i!} \nabla^i \boldsymbol{d}_j^{(n-s)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)\big(\underbrace{\tilde{\boldsymbol{\theta}}^{(n-s)} - \tilde{\boldsymbol{\theta}}^{(n)}}_{i \text{ times}}\big) \\
&\quad + \frac{1}{(r-1-j)!} \int_0^1 (1-t)^{r-1-j} \\
&\qquad \times \nabla^{r-j} \boldsymbol{d}_j^{(n-s)}\big(\tilde{\boldsymbol{\theta}}^{(n)} + t\big(\tilde{\boldsymbol{\theta}}^{(n-s)} - \tilde{\boldsymbol{\theta}}^{(n)}\big)\big)\big(\underbrace{\tilde{\boldsymbol{\theta}}^{(n-s)} - \tilde{\boldsymbol{\theta}}^{(n)}}_{r-j \text{ times}}\big)\, \mathrm{d}t \\
&\overset{\text{(a)}}{=} \sum_{i=0}^{r-1-j} \frac{1}{i!} \nabla^i \boldsymbol{d}_j^{(n-s)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)\big(\underbrace{\tilde{\boldsymbol{\theta}}^{(n-s)} - \tilde{\boldsymbol{\theta}}^{(n)}}_{i \text{ times}}\big) + O\big(s^{r-j} h^{r-j}\big) \\
&\overset{\text{(b)}}{=} \sum_{i=0}^{r-1-j} \frac{1}{i!} \nabla^i \boldsymbol{d}_j^{(n-s)}(\tilde{\boldsymbol{\theta}}^{(n)})\big(\underbrace{\tilde{\boldsymbol{\theta}}^{(n-s)} - \tilde{\boldsymbol{\theta}}^{(n)}}_{i \text{ times}}\big) + O\big(k^{r-j} h^{r-j}\big),
\end{aligned} \tag{55}$$

where in (a) we used that the derivatives of $\boldsymbol{d}_j^{(s)}(\boldsymbol{\theta})$ are bounded (Lemma B.1) and Lemma B.2; (b) is just because $s \leqslant k$. By the induction assumption,

$$\tilde{\boldsymbol{\theta}}^{(n-s)} = \tilde{\boldsymbol{\theta}}^{(n)} + \sum_{l=1}^{r-1-j} h^l \widetilde{\boldsymbol{d}}_l^{(n,s)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big) + O(k^{r-j} h^{r-j}),$$

which we insert into (55), giving

$$
\begin{aligned}
&\boldsymbol{d}_j^{(n-s)}\big(\tilde{\boldsymbol{\theta}}^{(n-s)}\big) \\
&\quad= \sum_{i=0}^{r-1-j} \frac{1}{i!}\nabla^i \boldsymbol{d}_j^{(n-s)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)\underbrace{\Bigg(\sum_{l=1}^{r-1-j} h^l \widetilde{\boldsymbol{d}}_l^{(n,s)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big) + O(k^{r-j}h^{r-j})\Bigg)}_{i \text{ times}} + O\big(k^{r-j}h^{r-j}\big) \\
&\quad\overset{\text{(a)}}{=} \sum_{i=0}^{r-1-j} \frac{1}{i!}\nabla^i \boldsymbol{d}_j^{(n-s)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)\underbrace{\Bigg(\sum_{l=1}^{r-1-j} h^l \widetilde{\boldsymbol{d}}_l^{(n,s)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)\Bigg)}_{i \text{ times}} + O\big(k^{r-j}h^{r-j}\big) \\
&\quad= \sum_{i=0}^{r-1-j} \frac{h^i}{i!}\nabla^i \boldsymbol{d}_j^{(n-s)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)\underbrace{\Bigg(\sum_{l=0}^{r-2-j} h^l \widetilde{\boldsymbol{d}}_{l+1}^{(n,s)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)\Bigg)}_{i \text{ times}} + O\big(k^{r-j}h^{r-j}\big) \\
&\quad\overset{\text{(b)}}{=} \sum_{i=0}^{r-1-j}\sum_{l=0}^{r-1-j-i} h^{i+l} \\
&\qquad\times \sum_{(i_0,\dots,i_l)\in\mathcal{K}_{i,l}} \frac{1}{i_0!\dots i_l!}\nabla^i \boldsymbol{d}_j^{(n-s)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)\Bigg(\underbrace{\widetilde{\boldsymbol{d}}_1^{(n,s)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)}_{i_0 \text{ times}},\dots,\underbrace{\widetilde{\boldsymbol{d}}_{l+1}^{(n,s)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)}_{i_l \text{ times}}\Bigg) \\
&\qquad+ O\big(k^{r-j}h^{r-j}\big)
\end{aligned}
$$

where in (a) we used that components of $\nabla^i \boldsymbol{d}_j^{(n-s)}(\boldsymbol{\theta})$ are bounded (Lemma B.1) and for $i \geqslant 1$ we have $(kh)^{i(r-j)} = O\big((kh)^{r-j}\big)$ because kh does not exceed T; in (b) we also used that $\widetilde{\boldsymbol{d}}_l^{(n,s)}(\boldsymbol{\theta}) = O(s^l)$ by Lemma B.1. Inserting this into (54) gives

$$
\begin{aligned}
&\tilde{\boldsymbol{\theta}}^{(n-k)} - \tilde{\boldsymbol{\theta}}^{(n)} \\
&\quad= -\sum_{s=1}^{k}\sum_{j=1}^{r-1} h^j \boldsymbol{d}_j^{(n-s)}\big(\tilde{\boldsymbol{\theta}}^{(n-s)}\big) + O(kh^r) \\
&\quad= -\sum_{s=1}^{k}\sum_{j=1}^{r-1} h^j \sum_{i=0}^{r-1-j}\sum_{l=0}^{r-1-j-i} h^{i+l}\times \\
&\qquad\times \sum_{(i_0,\dots,i_l)\in\mathcal{K}_{i,l}} \frac{1}{i_0!\dots i_l!}\nabla^i \boldsymbol{d}_j^{(n-s)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)\Bigg(\underbrace{\widetilde{\boldsymbol{d}}_1^{(n,s)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)}_{i_0 \text{ times}},\dots,\underbrace{\widetilde{\boldsymbol{d}}_{l+1}^{(n,s)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)}_{i_l \text{ times}}\Bigg) \\
&\qquad- \sum_{s=1}^{k}\sum_{j=1}^{r-1} h^j O\big(k^{r-j}h^{r-j}\big) + O(kh^r) \\
&\quad= -\sum_{m=1}^{r-1} h^m \sum_{s=1}^{k}\sum_{\substack{j\geqslant 1,i,l\geqslant 0\\ i+j+l=m}}\sum_{(i_0,\dots,i_l)\in\mathcal{K}_{i,l}} \frac{1}{i_0!\dots i_l!}\nabla^i \boldsymbol{d}_j^{(n-s)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)\Bigg(\underbrace{\widetilde{\boldsymbol{d}}_1^{(n,s)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)}_{i_0 \text{ times}},\dots,\underbrace{\widetilde{\boldsymbol{d}}_{l+1}^{(n,s)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)}_{i_l \text{ times}}\Bigg) \\
&\qquad+ O(k^r h^r) \\
&\quad\overset{\text{(a)}}{=} \sum_{m=1}^{r-1} h^m \widetilde{\boldsymbol{d}}_m^{(n,k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big) + O(k^r h^r),
\end{aligned}
$$

where in (a) we used (15). We have completed the induction step for (20). □

We will now conclude the proof of (18). Inserting (20) into (53) gives

$$
\begin{aligned}
&\nabla L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n-k)}\big) \\
&\quad= \sum_{i=0}^{\mathcal{R}-1} \frac{1}{i!} \nabla^{i+1} L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big) \Bigg(\underbrace{\sum_{m=1}^{\mathcal{R}-1} h^m \widetilde{\boldsymbol{d}}_m^{(n,k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big) + O(k^{\mathcal{R}} h^{\mathcal{R}})}_{i \text{ times}} \Bigg) + O(k^{\mathcal{R}} h^{\mathcal{R}}) \\
&\quad\overset{\text{(a)}}{=} \sum_{i=0}^{\mathcal{R}-1} \frac{1}{i!} \nabla^{i+1} L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big) \Bigg(\underbrace{\sum_{m=1}^{\mathcal{R}-1} h^m \widetilde{\boldsymbol{d}}_m^{(n,k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)}_{i \text{ times}} \Bigg) + O(k^{\mathcal{R}} h^{\mathcal{R}}) \\
&\quad= \sum_{i=0}^{\mathcal{R}-1} \frac{h^i}{i!} \nabla^{i+1} L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big) \Bigg(\underbrace{\sum_{m=0}^{\mathcal{R}-2} h^m \widetilde{\boldsymbol{d}}_{m+1}^{(n,k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)}_{i \text{ times}} \Bigg) + O(k^{\mathcal{R}} h^{\mathcal{R}}) \\
&\quad= \sum_{m=0}^{\mathcal{R}-1} h^m \sum_{\substack{i,l \geqslant 0 \\ i+l=m}} \sum_{(i_0,\ldots,i_l) \in \mathcal{K}_{i,l}} \frac{1}{i_0! \ldots i_l!} \nabla^{i+1} L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big) \Bigg(\underbrace{\widetilde{\boldsymbol{d}}_1^{(n,k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)}_{i_0 \text{ times}}, \ldots, \underbrace{\widetilde{\boldsymbol{d}}_{l+1}^{(n,k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)}_{i_l \text{ times}} \Bigg) \\
&\qquad + O(k^{\mathcal{R}} h^{\mathcal{R}})
\end{aligned}
$$

where in (a) we used that for $i \geqslant 1$ we have $(kh)^{i\mathcal{R}} = O\big((kh)^{\mathcal{R}}\big)$ because kh does not exceed T. Using exponential summation gives

$$
\begin{aligned}
&- \sum_{k=0}^{n} \beta^k \nabla L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n-k)}\big) \\
&\quad= -\nabla L^{(n)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big) \\
&\qquad - \sum_{m=0}^{\mathcal{R}-1} h^m \sum_{k=1}^{n} \beta^k \sum_{\substack{i,l \geqslant 0 \\ i+l=m}} \sum_{(i_0,\ldots,i_l) \in \mathcal{K}_{i,l}} \frac{1}{i_0! \ldots i_l!} \times \\
&\qquad\qquad \times \nabla^{i+1} L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big) \Bigg(\underbrace{\widetilde{\boldsymbol{d}}_1^{(n,k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)}_{i_0 \text{ times}}, \ldots, \underbrace{\widetilde{\boldsymbol{d}}_{l+1}^{(n,k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)}_{i_l \text{ times}} \Bigg) + O(h^{\mathcal{R}}) \\
&\quad= - \sum_{k=0}^{n} \beta^k \nabla L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big) \\
&\qquad - \sum_{m=1}^{\mathcal{R}-1} h^m \sum_{k=1}^{n} \beta^k \sum_{\substack{i,l \geqslant 0 \\ i+l=m}} \sum_{(i_0,\ldots,i_l) \in \mathcal{K}_{i,l}} \frac{1}{i_0! \ldots i_l!} \times \\
&\qquad\qquad \times \nabla^{i+1} L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big) \Bigg(\underbrace{\widetilde{\boldsymbol{d}}_1^{(n,k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)}_{i_0 \text{ times}}, \ldots, \underbrace{\widetilde{\boldsymbol{d}}_{l+1}^{(n,k)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big)}_{i_l \text{ times}} \Bigg) + O(h^{\mathcal{R}}) \\
&\quad= \sum_{m=0}^{\mathcal{R}-1} h^m \boldsymbol{d}_{m+1}^{(n)}\big(\tilde{\boldsymbol{\theta}}^{(n)}\big) + O(h^{\mathcal{R}}),
\end{aligned}
$$

as desired. Since (18) was sufficient for (16), we have proven (16) as well.

Lemma B.3. *Equation* (17) *follows from* (16).

Proof. The argument is standard. Define the error at the nth step

$$
\boldsymbol{e}^{(n)} := \boldsymbol{\theta}^{(n)} - \tilde{\boldsymbol{\theta}}^{(n)}
$$

and local error

$$\boldsymbol{\delta}^{(n)} := \tilde{\boldsymbol{\theta}}^{(n+1)} - \tilde{\boldsymbol{\theta}}^{(n)} + h\sum_{k=0}^{n}\beta^k \nabla L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n-k)}\big),$$

where $n \in [0 : \lfloor T/h \rfloor]$. By definition and Taylor's theorem, we have

$$\begin{aligned}\boldsymbol{e}^{(n+1)} &= \boldsymbol{e}^{(n)} - h\sum_{k=0}^{n}\beta^k\Big(\nabla L^{(n-k)}\big(\boldsymbol{\theta}^{(n-k)}\big) - \nabla L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n-k)}\big)\Big) - \boldsymbol{\delta}^{(n)} \\ &= \boldsymbol{e}^{(n)} - h\sum_{k=0}^{n}\beta^k \int_0^1 \nabla^2 L^{(n-k)}\big(\tilde{\boldsymbol{\theta}}^{(n-k)} + t\big(\boldsymbol{\theta}^{(n-k)} - \tilde{\boldsymbol{\theta}}^{(n-k)}\big)\big)\big(\boldsymbol{\theta}^{(n-k)} - \tilde{\boldsymbol{\theta}}^{(n-k)}\big)\,\mathrm{d}t - \boldsymbol{\delta}^{(n)}.\end{aligned}$$

Since the derivatives of the loss are bounded, this implies

$$\|\boldsymbol{e}^{(n+1)}\| \leqslant \|\boldsymbol{e}^{(n)}\| + hC_4\sum_{k=0}^{n}\beta^k\|\boldsymbol{e}^{(n-k)}\| + \|\boldsymbol{\delta}^{(n)}\|$$

with some constant C_4. Denoting $s^{(n)} := \max_{0\leqslant k\leqslant n}\|\boldsymbol{e}^{(k)}\|$, we have

$$s^{(n+1)} \leqslant s^{(n)} + hC_5 s^{(n)} + \|\boldsymbol{\delta}^{(n)}\| = (1+hC_5)s^{(n)} + \|\boldsymbol{\delta}^{(n)}\| \overset{\text{(a)}}{\leqslant} (1+hC_5)s^{(n)} + C_1 h^{\mathcal{R}+1},$$

with some constant C_5, where (a) is by (16). Applying this inequality iteratively, we obtain

$$\begin{aligned}s^{(n)} &\leqslant (1+hC_5)^n s^{(0)} + \frac{(1+hC_5)^n - 1}{C_5}C_1 h^{\mathcal{R}} \\ &= \frac{(1+hC_5)^n - 1}{C_5}C_1 h^{\mathcal{R}} \overset{\text{(a)}}{\leqslant} \frac{e^{C_5 nh} - 1}{C_5}C_1 h^{\mathcal{R}} \leqslant C_6 e^{C_5 nh} h^{\mathcal{R}} \leqslant C_6 e^{C_5 T} h^{\mathcal{R}},\end{aligned}$$

where in (a) we used the inequality $1 + x \leqslant e^x$ for any $x \geqslant 0$, C_6 is some constant. It is left to apply it with $n = \lfloor T/h \rfloor$ and put $C_2 = C_6 e^{C_5 T}$. □

We have proven both (16) and (17), concluding the proof of Theorem 3.1.

C Proof of Theorem 4.1

Proof of Lemma 4.2. We prove this by induction in m.

Induction base For the tree τ with vertices $\{1, 2\}$ where 1 is the root, we have

$$\begin{aligned}\boldsymbol{E}^{(n-l\to n)}_{\tau,\{2\},a} + \boldsymbol{E}^{(n-l\to n)}_{\tau,\varnothing,a} &= \sum_{b=0}^{n-l-a}\beta^b\nabla^2 L^{(n-l-a-b)}\sum_{l'=1}^{l}\boldsymbol{E}^{(n)}_{\bullet,l'} + \sum_{b=0}^{n-l-a}\beta^b\nabla^2 L^{(n-l-a-b)}\sum_{l'=1}^{a+b}\boldsymbol{E}^{(n-l)}_{\bullet,l'} \\ &= \sum_{b=0}^{n-l-a}\beta^b\nabla^2 L^{(n-l-a-b)}\sum_{l'=1}^{l+a+b}\boldsymbol{E}^{(n)}_{\bullet,l'} = \boldsymbol{E}^{(n)}_{\bullet,l+a},\end{aligned}$$

where we used

$$\begin{aligned}\sum_{l'=1}^{l}\boldsymbol{E}^{(n)}_{\bullet,l'} + \sum_{l'=1}^{a+b}\boldsymbol{E}^{(n-l)}_{\bullet,l'} &= \sum_{l'=1}^{l}\sum_{b_1=0}^{n-l'}\beta^{b_1}\nabla L^{(n-l'-b_1)} + \sum_{l'=1}^{a+b}\sum_{b_1=0}^{n-l-l'}\beta^{b_1}\nabla L^{(n-l-l'-b_1)} \\ &= \sum_{l'=1}^{l}\sum_{b_1=0}^{n-l'}\beta^{b}\nabla L^{(n-l'-b_1)} + \sum_{l'=l+1}^{l+a+b}\sum_{b_1=0}^{n-l'}\beta^{b}\nabla L^{(n-l'-b_1)} \\ &= \sum_{l'=1}^{l+a+b}\sum_{b_1=0}^{n-l'}\beta^{b_1}\nabla L^{(n-l'-b_1)} = \sum_{l'=1}^{l+a+b}\boldsymbol{E}^{(n)}_{\bullet,l'}.\end{aligned}$$

Induction step Let r be the root of τ, $v_1, \dots, v_\ell$ its children, $\tau_1, \dots, \tau_\ell$ the corresponding subtrees rooted at the children, $V_1, \dots, V_\ell$ their vertex sets (partitioning $[1\,{:}\,m] \setminus \{r\}$). From the definition,

$$\boldsymbol{E}^{(n-l\to n)}_{\tau,m,a} = \sum_{b=0}^{n-l-a} \beta^b \nabla^{\ell+1} L^{(n-l-a-b)} \left[\sum_{l'=1}^{l} \boldsymbol{E}^{(n)}_{\tau_s,l'} \mathbf{1}_{v_s \in m} + \sum_{l'=1}^{a+b} \boldsymbol{E}^{(n-l\to n)}_{\tau_s,m\cap V_s,l'} \mathbf{1}_{v_s \notin m} \right]_{s=1}^{\ell},$$

which means

$$\begin{aligned}
\sum_{\substack{m\in\mathcal{M}_\tau \\ v_1\notin m}} \boldsymbol{E}^{(n-l\to n)}_{\tau,m,a} &= \sum_{\substack{m\in\mathcal{M}_\tau \\ v_1\notin m}} \sum_{b=0}^{n-l-a} \beta^b \times \\
&\quad \times \nabla^{\ell+1} L^{(n-l-a-b)} \left[\sum_{l'=1}^{a+b} \boldsymbol{E}^{(n-l\to n)}_{\tau_1,m\cap V_1,l'}, \dots, \sum_{l'=1}^{l} \boldsymbol{E}^{(n)}_{\tau_\ell,l'} \mathbf{1}_{v_\ell \in m} + \sum_{l'=1}^{a+b} \boldsymbol{E}^{(n-l\to n)}_{\tau_\ell,m\cap V_\ell,l'} \mathbf{1}_{v_\ell \notin m} \right] \\
&= \sum_{m_1\in\mathcal{M}_{\tau_1}} \sum_{m_2\in\mathcal{M}_{\tau\setminus\tau_1}} \sum_{b=0}^{n-l-a} \beta^b \times \\
&\quad \times \nabla^{\ell+1} L^{(n-l-a-b)} \left[\sum_{l'=1}^{a+b} \boldsymbol{E}^{(n-l\to n)}_{\tau_1,m_1,l'}, \dots, \sum_{l'=1}^{l} \boldsymbol{E}^{(n)}_{\tau_\ell,l'} \mathbf{1}_{v_\ell \in m_2} + \sum_{l'=1}^{a+b} \boldsymbol{E}^{(n-l\to n)}_{\tau_\ell,m_2\cap V_\ell,l'} \mathbf{1}_{v_\ell \notin m_2} \right] \\
&= \sum_{m_2\in\mathcal{M}_{\tau\setminus\tau_1}} \sum_{b=0}^{n-l-a} \beta^b \times \\
&\quad \times \nabla^{\ell+1} L^{(n-l-a-b)} \left[\sum_{l'=1}^{a+b} \sum_{m_1\in\mathcal{M}_{\tau_1}} \boldsymbol{E}^{(n-l\to n)}_{\tau_1,m_1,l'}, \dots, \sum_{l'=1}^{l} \boldsymbol{E}^{(n)}_{\tau_\ell,l'} \mathbf{1}_{v_\ell \in m_2} + \sum_{l'=1}^{a+b} \boldsymbol{E}^{(n-l\to n)}_{\tau_\ell,m_2\cap V_\ell,l'} \mathbf{1}_{v_\ell \notin m_2} \right].
\end{aligned}$$

Here, we used the one-to-one correspondence between markings m not containing v_1 and pairs (m_1, m_2), where m_1 is a marking of τ_1 and m_2 is a marking of the tree $\tau \setminus \tau_1$. Now we apply the induction hypothesis (since τ_1 is a smaller tree) and replace $\sum_{m_1\in\mathcal{M}_{\tau_1}} \boldsymbol{E}^{(n-l\to n)}_{\tau_1,m_1,l'}$ with $\boldsymbol{E}^{(n)}_{\tau_1,l+l'}$:

$$\begin{aligned}
\sum_{\substack{m\in\mathcal{M}_\tau \\ v_1\notin m}} \boldsymbol{E}^{(n-l\to n)}_{\tau,m,a} &= \sum_{m_2\in\mathcal{M}_{\tau\setminus\tau_1}} \sum_{b=0}^{n-l-a} \beta^b \\
&\quad \times \nabla^{\ell+1} L^{(n-l-a-b)} \left[\sum_{l'=1}^{a+b} \boldsymbol{E}^{(n)}_{\tau_1,l+l'}, \dots, \sum_{l'=1}^{l} \boldsymbol{E}^{(n)}_{\tau_\ell,l'} \mathbf{1}_{v_\ell \in m_2} + \sum_{l'=1}^{a+b} \boldsymbol{E}^{(n-l\to n)}_{\tau_\ell,m_2\cap V_\ell,l'} \mathbf{1}_{v_\ell \notin m_2} \right].
\end{aligned}$$

We deal with the case $v_1 \in m$ similarly, obtaining

$$\begin{aligned}
\sum_{\substack{m\in\mathcal{M}_\tau \\ v_1\in m}} \boldsymbol{E}^{(n-l\to n)}_{\tau,m,a} &= \sum_{m_2\in\mathcal{M}_{\tau\setminus\tau_1}} \sum_{b=0}^{n-l-a} \beta^b \\
&\quad \times \nabla^{\ell+1} L^{(n-l-a-b)} \left[\sum_{l'=1}^{l} \boldsymbol{E}^{(n)}_{\tau_1,l'}, \dots, \sum_{l'=1}^{l} \boldsymbol{E}^{(n)}_{\tau_\ell,l'} \mathbf{1}_{v_\ell \in m_2} + \sum_{l'=1}^{a+b} \boldsymbol{E}^{(n-l\to n)}_{\tau_\ell,m_2\cap V_\ell,l'} \mathbf{1}_{v_\ell \notin m_2} \right].
\end{aligned}$$

Adding the latter two equations and using $\sum_{l'=1}^{a+b} \boldsymbol{E}^{(n)}_{\tau_1,l+l'} + \sum_{l'=1}^{l} \boldsymbol{E}^{(n)}_{\tau_1,l'} = \sum_{l'=1}^{l+a+b} \boldsymbol{E}^{(n)}_{\tau_1,l'}$ gives

$$\sum_{m\in\mathcal{M}_\tau} \boldsymbol{E}^{(n-l\to n)}_{\tau,m,a} = \sum_{m_2\in\mathcal{M}_{\tau\setminus\tau_1}} \sum_{b=0}^{n-l-a} \beta^b$$

$$\times \nabla^{\ell+1} L^{(n-l-a-b)} \Bigg[\sum_{l'=1}^{l+a+b} \boldsymbol{E}^{(n)}_{\tau_1,l'}, \dots, \sum_{l'=1}^{l} \boldsymbol{E}^{(n)}_{\tau_\ell,l'} \mathbf{1}_{v_\ell \in m_2} + \sum_{l'=1}^{a+b} \boldsymbol{E}^{(n-l\to n)}_{\tau_\ell,m_2\cap V_\ell,l'} \mathbf{1}_{v_\ell \notin m_2} \Bigg].$$

Continuing the same argument with $v_2, \dots, v_\ell$, we will arrive at

$$\sum_{m\in\mathcal{M}_\tau} \boldsymbol{E}^{(n-l\to n)}_{\tau,m,a} = \sum_{b=0}^{n-l-a} \beta^b \nabla^{\ell+1} L^{(n-l-a-b)} \Bigg[\sum_{l'=1}^{l+a+b} \boldsymbol{E}^{(n)}_{\tau_1,l'}, \dots, \sum_{l'=1}^{l+a+b} \boldsymbol{E}^{(n)}_{\tau_\ell,l'} \Bigg].$$

By definition, the right-hand side is equal to $\boldsymbol{E}^{(n)}_{\tau,l+a}$, completing the induction step. □

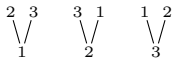
Proof of Lemma 4.3. Consider the following operation. Choose a partition of $[1:m]$ into $i+1$ disjoint non-empty sets $(V_0, \{V_1, \dots, V_i\})$, where $0 \leqslant i \leqslant m-1$ (one of the sets is privileged, but the order in the other ones does not matter); choose a labeled rooted tree $\tau_0 \in \mathcal{A}[V_0]$ with root $v_0 \in V_0$, another labeled rooted tree $\tau_1 \in \mathcal{A}[V_1]$ with root $v_1 \in V_1$, and so on, up $\tau_i \in \mathcal{A}[V_i]$ with root $v_i \in V_i$.

Now, assign to each of $v_1, \dots, v_i$ a parent among vertices of τ_0 in all possible ways, by choosing all mappings from $\{v_1, \dots, v_i\}$ to V_0. Write the following expression corresponding to the family of all such assignments:

$$\nabla^i \boldsymbol{E}^{(n-l)}_{\tau_0,a} \Bigg[\sum_{l'=1}^{l} \boldsymbol{E}^{(n)}_{\tau_1,l'}, \dots, \sum_{l'=1}^{l} \boldsymbol{E}^{(n)}_{\tau_i,l'} \Bigg]. \tag{56}$$

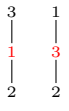
This way, we have constructed a family of labeled rooted trees $\tau \in \mathcal{A}[1:m]$ with roots of $\tau_1, \dots, \tau_i$ marked.

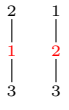
For example, consider $m = 3$. When $i = 0$, there is only one partition and three corresponding marked trees (with no vertices marked):

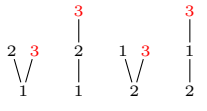
$V_0 = \{1,2,3\}$ 

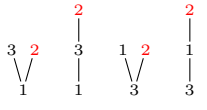
When $i = 1$, there are six partitions listed below with corresponding marked trees

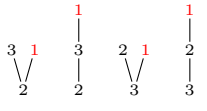
$V_0 = \{1\}, V_1 = \{2,3\}$ 

$V_0 = \{2\}, V_1 = \{1,3\}$ 

$V_0 = \{3\}, V_1 = \{1,2\}$ 

$V_0 = \{1,2\}, V_1 = \{3\}$ 

$V_0 = \{1,3\}, V_1 = \{2\}$ 

$V_0 = \{2,3\}, V_1 = \{1\}$ 

When $i = 2$, there are three partitions with corresponding marked trees

$$V_0 = \{1\}, V_1 = \{2\}, V_2 = \{3\} \quad \overset{2\ 3}{\backslash /}\ {}_1$$

$$V_0 = \{2\}, V_1 = \{1\}, V_2 = \{3\} \quad \overset{1\ 3}{\backslash /}\ {}_2$$

$$V_0 = \{3\}, V_1 = \{1\}, V_2 = \{2\} \quad \overset{1\ 2}{\backslash /}\ {}_3$$

Sum the expression (56) over the choices of a partition and the trees $\tau_0, \dots, \tau_i$:

$$\sum_{i=0}^{m-1} \sum_{\substack{(V_0,\{V_1\dots,V_i\}) \\ \text{partition of } [1:m]}} \sum_{\tau_0 \in \mathcal{A}[V_0],\dots,\tau_i \in \mathcal{A}[V_i]} \nabla^i \boldsymbol{E}^{(n-l)}_{\tau_0,a} \left[\sum_{l'=1}^{l} \boldsymbol{E}^{(n)}_{\tau_1,l'}, \dots, \sum_{l'=1}^{l} \boldsymbol{E}^{(n)}_{\tau_i,l'} \right].$$

Since all marked trees in $\mathcal{A}[1:m]$ can be constructed this way and are counted exactly once in this sum, this equals

$$\sum_{\tau \in \mathcal{A}[1:m]} \sum_{m \in \mathcal{M}_\tau} \boldsymbol{E}^{(n-l\to n)}_{\tau,m,a}.$$

But $\sum_{m\in\mathcal{M}_\tau} \boldsymbol{E}^{(n-l\to n)}_{\tau,m,a}$ is $\boldsymbol{E}^{(n)}_{\tau,l+a}$ by (25). Hence, we have obtained

$$\sum_{i=0}^{m-1} \sum_{\substack{(V_0,\{V_1\dots,V_i\}) \\ \text{partition of } [1:m]}} \sum_{\tau_0 \in \mathcal{A}[V_0],\dots,\tau_i \in \mathcal{A}[V_i]} \nabla^i \boldsymbol{E}^{(n-l)}_{\tau_0,a} \left[\sum_{l'=1}^{l} \boldsymbol{E}^{(n)}_{\tau_1,l'}, \dots, \sum_{l'=1}^{l} \boldsymbol{E}^{(n)}_{\tau_i,l'} \right] = \sum_{\tau \in \mathcal{A}[1:m]} \boldsymbol{E}^{(n)}_{\tau,l+a}.$$

Recall that $m!/\sigma(\tau)$ labeled rooted trees correspond to the same unlabeled rooted tree. Hence, this can be rewritten as

$$\begin{aligned} &\sum_{i=0}^{m-1} \sum_{\substack{(V_0,\{V_1\dots,V_i\}) \\ \text{partition of } [1:m]}} \sum_{\tau_0 \in \tilde{\mathcal{A}}[|V_0|],\dots,\tau_i \in \tilde{\mathcal{A}}[|V_i|]} \frac{|V_0|!}{\sigma(\tau_0)} \\ &\quad \times \nabla^i \boldsymbol{E}^{(n-l)}_{\tau_0,a} \left[\frac{|V_1|!}{\sigma(\tau_1)} \sum_{l'=1}^{l} \boldsymbol{E}^{(n)}_{\tau_1,l'}, \dots, \frac{|V_i|!}{\sigma(\tau_i)} \sum_{l'=1}^{l} \boldsymbol{E}^{(n)}_{\tau_i,l'} \right] = \sum_{\tau \in \tilde{\mathcal{A}}[m]} \frac{m!}{\sigma(\tau)} \boldsymbol{E}^{(n)}_{\tau,l+a}. \end{aligned} \tag{57}$$

Let us fix j, $(k_0, \dots, k_{m-j-i}) \in \mathcal{K}_{i,m-j-i}$ and count the number of partitions with $|V_0| = j$, where among $\{V_1, \dots, V_i\}$ there are k_0 sets of size 1, …, k_{m-j-i} sets of size $m-j-i+1$. First, we choose the elements of V_0 in $\binom{m}{j}$ ways, then we order the remaining elements in $(m-j)!$ ways and assign the first k_0 elements in this ordering as singletons, the next k_1 pairs as two-sets, and so on. Notice that each partition of $V \setminus V_0$ will be counted $k_0! \dots k_{m-j-i}! 1!^{k_0} \dots (m-j-i+1)!^{k_{m-j-i}}$ times (because the order of sets with the same length does not matter, and the order within each set does not matter). So, the required number of partitions is

$$\begin{aligned} &\binom{m}{j} \frac{(m-j)!}{k_0! \dots k_{m-j-i}! 1!^{k_0} \dots (m-j-i+1)!^{k_{m-j-i}}} \\ &\quad = \frac{m!}{j! k_0! \dots k_{m-j-i}! 1!^{k_0} \dots (m-j-i+1)!^{k_{m-j-i}}}. \end{aligned}$$

Thus, (57) can be rewritten as

$$\sum_{i=0}^{m-1} \sum_{j=1}^{m-i} \sum_{(k_0,\dots,k_{m-j-i}) \in \mathcal{K}_{i,m-j-i}} \frac{m!}{k_0! \dots k_{m-j-i}!} \sum_{\tau_0 \in \tilde{\mathcal{A}}[j]} \frac{1}{\sigma(\tau_0)}$$

$$\times \nabla^i \boldsymbol{E}^{(n-l)}_{\tau_0,a} \Bigg[\underbrace{\sum_{\tau\in\tilde{\mathcal{A}}[1]} \frac{1}{\sigma(\tau)} \sum_{l'=1}^{l} \boldsymbol{E}^{(n)}_{\tau,l'}}_{k_0 \text{ times}}, \dots, \underbrace{\sum_{\tau\in\tilde{\mathcal{A}}[m-j-i+1]} \frac{1}{\sigma(\tau)} \sum_{l'=1}^{l} \boldsymbol{E}^{(n)}_{\tau,l'}}_{k_{m-j-i} \text{ times}} \Bigg] = \sum_{\tau\in\tilde{\mathcal{A}}[m]} \frac{m!}{\sigma(\tau)} \boldsymbol{E}^{(n)}_{\tau,l+a},$$

where $j!1!^{k_0}\dots(m-j-i+1)!^{k_{m-j-i}}$ canceled out with $|V_0|!|V_1|!\dots|V_i|!$. Dividing both sides by $m!$ completes the proof. □

Proof of Theorem 4.1. We prove (27) and (28) by induction over $m \geqslant 2$.

For $m = 2$, they are already verified above. Note also that the second statement holds for $m = 1$ as well:

$$\widetilde{\boldsymbol{d}}^{(n,k)}_1 = \sum_{l=1}^{k} \boldsymbol{E}^{(n)}_{\bullet,l}.$$

By definition,

$$\boldsymbol{d}^{(n)}_m = -\beta \sum_{b=0}^{n-1} \beta^b \sum_{\ell=0}^{m-1} \sum_{(i_0,\dots,i_{m-1-\ell})\in\mathcal{K}_{\ell,m-1-\ell}} \frac{1}{i_0!\dots i_{m-1-\ell}!} \times$$
$$\times \nabla^{\ell+1} L^{(n-1-b)} \Bigg(\underbrace{\widetilde{\boldsymbol{d}}^{(n,b+1)}_1}_{i_0 \text{ times}}, \dots, \underbrace{\widetilde{\boldsymbol{d}}^{(n,b+1)}_{m-\ell}}_{i_{m-1-\ell} \text{ times}} \Bigg)$$

Insert the induction hypothesis (recall that (28) holds for $m = 1$ too):

$$\boldsymbol{d}^{(n)}_m = -\beta \sum_{b=0}^{n-1} \beta^b \sum_{\ell=0}^{m-1} \sum_{(i_0,\dots,i_{m-1-\ell})\in\mathcal{K}_{\ell,m-1-\ell}} \frac{1}{i_0!\dots i_{m-1-\ell}!}$$
$$\times \nabla^{\ell+1} L^{(n-1-b)} \Bigg(\underbrace{\sum_{l=1}^{b+1} \sum_{\tau\in\tilde{\mathcal{A}}[1]} \frac{1}{\sigma(\tau)} \boldsymbol{E}^{(n)}_{\tau,l}}_{i_0 \text{ times}}, \dots, \underbrace{\sum_{l=1}^{b+1} \sum_{\tau\in\tilde{\mathcal{A}}[m-\ell]} \frac{1}{\sigma(\tau)} \boldsymbol{E}^{(n)}_{\tau,l}}_{i_{m-1-\ell} \text{ times}} \Bigg)$$

The sum over $\tilde{\mathcal{A}}[s]$, repeated i_{s-1} times, generates a list of i_{s-1}-tuples of trees with s vertices. On this list, each multiset of i_{s-1} trees with multiplicities $\mu^s_1, \dots, \mu^s_{|\tilde{\mathcal{A}}[s]|}$ appears $\binom{i_{s-1}}{\mu^s_1 \dots \mu^s_{|\tilde{\mathcal{A}}[s]|}}$ times. Therefore, in the large sum above each multiset of ℓ trees with the total number of vertices $m-1$ and matching the vertex-count multiplicities $i_0, \dots, i_{m-1-\ell}$ (equivalently, each tree τ with m vertices whose root has ℓ children matching these vertex-count multiplicities) appears

$$\prod_{s=1}^{m-\ell} \frac{i_{s-1}!}{\mu^s_1(\tau)! \dots \mu^s_{|\tilde{\mathcal{A}}[s]|}(\tau)!} = \frac{i_0! \dots i_{m-1-\ell}!}{\prod_{s=1}^{m-\ell} \big(\mu^s_1(\tau)! \dots \mu^s_{|\tilde{\mathcal{A}}[s]|}(\tau)!\big)}$$

times, and we can rewrite

$$\boldsymbol{d}^{(n)}_m = -\beta \sum_{b=0}^{n-1} \beta^b \sum_{\ell=0}^{m-1} \sum_{(i_0,\dots,i_{m-1-\ell})\in\mathcal{K}_{\ell,m-1-\ell}} \sum_{\substack{\tau=[\tau_1,\dots,\tau_\ell]\in\tilde{\mathcal{A}}[m]: \\ \tau \text{ matches } i_0,\dots,i_{m-1-\ell}}} \frac{1}{\prod_{s=1}^{m-\ell} \big(\mu^s_1(\tau)! \dots \mu^s_{|\tilde{\mathcal{A}}[s]|}(\tau)!\big)}$$
$$\times \nabla^{\ell+1} L^{(n-1-b)} \Bigg(\frac{1}{\sigma(\tau_1)} \sum_{l=1}^{b+1} \boldsymbol{E}^{(n)}_{\tau_1,l}, \dots, \frac{1}{\sigma(\tau_\ell)} \sum_{l=1}^{b+1} \boldsymbol{E}^{(n)}_{\tau_\ell,l} \Bigg).$$

Using (10) simplifies this further:

$$\begin{aligned}
\boldsymbol{d}_m^{(n)} &= -\beta \sum_{b=0}^{n-1} \beta^b \sum_{\ell=0}^{m-1} \sum_{(i_0,\dots,i_{m-1-\ell})\in\mathcal{K}_{\ell,m-1-\ell}} \sum_{\substack{\tau=[\tau_1,\dots,\tau_\ell]\in\tilde{\mathcal{A}}[m]:\\ \tau \text{ matches } i_0,\dots,i_{m-1-\ell}}} \frac{1}{\sigma(\tau)} \\
&\quad \times \nabla^{\ell+1} L^{(n-1-b)}\left(\sum_{l=1}^{b+1} \boldsymbol{E}_{\tau_1,l}^{(n)}, \dots, \sum_{l=1}^{b+1} \boldsymbol{E}_{\tau_\ell,l}^{(n)}\right) \\
&= -\beta \sum_{b=0}^{n-1} \beta^b \sum_{\ell=0}^{m-1} \sum_{\tau=[\tau_1,\dots,\tau_\ell]\in\tilde{\mathcal{A}}[m]} \frac{1}{\sigma(\tau)} \nabla^{\ell+1} L^{(n-1-b)}\left(\sum_{l=1}^{b+1} \boldsymbol{E}_{\tau_1,l}^{(n)}, \dots, \sum_{l=1}^{b+1} \boldsymbol{E}_{\tau_\ell,l}^{(n)}\right) \\
&= -\beta \sum_{\tau\in\tilde{\mathcal{A}}[m]} \frac{1}{\sigma(\tau)} \boldsymbol{E}_{\tau,1}^{(n)}.
\end{aligned}$$

We have proven that under the induction hypothesis for smaller m, (27) holds.

By definition of the history terms in (15) and the induction hypothesis, to prove (28) it is enough to show (29).

By the induction hypothesis and (27) already proven, the left-hand side of (29) is

$$\sum_{i=0}^{m-1} \sum_{(k_0,\dots,k_{m-i-1})\in\mathcal{K}_{i,m-i-1}} \frac{1}{k_0!\dots k_{m-i-1}!} \tag{58}$$

$$\times \sum_{b=0}^{n-l} \beta^b \nabla^{i+1} L^{(n-l-b)}\Bigg[\underbrace{\sum_{l'=1}^{l} \sum_{\tau\in\tilde{\mathcal{A}}[1]} \frac{1}{\sigma(\tau)} \boldsymbol{E}_{\tau,l'}^{(n)}}_{k_0 \text{ times}}, \dots, \underbrace{\sum_{l'=1}^{l} \sum_{\tau\in\tilde{\mathcal{A}}[m-i]} \frac{1}{\sigma(\tau)} \boldsymbol{E}_{\tau,l'}^{(n)}}_{k_{m-i-1} \text{ times}}\Bigg] \tag{59}$$

$$+ \beta \sum_{j=2}^{m} \sum_{i=0}^{m-j} \sum_{(k_0,\dots,k_{m-i-j})\in\mathcal{K}_{i,m-i-j}} \frac{1}{k_0!\dots k_{m-i-j}!} \tag{60}$$

$$\times \nabla^i \sum_{\tau_0\in\tilde{\mathcal{A}}[j]} \frac{1}{\sigma(\tau_0)} \boldsymbol{E}_{\tau_0,1}^{(n-l)}\Bigg[\underbrace{\sum_{l'=1}^{l} \sum_{\tau\in\tilde{\mathcal{A}}[1]} \frac{1}{\sigma(\tau)} \boldsymbol{E}_{\tau,l'}^{(n)}}_{k_0 \text{ times}}, \dots, \underbrace{\sum_{l'=1}^{l} \sum_{\tau\in\tilde{\mathcal{A}}[m-i-j+1]} \frac{1}{\sigma(\tau)} \boldsymbol{E}_{\tau,l'}^{(n)}}_{k_{m-i-j} \text{ times}}\Bigg] \tag{61}$$

$$= \sum_{i=0}^{m-1} \sum_{(k_0,\dots,k_{m-i-1})\in\mathcal{K}_{i,m-i-1}} \frac{1}{k_0!\dots k_{m-i-1}!} \tag{62}$$

$$\times \sum_{b=0}^{n-l} \beta^b \nabla^{i+1} L^{(n-l-b)}\Bigg[\underbrace{\sum_{l'=1}^{l} \sum_{\tau\in\tilde{\mathcal{A}}[1]} \frac{1}{\sigma(\tau)} \boldsymbol{E}_{\tau,l'}^{(n)}}_{k_0 \text{ times}}, \dots, \underbrace{\sum_{l'=1}^{l} \sum_{\tau\in\tilde{\mathcal{A}}[m-i]} \frac{1}{\sigma(\tau)} \boldsymbol{E}_{\tau,l'}^{(n)}}_{k_{m-i-1} \text{ times}}\Bigg] \tag{63}$$

$$- \beta \sum_{i=0}^{m-1} \sum_{(k_0,\dots,k_{m-i-1})\in\mathcal{K}_{i,m-i-1}} \frac{1}{k_0!\dots k_{m-i-1}!} \tag{64}$$

$$\times \nabla^i \boldsymbol{E}_{\bullet,1}^{(n-l)}\Bigg[\underbrace{\sum_{l'=1}^{l} \sum_{\tau\in\tilde{\mathcal{A}}[1]} \frac{1}{\sigma(\tau)} \boldsymbol{E}_{\tau,l'}^{(n)}}_{k_0 \text{ times}}, \dots, \underbrace{\sum_{l'=1}^{l} \sum_{\tau\in\tilde{\mathcal{A}}[m-i]} \frac{1}{\sigma(\tau)} \boldsymbol{E}_{\tau,l'}^{(n)}}_{k_{m-i-1} \text{ times}}\Bigg] \tag{65}$$

$$+\beta\sum_{j=1}^{m}\sum_{i=0}^{m-j}\sum_{(k_0,\dots,k_{m-i-j})\in\mathcal{K}_{i,m-i-j}}\frac{1}{k_0!\dots k_{m-i-j}!} \tag{66}$$

$$\times\sum_{\tau_0\in\tilde{\mathcal{A}}[j]}\frac{1}{\sigma(\tau_0)}\nabla^i\boldsymbol{E}_{\tau_0,1}^{(n-l)}\Bigg[\underbrace{\sum_{l'=1}^{l}\sum_{\tau\in\tilde{\mathcal{A}}[1]}\frac{1}{\sigma(\tau)}\boldsymbol{E}_{\tau,l'}^{(n)},\dots,\sum_{l'=1}^{l}\sum_{\tau\in\tilde{\mathcal{A}}[m-i-j+1]}\frac{1}{\sigma(\tau)}\boldsymbol{E}_{\tau,l'}^{(n)}}_{k_{m-i-j}\text{ times}}\Bigg]. \tag{67}$$

Inserting $\boldsymbol{E}_{\bullet,1}^{(n-l)}=\sum_{b=0}^{n-l-1}\beta^b\nabla L^{(n-l-1-b)}$, we see that the sum in Eqs. (62) to (65) evaluates to

$$\begin{aligned}
&\sum_{i=0}^{m-1}\sum_{(k_0,\dots,k_{m-i-1})\in\mathcal{K}_{i,m-i-1}}\frac{1}{k_0!\dots k_{m-i-1}!}\\
&\quad\times\nabla^{i+1}L^{(n-l)}\Bigg[\underbrace{\sum_{l'=1}^{l}\sum_{\tau\in\tilde{\mathcal{A}}[1]}\frac{1}{\sigma(\tau)}\boldsymbol{E}_{\tau,l'}^{(n)}}_{k_0\text{ times}},\dots,\underbrace{\sum_{l'=1}^{l}\sum_{\tau\in\tilde{\mathcal{A}}[m-i]}\frac{1}{\sigma(\tau)}\boldsymbol{E}_{\tau,l'}^{(n)}}_{k_{m-i-1}\text{ times}}\Bigg]\\
&=\sum_{i=0}^{m-1}\sum_{\tau=[\tau_1,\dots,\tau_i]\in\tilde{\mathcal{A}}[m]}\frac{1}{\sigma(\tau)}\nabla^{i+1}L^{(n-l)}\Bigg[\sum_{l'=1}^{l}\boldsymbol{E}_{\tau_1,l'}^{(n)},\dots,\sum_{l'=1}^{l}\boldsymbol{E}_{\tau_i,l'}^{(n)}\Bigg],
\end{aligned}$$

By Lemma 4.3, the sum in Eqs. (66) and (67) is

$$\beta\sum_{\tau\in\tilde{\mathcal{A}}[m]}\frac{1}{\sigma(\tau)}\boldsymbol{E}_{\tau,l+1}^{(n)}.$$

We have obtained that the left-hand side of (29) is equal to

$$\sum_{i=0}^{m-1}\sum_{\substack{\tau\in\tilde{\mathcal{A}}[m]\\ \tau=[\tau_1,\dots,\tau_i]}}\frac{1}{\sigma(\tau)}\nabla^{i+1}L^{(n-l)}\Bigg[\sum_{l'=1}^{l}\boldsymbol{E}_{\tau_1,l'}^{(n)},\dots,\sum_{l'=1}^{l}\boldsymbol{E}_{\tau_i,l'}^{(n)}\Bigg]+\beta\sum_{\tau\in\tilde{\mathcal{A}}[m]}\frac{1}{\sigma(\tau)}\boldsymbol{E}_{\tau,l+1}^{(n)}.$$

Combining this with (26), we see that the left-hand side of (29) is equal to the right-hand side of (29). This completes the induction step and the whole proof of Theorem 4.1. □

D Proof of Corollaries

D.1 Proof of Corollary 5.1

Lemma D.1. *For all $\boldsymbol{\theta}\in\mathcal{D}$ and all $m\in[1:\mathcal{R}]$, we have*

$$\nabla^r\boldsymbol{f}_m^{(n)}(\boldsymbol{\theta})=O(1),\quad r\in[0:2\mathcal{R}-m].$$

Proof. If we express $\boldsymbol{f}_m^{(n)}(\boldsymbol{\theta})$ through $\{\boldsymbol{d}_j^{(n)}\}$ and their derivatives, the derivatives will be of the form $\nabla^l\boldsymbol{d}_j^{(n)}$ where $l+j\leqslant m$. Therefore, the derivatives in $\nabla^r\boldsymbol{f}_m^{(n)}(\boldsymbol{\theta})$ for $r\in[0:2\mathcal{R}-m]$ will be of the form $\nabla^l\boldsymbol{d}_j^{(n)}$ where $l+j\leqslant 2\mathcal{R}$. So the result follows immediately from Lemma B.1. □

Lemma D.2. *For all $r\in[1:\mathcal{R}]$*

$$\boldsymbol{\theta}(t_{n+1})=\boldsymbol{\theta}(t_n)+\sum_{j=1}^{r}h^j\boldsymbol{d}_j^{(n)}(\boldsymbol{\theta}(t_n))+O(h^{r+1}). \tag{68}$$

Proof. Differentiating, we get the exact equality on $t \in [t_n, t_{n+1}]$

$$\frac{\mathrm{d}^i \boldsymbol{\theta}}{\mathrm{d}t^i}(t) = \sum_{k_1,\dots,k_i=1}^{\mathcal{R}} h^{k_1+\dots+k_i-i} (D_{k_1}^{(n)} \dots D_{k_{i-1}}^{(n)} \boldsymbol{f}_{k_i}^{(n)})(\boldsymbol{\theta}(t)), \quad 1 \leqslant i \leqslant \mathcal{R}+1.$$

Taylor's theorem gives

$$\begin{aligned}
\boldsymbol{\theta}(t_{n+1}) &= \boldsymbol{\theta}(t_n) + \sum_{i=1}^{\mathcal{R}} \frac{h^i}{i!} \frac{\mathrm{d}^i \boldsymbol{\theta}}{\mathrm{d}t^i}(t_n^+) + \frac{h^{\mathcal{R}+1}}{(\mathcal{R}+1)!} \frac{\mathrm{d}^{\mathcal{R}+1} \boldsymbol{\theta}}{\mathrm{d}t^{\mathcal{R}+1}}(\tilde{t}) \\
&= \boldsymbol{\theta}(t_n) + \sum_{i=1}^{\mathcal{R}} \frac{1}{i!} \sum_{k_1,\dots,k_i=1}^{\mathcal{R}} h^{k_1+\dots+k_i} (D_{k_1}^{(n)} \dots D_{k_{i-1}}^{(n)} \boldsymbol{f}_{k_i}^{(n)})(\boldsymbol{\theta}(t_n)) \\
&\quad + \frac{1}{(\mathcal{R}+1)!} \sum_{k_1,\dots,k_{\mathcal{R}+1}=1}^{\mathcal{R}} h^{k_1+\dots+k_{\mathcal{R}+1}} (D_{k_1}^{(n)} \dots D_{k_{\mathcal{R}}}^{(n)} \boldsymbol{f}_{k_{\mathcal{R}+1}}^{(n)})(\boldsymbol{\theta}(\tilde{t})) \\
&= \boldsymbol{\theta}(t_n) + \sum_{i=1}^{\mathcal{R}} \frac{1}{i!} \sum_{\substack{1 \leqslant k_1,\dots,k_i \leqslant \mathcal{R} \\ i \leqslant k_1+\dots+k_i \leqslant \mathcal{R}}} h^{k_1+\dots+k_i} (D_{k_1}^{(n)} \dots D_{k_{i-1}}^{(n)} \boldsymbol{f}_{k_i}^{(n)})(\boldsymbol{\theta}(t_n)) \\
&\quad + \sum_{i=1}^{\mathcal{R}} \frac{1}{i!} \sum_{\substack{1 \leqslant k_1,\dots,k_i \leqslant \mathcal{R} \\ \mathcal{R}+1 \leqslant k_1+\dots+k_i \leqslant i\mathcal{R}}} h^{k_1+\dots+k_i} (D_{k_1}^{(n)} \dots D_{k_{i-1}}^{(n)} \boldsymbol{f}_{k_i}^{(n)})(\boldsymbol{\theta}(t_n)) \\
&\quad + \frac{1}{(\mathcal{R}+1)!} \sum_{k_1,\dots,k_{\mathcal{R}+1}=1}^{\mathcal{R}} h^{k_1+\dots+k_{\mathcal{R}+1}} (D_{k_1}^{(n)} \dots D_{k_{\mathcal{R}}}^{(n)} \boldsymbol{f}_{k_{\mathcal{R}+1}}^{(n)})(\boldsymbol{\theta}(\tilde{t})) \\
&\overset{\text{(a)}}{=} \boldsymbol{\theta}(t_n) + \sum_{i=1}^{\mathcal{R}} \frac{1}{i!} \sum_{j=i}^{\mathcal{R}} h^j \sum_{\substack{k_1,\dots,k_i \geqslant 1 \\ k_1+\dots+k_i=j}} (D_{k_1}^{(n)} \dots D_{k_{i-1}}^{(n)} \boldsymbol{f}_{k_i}^{(n)})(\boldsymbol{\theta}(t_n)) + O(h^{\mathcal{R}+1}) \\
&= \boldsymbol{\theta}(t_n) + \sum_{j=1}^{\mathcal{R}} h^j \sum_{i=1}^{j} \frac{1}{i!} \sum_{\substack{k_1,\dots,k_i \geqslant 1 \\ k_1+\dots+k_i=j}} (D_{k_1}^{(n)} \dots D_{k_{i-1}}^{(n)} \boldsymbol{f}_{k_i}^{(n)})(\boldsymbol{\theta}(t_n)) + O(h^{\mathcal{R}+1}) \\
&\overset{\text{(b)}}{=} \boldsymbol{\theta}(t_n) + \sum_{j=1}^{\mathcal{R}} h^j \boldsymbol{d}_j^{(n)}(\boldsymbol{\theta}(t_n)) + O(h^{\mathcal{R}+1}),
\end{aligned}$$

where $\tilde{t}$ is between t_n and t_{n+1}; in (a) we used Lemma D.1 and in (b) we used (32).

It is left to use the boundedness of $\boldsymbol{d}_j^{(n)}(\boldsymbol{\theta})$ in the region of interest (Lemma B.1). □

Lemma D.3. *The following global bound holds:*

$$\sup_{n \in [0:\lfloor T/h \rfloor]} \left\| \tilde{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}(t_n) \right\| \leqslant C_7 h^{\mathcal{R}},$$

where C_7 is some constant.

Proof. Note that $\boldsymbol{\theta}(0) = \tilde{\boldsymbol{\theta}}^{(0)}$, the local error bound is already proven in Lemma D.2, and each $\boldsymbol{d}_j^{(n)}(\cdot)$ is Lipschitz because their derivatives are bounded by Lemma B.1. So this result follows by the same standard argument as for Lemma B.3 with $\tilde{\boldsymbol{\theta}}^{(n)}$ replaced by $\boldsymbol{\theta}(t_n)$ and $\boldsymbol{\theta}^{(n)}$ replaced by $\tilde{\boldsymbol{\theta}}^{(n)}$. □

To conclude the proof of Corollary 5.1, it is left to combine Lemma D.3 with (17).

D.2 Proof of Corollary 5.3

Define $\{v_{m,l}^{(n)}\}$ by the recursion

$$\begin{aligned} v_{m,l}^{(n)} &= \sum_{b=0}^{n-l} \beta^b \sum_{l_1=1}^{l+b} v_{m-1,l_1}^{(n)}, \quad m \in \mathbb{Z}_{\geqslant 2}, \\ v_{1,l}^{(n)} &= \sum_{b=0}^{n-l} \beta^b. \end{aligned} \tag{69}$$

It is immediate from Theorem 4.1 that $v_m^{(n)} \equiv v_{m,1}^{(n)}$.

Next, define $c_m \in \tilde{\mathcal{A}}[m]$ the chain with m vertices (defined in Section 1.5). Applying (26) with $l = 1$ gives

$$\nabla^2 L \boldsymbol{E}_{c_{m-1},1}^{(n)} + \beta \boldsymbol{E}_{c_m,2}^{(n)} = \boldsymbol{E}_{c_m,1}^{(n)}.$$

By Lemma 4.2 with $l = a = 1$, we have also

$$\boldsymbol{E}_{c_m,2}^{(n)} = \sum_{m \in m_{c_m}} \boldsymbol{E}_{c_m,m,a}^{(n-1\to n)}.$$

Combining and using the definition of $\boldsymbol{E}_{c_m,m,a}^{(n-1\to n)}$ gives (36).

It is clear from (36) that for each fixed m the sequence $v_m^{(n)}$ can be bounded by a constant not depending on n (but depending on m), and from (69) that each $v_{m,l}^{(n)}$ (and in particular $v_m^{(n)}$) is non-decreasing in n. Hence, there is a limit $v_m^{(\infty)} := \lim_{n\to\infty} v_m^{(n)}$. The initial condition is $v_1^{(\infty)} = \lim_{n\to\infty} \sum_{b=0}^{n-1} \beta^b = (1-\beta)^{-1}$. Equation (37) is now immediate from (36) by taking the limit $n \to \infty$.

From (37), we get that the generating function $g_\beta(x)$ needs to satisfy the quadratic equation $g_\beta(x) - (1-\beta)^{-1} = (1-\beta)^{-1} x g_\beta(x) + \beta(1-\beta)^{-1} x g_\beta(x)^2$ and have $g_\beta(0) = (1-\beta)^{-1}$, which gives (38).

The Narayana polynomials defined above satisfy the recurrence (e. g. [14])

$$N_m(\beta) = (1+\beta) N_{m-1}(\beta) + \beta \sum_{k=1}^{m-2} N_k(\beta) N_{m-k-1}(\beta)$$

for all $m \geqslant 3$, which means that $\tilde{N}_m(\beta) := N_m(\beta)/(1-\beta)^{2m+1}$ satisfy

$$\tilde{N}_m(\beta) = \frac{1+\beta}{(1-\beta)^2} \tilde{N}_{m-1}(\beta) + \frac{\beta}{1-\beta} \sum_{k=1}^{m-2} \tilde{N}_k(\beta) \tilde{N}_{m-k-1}(\beta).$$

But (37) can be rewritten as

$$v_{m'+1}^{(\infty)} = \frac{1+\beta}{(1-\beta)^2} v_{(m'-1)+1}^{(\infty)} + \frac{\beta}{1-\beta} \sum_{j'=1}^{m'-2} v_{j'+1}^{(\infty)} v_{(m'-j'-1)+1}^{(\infty)}, \quad m' \geqslant 3,$$

so $\tilde{N}_m(\beta)$ and $v_{m+1}^{(\infty)}$ satisfy the same recurrence. It is easy to see that their elements with $m \in \{1, 2\}$ are equal, concluding the result.

D.3 Proof of Corollary 5.6

From Theorem 4.1,

$$q_{m,l}^{(n)} = \sum_{b=0}^{n-l} \beta^b \left(\sum_{l_1=1}^{l+b} \frac{1-\beta^{n-l_1+1}}{1-\beta} \right)^{m-1},$$

so their limit is

$$q_{m,l}^{(\infty)} = \frac{1}{(1-\beta)^{m-1}} \sum_{b=0}^{\infty} \beta^b (l+b)^{m-1}.$$

In particular, $q_{m+1,1}^{(\infty)} = \frac{1}{(1-\beta)^m} \sum_{b=0}^{\infty} \beta^b (1+b)^m = \frac{1}{(1-\beta)^m} \frac{1}{(1-\beta)^{m+1}} A_m(\beta)$.

D.4 Proof of Corollary 5.8

Lemma D.4. *Neglecting non-principal terms, the following formula holds:*

$$\boldsymbol{f}_m^{(n)}(\boldsymbol{\theta}) = \sum_{l=1}^{m} \frac{(-1)^{l+1}}{l} \sum_{\substack{k_1,\ldots,k_l \geqslant 1 \\ k_1+\ldots+k_l=m}} \nabla \boldsymbol{d}_{k_1}^{(n)} \ldots \nabla \boldsymbol{d}_{k_{l-1}}^{(n)} \boldsymbol{d}_{k_l}^{(n)}(\boldsymbol{\theta}) + \text{NPT}.$$

Proof. In other words, we need to show

$$\boldsymbol{f}_m^{(n)}(\boldsymbol{\theta}) = \sum_{\mathfrak{s}: w(\mathfrak{s})=m} C_{l(\mathfrak{s})} \mathfrak{s} + \text{NPT},$$

with

$$C_l := \begin{cases} 1 & \text{if } l = 1, \\ (-1)^{l+1} \sum_{i=2}^{l} \frac{(-1)^i}{i!} \sum_{\substack{l_1,\ldots,l_i \geqslant 1 \\ l_1+\ldots+l_i=l}} \frac{1}{l_1 \ldots l_i} & \text{if } l \geqslant 2, \end{cases} \tag{70}$$

where by $\mathfrak{s}$ we denote expressions of the form $\nabla \boldsymbol{d}_{k_1}^{(n)} \ldots \nabla \boldsymbol{d}_{k_{l-1}}^{(n)} \boldsymbol{d}_{k_l}^{(n)}(\boldsymbol{\theta})$, by $w(\mathfrak{s})$ their weight, which is defined as $k_1 + \ldots + k_l$, and by $l(\mathfrak{s})$ their length, which is the number of nodes $\boldsymbol{d}_{k_i}^{(n)}$ (l in this case). For two such expressions $\mathfrak{s}_1$ and $\mathfrak{s}_2$, we will write $\mathfrak{s}_1\mathfrak{s}_2$ for their concatenation, for example if $\mathfrak{s}_1 = \boldsymbol{d}_3^{(n)}(\boldsymbol{\theta})$ and $\mathfrak{s}_2 = \nabla \boldsymbol{d}_1^{(n)}(\boldsymbol{\theta}) \boldsymbol{d}_2^{(n)}(\boldsymbol{\theta})$, then $\mathfrak{s}_1\mathfrak{s}_2 = \nabla \boldsymbol{d}_3^{(n)}(\boldsymbol{\theta}) \nabla \boldsymbol{d}_1^{(n)}(\boldsymbol{\theta}) \boldsymbol{d}_2^{(n)}(\boldsymbol{\theta})$.

We will argue by induction over m. For $m = 1$ the statement is obvious. Ignoring non-principal terms, we can write

$$\boldsymbol{f}_m^{(n)}(\boldsymbol{\theta}) = \boldsymbol{d}_m^{(n)}(\boldsymbol{\theta}) - \sum_{i=2}^{m} \frac{1}{i!} \sum_{\substack{k_1,\ldots,k_i \geqslant 1 \\ k_1+\ldots+k_i=m}} \nabla \boldsymbol{f}_{k_1}^{(n)} \ldots \nabla \boldsymbol{f}_{k_{i-1}}^{(n)} \boldsymbol{f}_{k_i}^{(n)}(\boldsymbol{\theta}) + \text{NPT}.$$

Now using the induction assumption, we rewrite it as

$$\begin{aligned} \boldsymbol{f}_m^{(n)}(\boldsymbol{\theta}) &= \boldsymbol{d}_m^{(n)}(\boldsymbol{\theta}) - \sum_{i=2}^{m} \frac{1}{i!} \sum_{\substack{k_1,\ldots,k_i \geqslant 1 \\ k_1+\ldots+k_i=m}} \sum_{\substack{\mathfrak{s}_1,\ldots,\mathfrak{s}_i: \\ w(\mathfrak{s}_1)=k_1,\ldots,w(\mathfrak{s}_i)=k_i}} C_{l(\mathfrak{s}_1)} \ldots C_{l(\mathfrak{s}_i)} \mathfrak{s}_1 \ldots \mathfrak{s}_i + \text{NPT} \\ &\overset{\text{(a)}}{=} \boldsymbol{d}_m^{(n)}(\boldsymbol{\theta}) - \sum_{i=2}^{m} \frac{1}{i!} \sum_{\substack{k_1,\ldots,k_i \geqslant 1 \\ k_1+\ldots+k_i=m}} \sum_{\substack{\mathfrak{s}_1,\ldots,\mathfrak{s}_i: \\ w(\mathfrak{s}_1)=k_1,\ldots,w(\mathfrak{s}_i)=k_i}} \frac{(-1)^{l(\mathfrak{s}_1)+\ldots+l(\mathfrak{s}_i)+i}}{l(\mathfrak{s}_1)\ldots l(\mathfrak{s}_i)} \mathfrak{s}_1 \ldots \mathfrak{s}_i + \text{NPT} \\ &= \boldsymbol{d}_m^{(n)}(\boldsymbol{\theta}) - \sum_{\mathfrak{s}: w(\mathfrak{s})=m} \sum_{i=2}^{m} \frac{(-1)^i}{i!} \sum_{\substack{\mathfrak{s}_1,\ldots,\mathfrak{s}_i \text{ non-empty:} \\ \mathfrak{s}=\mathfrak{s}_1\ldots\mathfrak{s}_i}} \frac{(-1)^{l(\mathfrak{s}_1)+\ldots+l(\mathfrak{s}_i)}}{l(\mathfrak{s}_1)\ldots l(\mathfrak{s}_i)} \mathfrak{s} + \text{NPT} \end{aligned}$$

$$\overset{\text{(b)}}{=} \sum_{\mathfrak{s}:w(\mathfrak{s})=m} C_{l(\mathfrak{s})}\mathfrak{s} + \text{NPT}.$$

where in (a) we used that $C_l = (-1)^{l+1}/l$ which is proven below, in (b) we used the definition of C_l given in (70).

It is left to prove $C_l = (-1)^{l+1}/l$, or, equivalently, for $l \geqslant 2$ we have

$$\sum_{i=1}^{l} \frac{(-1)^i}{i!} \sum_{\substack{l_1,\dots,l_i\geqslant 1\\ l_1+\dots+l_i=l}} \frac{1}{l_1\dots l_i} \overset{?}{=} 0. \tag{71}$$

To do this, note that

$$-\ln(1-x) = \sum_{n=1}^{\infty} \frac{x^n}{n},$$

which means that the coefficient before x^l in the power series of $\big[-\ln(1-x)\big]^i$ is

$$\sum_{\substack{l_1,\dots,l_i\geqslant 1\\ l_1+\dots+l_i=l}} \frac{1}{l_1\dots l_i},$$

and therefore the left-hand side of (71) is the coefficient before x^l in the power series

$$\sum_{i=1}^{\infty} \frac{(-1)^i}{i!}\big[-\ln(1-x)\big]^i = \exp\{\ln(1-x)\} - 1 = -x,$$

which is zero for $l \geqslant 2$. □

Equation (40) is immediate from Lemma D.4 and Corollary 5.3. Taking the limit as $n \to \infty$,

$$z_m^{(\infty)} := \lim_{n\to\infty} z_m^{(n)} = \sum_{l=1}^{m} \frac{(-1)^{l+1}}{l} \sum_{\substack{k_1,\dots,k_l\geqslant 1\\ k_1+\dots+k_l=m}} p_{k_1}^{(\infty)} \dots p_{k_l}^{(\infty)},$$

where

$$p_k^{(\infty)} := \begin{cases} -(1-\beta)^{-1}, & \text{if } k=1,\\ -\beta v_k^{(\infty)}, & \text{if } k \geqslant 2. \end{cases}$$

The generating function $\bar{g}_\beta(x) := \sum_{k=0}^{\infty} z_{k+1}^{(\infty)} x^k$ satisfies

$$[\bar{g}_\beta(x)]_k = \sum_{l=0}^{k} \frac{(-1)^l}{l+1}\big[\{-1-\beta g_\beta(x)\}^{l+1}\big]_{k-l},$$

where $[g(x)]_k$ denotes the coefficient before x^k in the power series of $g(x)$. Multiplying both sides by x^k, summing over k and changing the order of summation gives

$$\begin{aligned}\sum_{k=0}^{\infty}[\bar{g}_\beta(x)]_k x^k &= -\sum_{k=0}^{\infty}\sum_{l=0}^{k} \frac{1}{l+1}\big[\{1+\beta g_\beta(x)\}^{l+1}\big]_{k-l} x^k\\ &= -\sum_{l=0}^{\infty} \frac{1}{l+1} x^l \sum_{k=l}^{\infty}\big[\{1+\beta g_\beta(x)\}^{l+1}\big]_{k-l} x^{k-l}\end{aligned}$$

$$= -\sum_{l=0}^{\infty} \frac{1}{l+1} x^l \{1+\beta g_\beta(x)\}^{l+1} = -\frac{1}{x}\sum_{l=1}^{\infty} \frac{1}{l}\{x+\beta x g_\beta(x)\}^l$$
$$= \frac{1}{x}\ln\left(\frac{1+\beta-x+\sqrt{(1-\beta-x)^2-4\beta x}}{2}\right),$$

as desired.

E Averaging over Dataset Permutations

Let $M := n+1$ be the batch count.

Recall that

$$L^{(k)}(\boldsymbol{\theta}) = \frac{1}{B}\sum_{r=kB+1}^{kB+B} \ell_{\pi(r)}(\boldsymbol{\theta}), \quad k \in [0\,{:}\,n],$$

where $\{\ell_s\}_{s=1}^{(n+1)B}$ are per-sample losses and π is a random permutation of $[1\,{:}\,(n+1)B]$, distributed uniformly over all $((n+1)B)!$ such permutations.

Note that

$$\mathbb{E}_\pi \nabla^2(\ell_{\pi(1)} - L)\nabla(\ell_{\pi(2)} - L) = -\frac{\nabla \operatorname{tr}\boldsymbol{\Sigma}}{2(MB-1)},$$
$$\mathbb{E}_\pi \nabla^2(\ell_{\pi(1)} - L)\nabla(\ell_{\pi(1)} - L) = \frac{\nabla \operatorname{tr}\boldsymbol{\Sigma}}{2},$$

where $\boldsymbol{\Sigma}$ is the empirical covariance matrix of per-sample gradients (7).

Lemma E.1. *The following expressions hold:*

$$\mathbb{E}_\pi \nabla^2 L^{(0)}\nabla L^{(1)} = \nabla^2 L \nabla L - \frac{\nabla \operatorname{tr}\boldsymbol{\Sigma}}{2(MB-1)}, \quad \mathbb{E}_\pi \nabla^2 L^{(0)}\nabla L^{(0)} = \nabla^2 L \nabla L + \frac{M-1}{2(MB-1)}\nabla \operatorname{tr}\boldsymbol{\Sigma}.$$

Proof. Indeed,

$$\begin{aligned}\mathbb{E}_\pi \nabla^2 L^{(0)}\nabla L^{(1)} &= \mathbb{E}_\pi \frac{1}{B}\sum_{r=1}^{B} \nabla^2 \ell_{\pi(r)} \frac{1}{B}\sum_{s=B+1}^{2B} \nabla \ell_{\pi(s)} \\ &= \mathbb{E}_\pi \nabla^2 \ell_{\pi(1)} \nabla \ell_{\pi(2)} = \nabla^2 L \nabla L + \mathbb{E}_\pi \nabla^2(\ell_{\pi(1)} - L)\nabla(\ell_{\pi(2)} - L) \\ &= \nabla^2 L \nabla L - \frac{\nabla \operatorname{tr}\boldsymbol{\Sigma}}{2(MB-1)}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_\pi \nabla^2 L^{(0)}\nabla L^{(0)} &= \mathbb{E}_\pi \frac{1}{B}\sum_{r=1}^{B} \nabla^2 \ell_{\pi(r)} \frac{1}{B}\sum_{s=1}^{B} \nabla \ell_{\pi(s)} \\ &= \frac{1}{B}\mathbb{E}_\pi \nabla^2 \ell_{\pi(1)} \nabla \ell_{\pi(1)} + \frac{B-1}{B}\mathbb{E}_\pi \nabla^2 \ell_{\pi(1)} \nabla \ell_{\pi(2)} \\ &= \frac{1}{B}\left(\nabla^2 L \nabla L + \frac{\nabla \operatorname{tr}\boldsymbol{\Sigma}}{2}\right) + \frac{B-1}{B}\left(\nabla^2 L \nabla L - \frac{\nabla \operatorname{tr}\boldsymbol{\Sigma}}{2(MB-1)}\right) \\ &= \nabla^2 L \nabla L + \frac{M-1}{2(MB-1)}\nabla \operatorname{tr}\boldsymbol{\Sigma}\end{aligned}$$

completing the proof. □

Lemma E.2. *We have*

$$\mathbb{E}_\pi \boldsymbol{d}_2^{(n)}(\boldsymbol{\theta}) = (C_{0,0}^{(n)}(\beta) + C_{0,1}^{(n)}(\beta))\nabla^2 L \nabla L + \left(C_{0,0}^{(n)}(\beta)\frac{M-1}{2(MB-1)} - C_{0,1}^{(n)}(\beta)\frac{1}{2(MB-1)}\right)\nabla \operatorname{tr}\boldsymbol{\Sigma}$$
$$= \left(-\frac{\beta}{(1-\beta)^3} + o_n(1)\right)\nabla^2 L \nabla L + \left(-\frac{\beta}{2(1-\beta)^2(1+\beta)} + o_n(1)\right)\frac{\nabla \operatorname{tr}\boldsymbol{\Sigma}}{B},$$

where $o_n(1)$ are terms that go to zero as $n \to \infty$ (for fixed β) regardless of $B \in [1:n+1]$, and

$$C_{0,0}^{(n)}(\beta) := -\frac{\beta[1-\beta^n(1+\beta)+\beta^{2n+1}]}{(1-\beta)^2(1+\beta)}, \tag{72}$$
$$C_{0,1}^{(n)}(\beta) := \frac{-2\beta^2 + 2n(1-\beta^2)\beta^{n+1} + 2\beta^{2n+2}}{(1-\beta)^3(1+\beta)}. \tag{73}$$

Proof. Using Lemma E.1, we can write

$$\begin{aligned}
\mathbb{E}_\pi \boldsymbol{d}_2^{(n)}(\boldsymbol{\theta}) &= -\beta \mathbb{E}_\pi \sum_{b=0}^{n-1}\beta^b \sum_{l'=1}^{b+1}\sum_{b'=0}^{n-l'}\beta^{b'}\nabla^2 L^{(n-1-b)}\nabla L^{(n-l'-b')} \\
&= C_{0,0}^{(n)}(\beta)\mathbb{E}_\pi \nabla^2 L^{(0)}\nabla L^{(0)} + C_{0,1}^{(n)}(\beta)\mathbb{E}_\pi \nabla^2 L^{(0)}\nabla L^{(1)} \\
&= C_{0,0}^{(n)}(\beta)\left(\nabla^2 L \nabla L + \frac{M-1}{2(MB-1)}\nabla \operatorname{tr}\boldsymbol{\Sigma}\right) + C_{0,1}^{(n)}(\beta)\left(\nabla^2 L \nabla L - \frac{\nabla \operatorname{tr}\boldsymbol{\Sigma}}{2(MB-1)}\right) \\
&= (C_{0,0}^{(n)}(\beta) + C_{0,1}^{(n)}(\beta))\nabla^2 L \nabla L + \left(C_{0,0}^{(n)}(\beta)\frac{M-1}{2(MB-1)} - C_{0,1}^{(n)}(\beta)\frac{1}{2(MB-1)}\right)\nabla \operatorname{tr}\boldsymbol{\Sigma} \\
&= \left(-\frac{\beta}{(1-\beta)^3} + o_n(1)\right)\nabla^2 L \nabla L + \left(-\frac{\beta}{2(1-\beta)^2(1+\beta)} + o_n(1)\right)\frac{\nabla \operatorname{tr}\boldsymbol{\Sigma}}{B},
\end{aligned}$$

where $C_{0,0}^{(n)}(\beta)$ and $C_{0,1}^{(n)}(\beta)$ can be calculated as

$$C_{0,0}^{(n)}(\beta) := -\beta\sum_{b=0}^{n-1}\beta^b\sum_{l'=1}^{b+1}\beta^{b+1-l'} = -\frac{\beta[1-\beta^n(1+\beta)+\beta^{2n+1}]}{(1-\beta)^2(1+\beta)} \underset{n\to\infty}{\longrightarrow} -\frac{\beta}{(1-\beta)^2(1+\beta)},$$
$$\begin{aligned}
C_{0,1}^{(n)}(\beta) &:= -\beta\sum_{b=0}^{n-1}\beta^b\sum_{l'=1}^{b+1}\sum_{b'=0}^{n-l'}\beta^{b'} - C_{0,0}^{(n)}(\beta) \\
&= \frac{-2\beta^2 + 2n(1-\beta^2)\beta^{n+1} + 2\beta^{2n+2}}{(1-\beta)^3(1+\beta)} \underset{n\to\infty}{\longrightarrow} -\frac{2\beta^2}{(1-\beta)^3(1+\beta)}.
\end{aligned}$$

□

References

[1] C. Bai, L. Guo, Y. Sheng, and R. Tang. Post-groups, (lie-)butcher groups and the yang–baxter equation. *Mathematische Annalen*, 388(3):3127–3167, 2024. doi: 10.1007/s00208-023-02592-z.

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

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

[4] J. Bernstein and L. Newhouse. Old optimizer, new norm: An anthology. In *OPT 2024: Optimization for Machine Learning*, 2024. URL https://openreview.net/forum?id=ux18f5nOpD.

[5] W.-J. Beyn. Numerical methods for dynamical systems. *Advances in Numerical Analysis*, 1:175–236, 1991.

[6] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

[7] J. C. Butcher. An algebraic theory of integration methods. *Mathematics of Computation*, 26(117):79–106, 1972.

[8] M. Calvo, A. Murua, and J. Sanz-Serna. Modified equations for odes. *Contemporary Mathematics*, 172:63–63, 1994.

[9] M. D. Cattaneo and B. Shigida. How memory in optimization algorithms implicitly modifies the loss, 2025. URL https://arxiv.org/abs/2502.02132.

[10] M. D. Cattaneo, J. M. Klusowski, and B. Shigida. On the implicit bias of Adam. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 5862–5906. PMLR, 2024. URL https://proceedings.mlr.press/v235/cattaneo24a.html.

[11] P. Chartier, E. Hairer, and G. Vilmart. Numerical integrators based on modified differential equations. *Mathematics of Computation*, 76(260):1941–1953, 2007.

[12] P. Chartier, E. Hairer, and G. Vilmart. Algebraic structures of b-series. *Foundations of Computational Mathematics*, 10(4):407–427, 2010.

[13] J. Cohen, S. Kaur, Y. Li, J. Z. Kolter, and A. Talwalkar. 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.

[14] C. Coker. Enumerating a class of lattice paths. *Discrete mathematics*, 271(1-3):13–28, 2003.

[15] T. Eirola. Aspects of backward error analysis of numerical odes. *Journal of Computational and Applied Mathematics*, 45(1-2):65–73, 1993.

[16] C. L. Ernst Hairer and G. Wanner. *Geometric numerical integration*. Springer-Verlag, Berlin, 2 edition, 2006. ISBN 978-3-540-30666-5. doi: 10.1007/3-540-30666-8.

[17] M. Farazmand. Multiscale analysis of accelerated gradient methods. *SIAM Journal on Optimization*, 30(3):2337–2354, 2020. URL https://epubs.siam.org/doi/abs/10.1137/18M1203997.

[18] W. G. Faris. Rooted tree graphs and the butcher group: Combinatorics of elementary perturbation theory, 2021. URL https://arxiv.org/abs/2101.09364.

[19] K. Feng. Formal power series and numerical algorithms for dynamical systems. In *Proc. International Conference on Scientific Computation, Hangzhou, China*, pages 28–35, 1991.

[20] N. Fenichel and J. Moser. Persistence and smoothness of invariant manifolds for flows. *Indiana University Mathematics Journal*, 21(3):193–226, 1971.

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

[22] A. Ghosh, H. Lyu, X. Zhang, and R. Wang. Implicit regularization in heavy-ball momentum accelerated stochastic gradient descent. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ZzdBhtEH9yB.

[23] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. The MIT Press, 2016.

[24] D. F. Griffiths and J. M. Sanz-Serna. On the scope of the method of modified equations. *SIAM Journal on Scientific and Statistical Computing*, 7(3):994–1008, 1986.

[25] V. Gupta, T. Koren, and Y. Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1842–1850. PMLR, 2018. URL https://proceedings.mlr.press/v80/gupta18a.html.

[26] E. Hairer. Backward analysis of numerical integrators and symplectic methods. *Annals of Numerical Mathematics*, 1:107–132, 1994. URL https://archive-ouverte.unige.ch/unige:12640.

[27] E. Hairer and G. Wanner. On the butcher group and general multi-value methods. *Computing*, 13(1):1–15, 1974.

[28] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[29] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 558–567, 2019.

[30] M. W. Hirsch, C. C. Pugh, and M. Shub. Invariant manifolds. *Bulletin of the American Mathematical Society*, 76(5):1015–1019, 1970.

[31] A. Kerber. *Applied Finite Group Actions*, volume 19 of *Algorithms and Combinatorics*. Springer Berlin Heidelberg, Berlin and Heidelberg, 2 edition, 1999. ISBN 978-3-540-65941-9. doi: 10.1007/978-3-662-11167-3.

[32] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=H1oyRlYgg.

[33] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2017. URL https://arxiv.org/abs/1412.6980.

[34] N. B. Kovachki and A. M. Stuart. Continuous time analysis of momentum methods. *Journal of Machine Learning Research*, 22(17):1–40, 2021. URL http://jmlr.org/papers/v22/19-466.html.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[36] D. Masters and C. Luschi. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*, 2018.

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

[38] B. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(64)90137-5. URL https://www.sciencedirect.com/science/article/pii/0041555364901375.

[39] S. J. Prince. *Understanding Deep Learning*. The MIT Press, 2023. URL http://udlbook.com.

[40] S. Reich. Backward error analysis for numerical integrators. *SIAM Journal on Numerical Analysis*, 36(5):1549–1570, 1999.

[41] M. Rosca, Y. Wu, C. Qin, and B. Dherin. On a continuous time model of gradient descent dynamics and instability in deep learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=EYrRzKPinA.

[42] B. Shi, S. S. Du, M. I. Jordan, and W. J. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, 195(1):79–148, 2022.

[43] S. Smith, E. Elsen, and S. De. On the generalization benefit of noise in stochastic gradient descent. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9058–9067. PMLR, 2020. URL https://proceedings.mlr.press/v119/smith20a.html.

[44] S. L. Smith, B. Dherin, D. Barrett, and S. De. 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.

[45] A. M. Stuart and A. R. Humphries. *Dynamical systems and numerical analysis*, volume 2. Cambridge University Press, 1998.

[46] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 2013. PMLR. URL https://proceedings.mlr.press/v28/sutskever13.html.

[47] R. F. Warming and B. J. Hyett. The modified equation approach to the stability and accuracy analysis of finite-difference methods. *Journal of Computational Physics*, 14(2): 159–179, 1974.

[48] S. Wiggins. *Normally hyperbolic invariant manifolds in dynamical systems*, volume 105. Springer Science & Business Media, 2013.

[49] J. H. Wilkinson. Error analysis of floating-point computation. *Numerische Mathematik*, 2 (1):319–340, 1960.

[50] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017.

[51] Y. Zhao, H. Zhang, and X. Hu. Penalizing gradient norm for efficiently improving generalization in deep learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26982–26992. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhao22i.html.