Convergence of Online Adaptive and Recurrent Optimization Algorithms

Pierre-Yves Massé, Yann Ollivier

Convergence of Online Adaptive and Recurrent Optimization Algorithms

Pierre-Yves Massé [*] Yann Ollivier [†]

December 16, 2020

Abstract

We prove local convergence of several notable gradient descent algorithms used in machine learning, for which standard stochastic gradient descent theory does not apply directly. This includes, first, online algorithms for recurrent models and dynamical systems, such as *Real-time recurrent learning* (RTRL) [Jaeger, 2002; Pearlmutter, 1995] and its computationally lighter approximations NoBackTrack [Ollivier et al., 2015] and UORO [Tallec and Ollivier, 2018]; second, several adaptive algorithms such as RMSProp, online natural gradient, and Adam with $\beta^2 \to 1$.

Despite local convergence being a relatively weak requirement for a new optimization algorithm, no local analysis was available for these algorithms, as far as we knew. Analysis of these algorithms does not immediately follow from standard stochastic gradient (SGD) theory. In fact, Adam has been proved to lack local convergence in some simple situations [Reddi et al., 2018]. For recurrent models, online algorithms modify the parameter while the model is running, which further complicates the analysis with respect to simple SGD.

Local convergence for these various algorithms results from a single, more general set of assumptions, in the setup of learning dynamical systems online. Thus, these results can cover other variants of the algorithms considered.

We adopt an "ergodic" rather than probabilistic viewpoint, working with empirical time averages instead of probability distributions. This is more data-agnostic and creates differences with respect to standard SGD theory, especially for the range of possible learning rates. For instance, with cycling or per-epoch reshuffling over a finite dataset instead of pure i.i.d. sampling with replacement, empirical averages of gradients converge at rate $1/T$ instead of $1/\sqrt{T}$ (cycling acts as a variance reduction method), theoretically allowing for larger learning rates than in SGD.

Contents

[*] Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague

[†] Facebook Artificial Intelligence Research, Paris

1 Introduction

We consider, from a machine learning perspective, the problem of optimizing in real time the parameters of a dynamical system so that its behavior optimizes some criterion over time. This problem has a longstanding history, especially for linear systems of small to moderate dimension [Ljung and Söderström, 1984], encompassing many classical recursive control problems like the steering of a ship, the short-term prediction of power demand or the transmission of speech through limited capacity transmission channels [Ljung and Söderström, 1984]. Examples that have attracted more recent attention include recurrent models in machine learning (recurrent neural networks), used to represent time-structured or sequentially-structured data. Even when the data have no time structure, a dynamical system can also represent the internal state of a machine learning algorithm, such as momentum variables in extensions of stochastic gradient descent.

We focus on *online* (or *real-time*) algorithms, that are able to update their state or predictions as each new observation arrives, at an algorithmic and memory cost that does not grow with the amount of data processed. Quoting Pearlmutter [1995], "An online, exact, and stable, but computationally expensive, procedure for determining the derivatives of functions of the states of a dynamic system with respect to that system's internal parameters has been discovered and applied to recurrent neural networks a number of times [...], *real time recurrent learning*, RTRL. Like BPTT, the technique was known and applied to other sorts of systems since the 1950s".

Thus, RTRL is the algorithm that adapts the parameters of a dynamical system by gradient descent over some criterion at each time step, in real time while the system is running. RTRL has both practical and theoretical shortcomings: First, its computational burden is prohibitive even for moderately-dimensioned systems. This has led to several lightweight approximations based on stochastic approximation, such as *NoBackTrack* and its extensions *UORO* and *Kronecker-factored RTRL* [Ollivier et al., 2015; Tallec and Ollivier, 2018; Mujika et al., 2018]. For relatively short data sequences (such as sentences in natural language processing), non-online algorithms such as backpropagation through time (BPTT) are usually preferred. *Truncated BPTT* is an approximation of BPTT that works online by maintaining a fixed-length memory of recent data.

Second, as far as we know, no proof of convergence, even local, has been given for these algorithms. A key feature of online algorithms is that the parameters of the dynamical system are updated while the system is running. Intuitively this is only a second-order phenomenon if learning rates are small; but this still complicates the analysis substantially.

We provide such a proof of local convergence for RTRL, and for some of its variants. Moreover, the results carry over to other non-recurrent machine learning algorithms, such as RMSProp, Adam, or online natural gradient. The dynamical system viewpoint is used to handle the internal state of these algorithms.

More precisely, we prove local convergence of various algorithms for recurrent and non-recurrent systems:

1. Real-time recurrent learning (RTRL) (Theorem 2.28);

2. Truncated backpropagation through time (TBPTT), provided the truncation length is slowly increased at a rate related to the main learning rate (Theorem 3.14);

3. Unbiased stochastic approximations to RTRL: NoBackTrack and UORO (Corollary 3.23);

4. Stochastic gradient descent with momentum and any parameter-dependent or adaptive preconditioning, where a definite positive preconditioning matrix is estimated online from the data (Corollary 3.10). This covers algorithms such as RMSProp and Adam with the preconditioner updated at the same rate as the main learning rate (Corollaries 3.8, 3.10), a natural gradient descent with the Fisher matrix estimated online at the same rate as the main learning rate (Corollary 3.8), or the extended Kalman filter in the static case (for estimating the state of a fixed system via nonlinear noisy measurements). Results for RMSProp and Adam are known (e.g., Zou et al. [2019]); our result is less precise but more general as it covers any kind of adaptive preconditioning rather than a specific algorithm.

We give a more precise overview of results in Section 2.

Although local convergence is a relatively weak property for an algorithm (compared to global convergence results obtained in convex situations), no local analysis was available for these algorithms apart from RMSProp and Adam, as far as we know. Our original project was to prove local convergence for NoBackTrack and UORO based on a convergence proof for RTRL, but we could locate no such existing proof. Convergence of these algorithms does not immediately follow from

standard stochastic gradient (SGD) theory. In fact, Adam has been proved to lack local convergence if its hyperparameter β^2 is fixed [Reddi et al., 2018] (convergence occurs with a time-dependent $\beta^2 \to 1$ so that the preconditioner is averaged over more and more samples).

Importantly, we prove local convergence under *local* assumptions: we do not assume that the model or system is well-behaved out of some ball of finite radius. We believe this reflects problems encountered in practice, when large steps can be difficult to recover from if the system parameters reach an unsafe zone. Thus, local convergence under local assumptions can be harder to prove than global convergence under global assumptions.

Most data to which recurrent models are applied cannot reasonably be assumed to be fully Markovian (natural text has arbitrary long-term dependencies, time series may be non-time-homogeneous). So we adopt a more data-agnostic viewpoint, reasoning on ergodic properties of an individual data sequence rather than on expectations. A local minimum is defined as a parameter value that achieves locally best loss on average over time (Assumption 2.11.a). Ergodic properties of gradients, averaged over time, replace expectations, and the standard stochastic case is recovered by proving that the assumptions hold with probability one. This per-trajectory viewpoint with local assumptions leads to several differences with respect to standard SGD theory, mostly relating to learning rates:

- When dealing with finite datasets, the per-trajectory viewpoint emphasizes specific properties of cycling through the data samples or reshuffling at every epoch, as opposed to the pure SGD method of selecting a sample at random at every step: cycling acts as a variance reduction method (ensuring each sample is selected exactly once within N steps, where N is the size of the dataset). This results in larger possible learning rates: with cycling or random reshuffling, learning rates $\eta_t \propto 1/t^b$ with any $0 < b \leq 1$ are suitable, as opposed to $1/2 < b \leq 1$ in classical Robbins–Monro theory (Corollary 3.4). This opens the door to more elaborate variance reduction methods in SGD.

- On the contrary, in a non-recurrent, online i.i.d. setting with an infinite dataset, our results are sometimes suboptimal: depending on which moments of the noise are finite, we may get more constraints on the learning rate (Section 3.1.6). This is presumably because the ergodic Assumption 2.11.a does not capture the full randomness of an i.i.d. sequence of samples.

- In a dynamical system setting, the stepsizes η_t for the gradient descent must vary smoothly in time, to avoid spurious correlations between the stepsize and the state of the system, which would bias the gradient descent. This is stricter than the classical Robbins–Monro criterion [Robbins and Monro, 1951]. (For instance, if a dynamical system exhibits periodic phenomena of period 2, and if η_t vanishes for even values of t, the gradient descent using η_t may be strongly biased and diverge.) We avoid this issue by assuming the learning rates behave like $1/t^b$ for some $b > 0$. (A more general homogeneity condition on the learning rates is given in Assumption 4.19.)

Finally, we treat adaptive preconditioning (RMSProp, Adam, online natural gradient...) by viewing the preconditioner as part of the parameter to be estimated. The corresponding update does not follow the gradient of a loss function; indeed, unlike a Hessian, the Jacobian Λ of the expected update is not a symmetric, definite

positive matrix. But its eigenvalues still have positive real part (Sections 3.1.4–3.1.5), which is sufficient to apply the standard Lyapunov theory for stable matrices (Appendix A), and prove local convergence. A reminder on positive-stable matrices is included in Appendix A. Focusing on positive-stable matrices instead of positive-definite Hessians is not new in machine learning: see for instance the classical paper Polyak and Juditsky [1992] on averaged stochastic gradient descent.

Some related work. Learning of recurrent models and dynamical systems is not a new topic (see historical references in Pearlmutter [1995]; Ljung and Söderström [1984]), and it is impossible to be exhaustive. For dynamical systems, an in-depth reference is Ljung and Söderström [1984], which discusses algorithms for learning a dynamical system online, largely focusing on the linear case. For *linear* dynamical systems, more precise results are available. For instance, Hardt et al. [2016] prove global convergence of non-online stochastic gradient descent on linear systems, provided the matrix defining the system is parameterized in a particular way based on its characteristic polynomial. For nonlinear systems, Benveniste et al. [1990] present results for stochastic gradient descent in very general time-dependent systems under strong Markovian assumptions, but it is not clear how to cast the algorithms studied here in their framework and how to check the technical assumptions.

Our overall approach to the proofs follows the classical ODE method for the analysis of SGD around a local optimum [Ljung, 1977; Benveniste et al., 1990; Borkar and Meyn, 2000; Kushner and Yin, 2003; Borkar, 2009]. The ODE approach views the optimization process on the parameter as an approximation of a continuous-time, noise-free "ideal" gradient descent, whose timescale is defined by the step sizes of the algorithm. Thus, our analysis is based on bounding the difference between the true system and an idealized system, linearized close to the optimum and with the noise averaged out. A central role is played by the Jacobian Λ of the optimization algorithm around the local optimum: this is the Hessian of the loss for simple SGD, but is a more complicated, non-symmetric matrix in adaptive algorithms such as Adam (Sections 3.1.4 and 3.1.5). Following the standard theory of dynamical systems, the idealized system on the parameter will converge when all eigenvalues of this matrix have positive real part (namely, in the simplest case, when the Hessian of the loss is positive definite).

For simple, non-recurrent SGD on general (non-convex) loss functions, one of the cleanest results is probably still Bertsekas and Tsitsiklis [2000], which proves convergence to a local minimum (which may be at infinity) under mild global assumptions (globally Lipschitz gradients, noise bounded by the gradient norm): namely, the loss converges and the gradient of the loss converges to 0. This does not cover either dynamical systems or algorithms other than simple SGD. Moreover, contrary to this work, we only make local assumptions.

For adaptive gradient descent algorithms such as Adam and RMSProp, convergence results already exist. Our result (Corollary 3.10) is less precise but more general, in that it covers any kind of adaptive preconditioning rather than specific algorithms, also covering the online natural gradient, for example. Among others, Zou et al. [2019, Corollary 10] prove a convergence result for Adam and RMSProp over a wide range of hyperparameters, together with finite-time bounds in expectation. We refer to Défossez et al. [2020] for more up-to-date finite-time bounds for Adam, and for additional references. These results, and ours, use a time-dependent Adam hyperparameter $\beta^2 \to 1$ so that square gradients are averaged over more and

more samples. On the other hand, Reddi et al. [2018] show divergence of Adam with fixed hyperparameters β^1 and β^2 when cycling over a finite dataset, contradicting an earlier convergence claim in Kingma and Ba [2014].

Convergence of algorithms with adaptive preconditioners (RMSProp, Adam, online natural gradient) with $\beta^2 \to 1$ could also probably be proved using two-timescale methods (see for instance Tadic [2004]). However, two-timescale methods, as the name suggests, require different timescales for the learning rate and the adaptive preconditioner: the main learning rate should be smaller than the rate at which the preconditioner is updated (which itself should tend to 0). Our result (Corollary 3.8, Corollary 3.10) lifts this restriction by letting the main learning rate be as large as the update rate of the preconditioner.

Finally, empirical differences between cycling over a dataset or random per-epoch reshuffling as opposed to pure i.i.d. sampling from the dataset have been observed for some time [Bottou, 2009]. Some quantitative results for convex functions are available [Gürbüzbalaban et al., 2015], showing improved convergence for random reshuffling compared to SGD. But these results still require learning rates smaller than $1/\sqrt{t}$, contrary to ours.

Structure of the text. In Section 2, we present an overview of the results, introduce the notation for dynamical systems, and present the standard RTRL algorithm as well as several generalizations that will encompass more algorithms. We then state the local convergence theorem for these extended RTRL algorithms, after discussing the technical assumptions. Section 3 contains several examples and applications, both recurrent and non-recurrent: simple SGD and the influence on learning rates of cycling over a dataset versus pure i.i.d. sampling, SGD with adaptive preconditioning and with momentum (including Adam), the original RTRL algorithm, truncated backpropagation through time with increasing truncation, and the NoBackTrack and UORO algorithms. We then proceed to the proof: in Sections 4 and 5 we go to a more abstract setting using an extended dynamical system that contains all the variables maintained by an algorithm; in this more abstract setting, we use the ODE method to quantify the discrepancy between the ideal continuous-time, noise-free gradient descent and the actual online gradient descent for the dynamical system. In Sections 6 and 7 we bridge the abstract setting and the concrete algorithms; especially, we check that all properties needed for Section 4 are indeed satisfied for the practical algorithms.

Acknowledgements. The authors would like to thank Léon Bottou, Joan Bruna, and Aaron Defazio for pointing us to relevant references. The work of the first author was partially supported by the European Regional Development Fund under the project IMPACT (reg. no. CZ.02.1.01/0.0/0.0/15_003/0000468).

2 Recurrent Models and the RTRL Algorithm

2.1 Overview of RTRL

We consider a dynamical system parameterized by $\theta \in \Theta$, whose state $s_t \in \mathcal{S}_t$ at time $t \geq 1$ is subjected to the evolution equation

$$s_t = \mathbf{T}_t(s_{t-1}, \theta), \tag{1}$$

with some transition operator $\mathbf{T}_t$. At each time, we are given a loss function $\mathcal{L}_t(s_t)$, and our objective is to optimize the parameter θ as to minimize the average loss function $\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_t(s_t)$ over some large time interval $T \to \infty$, in an online manner. Formal definitions are given in Section 2.3 below.

This formalism encompasses non-recurrent situations, by letting $\mathbf{T}_t$ be independent of s_{t-1}. For instance, consider a regression problem $y = F_\theta(x)$, with training dataset $(x_t, y_t)_{t\in[1;T]}$, and a loss function $\ell(y, y_t)$ such as $\ell(y, y_t) = \|y - y_t\|^2$. This can be represented by identifying the state s with y, namely, setting

$$\mathbf{T}_t(s_{t-1}, \theta) := F_\theta(x_t), \qquad \mathcal{L}_t(s) := \ell(s, y_t). \tag{2}$$

The operators $\mathbf{T}_t$ and $\mathcal{L}_t$ depend on the data. In this non-recurrent case, the RTRL algorithm will reduce to standard stochastic gradient descent.

Another typical system we have in mind is a recurrent model with internal state s_t, where the time-dependent transition operator

$$\mathbf{T}_t(s_{t-1}, \theta) := F_\theta(s_{t-1}, x_t) \tag{3}$$

is defined via a time-independent function F with some input x_t as an argument.[1] Once more, we define a loss function $\mathcal{L}_t(s) := \mathcal{L}(s, y_t)$ where $\mathcal{L}(s, y_t)$ typically measures the loss between a value y_t to be predicted, and some part of the state s that encodes the prediction on y_t.

Thus, the data (x_t, y_t) is encoded in (1) via the time dependency of $\mathbf{T}_t$ and $\mathcal{L}_t$. Recurrent neural networks (RNNs) fit this framework; for instance, a simple RNN model is

$$s_t = \text{sigmoid}\left(W s_{t-1} + W' x_t + B\right), \tag{4}$$

where W, W' and B are matrices or vectors of suitable dimensions, and where $\theta = (W, W', B)$.

Thus, when $\mathbf{T}_t$ is defined this way, we assume the sequence of inputs to be fixed once and for all,[2] and make no direct assumption on its nature. In particular, we do not make explicit stochastic assumptions on the data, but we assume they satisfy ergodic-like properties, expressed as empirical averages over time (see Assumption 2.11.a).

Jaeger's tutorial [Jaeger, 2002] presents several classical recurrent training algorithms. The most widely used is backpropagation through time. One of its important drawbacks is the need to store and pass through the complete sequence of past observations every time a new observation (x_{t+1}, y_{t+1}) becomes available: it is not possible to process online newly arrived inputs coming from a stream of data. On the other hand, the RTRL algorithm may be used online, but has much heavier computational and memory requirements. Let us now describe it.

The RTRL algorithm conducts an approximate gradient descent on the parameter of the dynamical system to be trained. The state s_t of the system at each time depends on the parameter used and on the initial state. By composition, the loss above on s_t may thus be viewed as a loss on the parameter and the initial state.

[1] This describes an online system with unbounded time. Finite-length training sequences are covered by separating them by end-of-sentence input symbol $x_t^\dagger$ and defining $F_\theta(s_{t-1}, x_t^\dagger) := s_0^*$ to reset the system to state s_0^* after each sequence, with notation as in (3). This preserves all our assumptions below.

[2] This means in particular that the system is non-adversarial: the inputs and targets do not change based on the behavior of the algorithm.

(We will omit the initial state for now.) We write $\mathcal{L}_t(s_t)$ for the original loss on the state of the system at time t, and $\mathcal{L}_{\rightsquigarrow t}(\theta)$ for the resulting loss at time t, seen as a function of the parameter via running the system up to time t with parameter θ (Definition 2.7). In computational terms, $\mathcal{L}_{\rightsquigarrow t}$ corresponds to the loss of the whole computational graph leading to $\mathcal{L}_t$.

The derivative of $\mathcal{L}_{\rightsquigarrow t}$ with respect to the parameter can be computed by induction, by direct differentiation of the recurrent equation (1) that defines the system. Informally, by the chain rule,[3]

$$\frac{\partial \mathcal{L}_{\rightsquigarrow t}}{\partial \theta} = \frac{\partial \mathcal{L}_t}{\partial s_t} \cdot \frac{\partial s_t}{\partial \theta} \tag{5}$$

where $\frac{\partial s_t}{\partial \theta}$ is the Jacobian matrix of the state s_t as a function of θ. Then by differentiating the evolution equation (1),

$$\frac{\partial s_t}{\partial \theta} = \frac{\partial \mathbf{T}_t}{\partial s_{t-1}} \cdot \frac{\partial s_{t-1}}{\partial \theta} + \frac{\partial \mathbf{T}_t}{\partial \theta}. \tag{6}$$

This allows for computing $\frac{\partial s_t}{\partial \theta}$ by induction in an online manner: store the value of the Jacobian $\frac{\partial s_t}{\partial \theta}$ in a variable J_t, and update J_t via (6) at each time step, namely,

$$J_t = \frac{\partial \mathbf{T}_t}{\partial s_{t-1}} \cdot J_{t-1} + \frac{\partial \mathbf{T}_t}{\partial \theta}$$

after which the stored value J_{t-1} can be discarded. This is the core of the RTRL algorithm. The derivative (5) is then used to obtain the parameter via a gradient descent step

$$\theta \leftarrow \theta - \eta_t \left(\frac{\partial \mathcal{L}_t}{\partial s_t} \cdot J_t \right)$$

with learning rate η_t. In the non-recurrent case (2), $\mathbf{T}_t$ does not depend on s_{t-1}, and RTRL reduces to standard online gradient descent on $\mathcal{L}_t$.

However, updating the parameter at every step breaks the validity of the computations (5)–(6), because RTRL will use values of J_{t-1} stored and computed on previous values of the parameter θ, thus mixing partial derivatives taken at different parameter values. The magnitude of the error at each step is $O(\eta_t)$ (since the parameter changes only by $O(\eta_t)$), so intuitively this should not matter too much for small learning rates. But this is a core difficulty in the analysis of RTRL.

The RTRL algorithm is computationally heavy for large-dimensional systems, since the Jacobian J_t is an element of the space $\mathrm{L}(\Theta, \mathcal{S}_t)$, so that even storing it requires memory $\dim(\theta) \times \dim(s_t)$, not to mention performing the multiplication $\frac{\partial \mathbf{T}_t}{\partial s} J_{t-1}$. This justifies the practical preference for backpropagation through time in non-online setups, and the introduction of approximations such as UORO and NoBackTrack in online setups.

[3] For Jacobians, we use the standard convention from differential geometry, namely, if x and y are multidimensional variables then $\frac{\partial y}{\partial x}$ is the matrix with entries $\frac{\partial y_i}{\partial x_j}$. With this convention the chain rules writes $\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} \frac{\partial y}{\partial x}$. This makes $\frac{\partial \mathcal{L}_t}{\partial s_t}$ a *row* vector. When working with standard RTRL, we abuse notations by omitting the transpose around $\frac{\partial \mathcal{L}_t}{\partial s} \cdot \frac{\partial s_t}{\partial \theta}$ in expressions of the form $\theta \leftarrow \theta - \frac{\partial \mathcal{L}_t}{\partial s} \cdot \frac{\partial s_t}{\partial \theta}$.

2.2 Overview of Results

We provide here a semi-technical account of the main results of the text; the full definitions and statements appear in the next sections. We start with the most general statements covering RTRL, then provide some corollaries for local convergence of various recurrent and non-recurrent existing algorithms: stochastic gradient descent with adaptive preconditioning (RMSProp, Adam, online natural gradient...), truncated backpropagation through time, and RTRL approximations such as UORO and NoBackTrack.

These results take the general form: if the parameter is initialized close enough to some local optimum, then the learning algorithm converges to that optimum. Such a local convergence property is relatively weak, but for most algorithms considered, we could not locate a proof of local convergence.[4] Moreover, we only rely on local assumptions. We do not explicitly assume a random data model. For randomized algorithms, the assumptions are satisfied with probability 1; this results in convergence with probability tending to 1 as the overall learning rate tends to 0 (Section 2.5).

General results: RTRL and extended RTRL algorithms. The general setting is a dynamical system parameterized by $\theta \in \Theta = \mathbb{R}^{\dim(\theta)}$, whose state $s_t \in \mathcal{S}_t = \mathbb{R}^{\dim(s_t)}$ at time $t \geq 1$ is subjected to the evolution equation

$$s_t = \mathbf{T}_t(s_{t-1}, \theta)$$

given some time-dependent, C^2 transition operator $\mathbf{T}_t$. An important example is $\mathbf{T}_t(s_{t-1}, \theta) = F_\theta(s_{t-1}, x_t)$ using a time-independent function F and a sequence of external inputs (x_t): this covers, for instance, recurrent neural networks or general dynamical systems with inputs x_t. A further example is the *non-recurrent* case where s_{t-1} is discarded, namely, $\mathbf{T}_t(s_{t-1}, \theta) = F_\theta(x_t)$, where again $\mathbf{T}_t$ depends on t via x_t. In this latter case one has $s_t = F_\theta(x_t)$: thus, this covers standard parametric interpolation problems, such as feedforward neural networks.

Denote $\mathbf{s}_t(\theta)$ the state obtained at time t by running the system from time 0 to t with parameter θ. (In this overview, we assume s_0 is fixed for simplicity, and omit it.) We assume that we are given a C^2 loss function $\mathcal{L}_t \colon \mathcal{S}_t \to \mathbb{R}$ for each time t. The goal is to minimize the average loss

$$\frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_{\rightsquigarrow t}(\theta), \qquad \mathcal{L}_{\rightsquigarrow t}(\theta) := \mathcal{L}_t(\mathbf{s}_t(\theta))$$

as a function of θ, when $T \to \infty$. A typical loss function would be $\mathcal{L}_t(s_t) = \ell(s_t, y_t)$ where ℓ is a fixed loss function between s_t and a desired output y_t at time t.

In the non-recurrent case with a random i.i.d. sample taken at each time, there is no difference between minimizing the expected loss and minimizing the temporal average of the loss (thanks to the law of large numbers). However, with a dynamical system and with no random data model, we define an optimum based on such

[4] Although this is not treated in this work, we believe that convergence to each local optimum θ^* can be extended to the whole basin of attraction of θ^* for the "ideal" infinitesimal-learning-rate gradient descent $\mathrm{d}\theta_t / \mathrm{d}t = -\partial_\theta \mathcal{L}_{\rightsquigarrow t}(\theta_t)$ using the same proof technique, assuming the learning rates are small enough. Indeed, our whole analysis is based on deviations from this infinitesimal-learning-rate setting, using a suitable Lyapunov function for convergence. We give a more precise argument in Section 2.6.

temporal averages instead of expectations, thus relying on stationarity or ergodicity properties.

Thus, we define a local optimum for this problem as a parameter value $\theta^* \in \Theta$ such that the average derivative of the loss with respect to θ^* vanishes, and such that the average Hessian of the loss at θ^* is positive definite. For a given local optimum, the rate at which these averages converge will affect possible learning rates for convergence towards that optimum. (This is useful for improving learning rates when cycling over a dataset, for instance.) Therefore, more precisely, we say (Assumption 2.11.a) that θ^* is a local optimum with exponent $0 < a < 1$ if gradients of the loss at θ^* average to 0 at rate t^a/t:

$$\frac{1}{T}\sum_{t=1}^{T} \frac{\partial}{\partial \theta} \mathcal{L}_{\rightsquigarrow t}(\theta^*) = O(T^a/T),$$

and if on average, Hessians of the loss at θ^* converge to a positive definite matrix, at rate t^a/t: there is a positive definite matrix H such that

$$\frac{1}{T}\sum_{t=1}^{T} \frac{\partial^2}{\partial \theta^2} \mathcal{L}_{\rightsquigarrow t}(\theta^*) = H + O(T^a/T).$$

For instance, consider a non-recurrent linear regression problem with bounded or Gaussian centered noise ε_t, namely, input x_t, prediction model $s_t = \theta \cdot x_t$, observations $y_t = \theta^* \cdot x_t + \varepsilon_t$, and quadratic losses $\ell_t(s_t) = (s_t - y_t)^2$. Then the derivatives of the loss at θ^* are equal to $-2\varepsilon_t x_t$. So with bounded x_t, the assumption on gradients is satisfied for any $a > 1/2$ by the law of the iterated logarithm (and likewise for deviations from the average Hessian $H = \lim \frac{1}{T}\sum_{t=1}^{T} x_t x_t^\top$ assuming this limit exists). This can be improved if cycling over a finite dataset instead of picking i.i.d. samples: then empirical averages converge at rate $1/T$, so the assumption is satisfied for any $a > 0$ instead of just $a > 1/2$.

We assume (Assumption 2.13) that the dynamical system is stable at first order around θ^*. Remember that a *linear* dynamical system $s_t = As_{t-1} + B\theta + Cx_t$ is stable if and only if A has spectral radius less than 1, namely, if and only if A^k is contracting for some $k \geq 1$. Here the system may be nonlinear. Define $A_t := \frac{\partial \mathbf{T}_t}{\partial s}(\mathbf{s}_{t-1}(\theta^*), \theta^*)$: intuitively this represents the value of $\partial s_t / \partial s_{t-1}$ along the trajectory defined by θ^*. We assume that the product of a sufficiently large number of consecutive A_t is contracting (Assumption 2.13). For a linear system, this is equivalent to standard stability. Note that this is assumed only at θ^*. If this assumption is not satisfied, then even *running* the system with fixed parameter θ^* is numerically unstable, so there is little interest in trying to learn θ^*. In the non-recurrent case, $\mathbf{T}_t$ does not depend on s_{t-1} so that $A_t = 0$ and the assumption is automatically satisfied.

Finally, we have a series of more "technical" assumptions (technical in the sense that they are always satisfied over a finite dataset for a smooth feedforward model): the transition functions $\mathbf{T}_T$ are uniformly C^2 around the trajectory defined by θ^* (Assumption 2.23), the first and second derivatives of the loss function with respect to s_t grow at most like t^γ for some $0 \leq \gamma < 1$ along the trajectory defined by θ^* (Assumption 2.24), and the Hessians of the loss with respect to θ are uniformly continuous in time around θ^* (Assumption 2.25.a, always satisfied if all the functions involved are C^3 with uniformly bounded first, second and third derivatives). For instance, if gradients and Hessians of the loss are bounded over time close to θ^* (e.g., if working with a finite dataset), then $\gamma = 0$.

Our first result is local convergence of the RTRL algorithm under these assumptions: if the parameter is initialized close enough to the local optimum, then RTRL converges to that optimum. The possible range of learning rates depends on the various exponents in the assumptions, allowing for a larger range than the classical Robbins–Monro criterion when cycling over a finite dataset, for instance.

Theorem 2.1 (informal, see Theorem 2.28)**.** *Consider a parameterized dynamical system $s_t = \mathbf{T}_t(s_{t-1}, \theta)$ with loss function $\mathcal{L}_t$ as above, satisfying all the assumptions above. Let θ^* be a local optimum of the empirical loss, in the sense above.*

Let $(\eta_t)_{t\geq 0}$ be a non-increasing stepsize sequence satisfying $\eta_t = \overline{\eta}\, t^{-b}\ (1 + o\,(1/t^\gamma))$, where $\overline{\eta} > 0$ is the overall learning rate*, and b is any exponent such that $\max(a, \gamma) + 2\gamma < b \leq 1$, where a and γ are the exponents from the assumptions above, respectively about convergence of time averages and growth of losses.*

Then there exists a neighborhood $\mathcal{N}_{\theta^*}$ of θ^*, a neighborhood $\mathcal{N}_0^J$ of 0, and an overall learning rate $\overline{\eta}_{\text{conv}} > 0$ such that for any overall learning rate $\overline{\eta} < \overline{\eta}_{\text{conv}}$, the following convergence holds.

For any initial parameter $\theta_0 \in \mathcal{N}_{\theta^*}$ and any initial differential $J_0 \in \mathcal{N}_0^J$, the RTRL learning trajectory

$$\begin{cases} s_t = \mathbf{T}_t\left(s_{t-1},\, \theta_{t-1}\right), \\ J_t = \dfrac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial s}\, J_{t-1} + \dfrac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial \theta}, \\ \theta_t = \theta_{t-1} - \eta_t \left(\dfrac{\partial \mathcal{L}_t(s_t)}{\partial s} \cdot J_t\right) \end{cases}$$

satisfies $\theta_t \to \theta^*$ as $t \to \infty$.

As far as we know, this is the first general convergence result for RTRL.

Extended RTRL algorithms. This theorem generalizes to more complex update rules for θ: these can cover, for instance, adaptive preconditioners such as Adam or online natural gradient (Section 3), by considering the preconditioner as part of the parameter θ to be estimated.

In that case, the passage from θ_{t-1} to θ_t is not necessarily a gradient step for some loss, so we will consider more general update rules. Assume that the update of θ_t Theorem 2.1 is replaced with

$$\theta_t = \Phi(\theta_{t-1}, \eta_t\, v_t), \qquad v_t = \mathcal{U}_t\left(\frac{\partial \mathcal{L}_t(s_t)}{\partial s} \cdot J_t,\ s_t,\ \theta_{t-1}\right)$$

for some operators $\mathcal{U}_t$ and Φ. Namely, $\mathcal{U}_t$ computes an update direction from the RTRL gradients and the current state and parameter, then Φ applies the update with stepsize η_t.

A typical example for $\mathcal{U}_t$ is preconditioning: $\mathcal{U}_t\,(v,\, s,\, \theta) = P(\theta)\, v$ for some matrix-valued P. For preconditioners P estimated online by collecting some statistics, the quantities used to estimate P can be treated as part of θ (see examples in Section 3).

We assume that $\Phi(\theta, v)$ coincides with $\theta - v$ at first order in $\|v\|$ (Assumption 2.17): this covers for instance capped gradient steps such as $\theta - \frac{v}{\max(1, \|v\|)}$, or Riemannian exponentials $\exp_\theta(v)$ expressed in a coordinate system.

We do not make assumptions on the general form of $\mathcal{U}_t$ except for technical assumptions (Assumption 2.14): $\mathcal{U}_t$ is C^1, at most linear with respect to its first

argument, and with bounded derivatives close to the optimal parameter θ^*. This covers preconditioned updates $\mathcal{U}_t(v,\, s,\, \theta) = P(\theta)\, v$ (Remark 2.15).

However, changing the update rule for θ changes the definition of a local optimum: a local optimum becomes a value θ^* such that the average update $\mathcal{U}_t$ is 0. This is a joint property of the dynamical system and the update rule $\mathcal{U}_t$. More precisely (Assumption 2.11.b), we define a "local optimum" for such extended update rules, as a value θ^* such that the average update computed at θ^* tends to 0 at rate T^a/T:

$$\frac{1}{T}\sum_{t=1}^{T}\mathcal{U}_t\left(\frac{\partial}{\partial\theta}\mathcal{L}_{\rightsquigarrow t}(\theta^*), \mathbf{s}_t(\theta^*), \theta^*\right) = O(T^a/T).$$

The second-order condition for a local optimum (positivity of the Hessian) involves the "extended Hessians", defined as the Jacobian of the update direction with respect to θ. Setting

$$\mathcal{H}_t(\theta) := \frac{\partial}{\partial\theta}\left(\theta \mapsto \mathcal{U}_t\left(\frac{\partial}{\partial\theta}\mathcal{L}_{\rightsquigarrow t}(\theta), \mathbf{s}_t(\theta), \theta\right)\right),$$

the assumption states that the average extended Hessian at θ^* converges at rate T^a/T,

$$\frac{1}{T}\sum_{t=1}^{T}\mathcal{H}_t(\theta^*) = \Lambda + O(T^a/T)$$

to some matrix Λ all of whose eigenvalues have positive real part. The standard case is $\mathcal{U}_t(v, s, \theta) = v$: then these conditions reduce to the average gradient being 0 and the average Hessian being positive definite. For a preconditioning $\mathcal{U}_t(v,\, s,\, \theta) = P(\theta)\, v$ with known (non-adaptive) matrix $P(\theta)$, these conditions hold if the average gradient is 0, the average Hessian at θ^* is positive definite, and $P(\theta^*) + P(\theta^*)^\top$ is positive definite (Section 3.1.2).

For adaptive algorithms, we include other quantities as part of the parameter θ to be estimated (such as the average square gradients in Adam). Then the update of θ is not a gradient udpate anymore, and the "extended Hessian" is not a symmetric matrix anymore. Considering the analogous continuous-time dynamical system $\theta' = -\mathcal{U}(\theta)$, it is known that stability of a fixed point θ^* does not require the Jacobian $\partial_\theta\mathcal{U}(\theta^*)$ of the update to be symmetric definite positive, only for its eigenvalues to have positive real part, and this is what we will use.

Under these assumptions on $\mathcal{U}_t$ and Φ, and under the same conditions as in Theorem 2.1, the learning trajectories of the extended RTRL algorithm

$$\begin{cases} s_t = \mathbf{T}_t\left(s_{t-1},\, \theta_{t-1}\right), \\ J_t = \dfrac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial s}\, J_{t-1} + \dfrac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial \theta}, \\ v_t = \mathcal{U}_t\left(\dfrac{\partial \mathcal{L}_t(s_t)}{\partial s}\cdot J_t,\, s_t,\, \theta_{t-1}\right), \\ \theta_t = \Phi(\theta_{t-1}, \eta_t\, v_t), \end{cases}$$

satisfy $\theta_t \to \theta^*$ as $t \to \infty$ (Theorem 2.28).

Corollaries for non-recurrent situations: cycling over samples, adaptive preconditioning... Here we present some consequences of these results for non-recurrent models, in a standard setting for machine learning applications. Namely, we consider a finite dataset $D = (x_n, y_n)_{n\in[1;N]}$ of inputs and labels (with values

in any sets), together with a loss function $\ell(x, y, \theta)$ for an input-label pair (x, y), depending on a parameter θ. We assume ℓ is C^3 with respect to θ.

A *strict local optimum* for this problem is a local optimum of the average loss with positive definite Hessian, namely, a parameter θ^* such that

$$\frac{1}{N}\sum_{n=1}^{N} \partial_\theta \ell(x_n, y_n, \theta^*) = 0, \qquad \frac{1}{N}\sum_{n=1}^{N} \partial_\theta^2 \ell(x_n, y_n, \theta^*) \succ 0.$$

We say that an algorithm to learn θ^* *converges locally* if there is a neighborhood of θ^* and a maximal overall learning rate $\overline{\eta}_{\max}$ (with learning rates as in Theorem 2.1) such that, if the parameter is initialized in this neighborhood and the overall learning rate $\overline{\eta}$ is smaller than $\overline{\eta}_{\max}$, then the sequence of parameters produced by the algorithm converges to θ^*.

First, the "ergodic" viewpoint used to define local optima in the recurrent case illustrates the different behavior of different data sampling strategies for stochastic gradient descent in the non-recurrent case. In pure i.i.d. sampling, at each step a sample from the dataset is selected at random with replacement. In that case, an empirical average of some quantity over T samples converges to the average over the dataset at rate $1/\sqrt{T}$ (variance $1/T$), so that the ergodic assumption above is satisfied with exponent $a > 1/2$. On the other hand, if cycling over all examples in the dataset, or if randomly reshuffling the dataset before each pass on the dataset, then empirical averages over T samples converge to the dataset average at rate $1/T$, so the ergodic assumption is satisfied with any exponent $a > 0$. Cycling or reshuffling acts as a variance reduction method.

Since admissible learning rates in Theorem 2.1 depend on a, this leads to the following.

Corollary 2.2 (informal; see Section 3.1.1)**.** *Consider ordinary stochastic gradient descent*

$$\theta_t = \theta_{t-1} - \eta_t \, \partial_\theta \ell(x_{i_t}, y_{i_t}, \theta)$$

over a finite dataset D with loss ℓ, with i_t the sample selected at step t. Assume the learning rates satisfy $\eta_t \propto t^{-b}$ with

$$\begin{cases} 0 < b \leq 1 & \text{for cycling over } D \text{ or random reshuffling;} \\ 1/2 < b \leq 1 & \text{for i.i.d. sampling of } i_t. \end{cases}$$

Then this algorithm is locally convergent.

Thus, cycling or reshuffling allows for larger learning rates than the classical Robbins–Monro criterion. However, it is unclear if such a variance reduction is desirable from a statistical learning perspective: the variance introduced by i.i.d. resampling is a form of bootstrap and may be helpful to represent the inherent variance from a finite dataset.

Next, adaptive preconditioning can be treated via the "extended" RTRL algorithm using $\mathcal{U}_t$ above. We give several examples in Sections 3.1.2–3.1.5 (RMSProp, Adam with $\beta^2 \to 1$, natural gradient, online natural gradient). In fact, Corollary 3.10 proves local convergence for stochastic gradient descent with momentum and any parameter-dependent adaptive preconditioning matrix P estimated online from the data, provided $P + P^\top$ is positive definite when computed at θ^* and on average over the dataset.

Let us give the example of Adam here. Removing momentum and replacing the entrywise square with a tensor square produces a similar result for the online natural gradient (Section 3.1.4). The extended Kalman filter in the "static" case (for estimating a fixed state via noisy nonlinear measurements) is strictly equivalent to a particular case of online natural gradient via a nontrivial correspondence [Ollivier, 2018], and is covered as a consequence.

Corollary 2.3 (informal, see Corollary 3.10)**.** *Consider a finite dataset $D = (x_n, y_n)$ as above. Take learning rates η_t as in Corollary 2.2 depending on the sample selection scheme.*

Consider a preconditioned gradient descent algorithm with momentum, that maintains a momentum variable J together with square gradient statistics ψ updated via moving averages:

$$\begin{aligned}
J_t &= \beta^1 J_{t-1} + (1-\beta^1)\,\partial_\theta \ell(x_{i_t}, y_{i_t}, \theta_{t-1}) \\
\psi_t &= \beta_t^2\,\psi_{t-1} + (1-\beta_t^2)\,(\partial_\theta \ell(x_{i_t}, y_{i_t}, \theta_{t-1}))^{\odot 2} \\
P_t &= \operatorname{diag}(\psi_t + \varepsilon)^{-1} \\
\theta_t &= \theta_{t-1} - \eta_t P_t J_t \qquad \text{or} \qquad \theta_t = \theta_{t-1} - \eta_t P_{t-1} J_t
\end{aligned}$$

where i_t is the data sampled at step t, where $0 \leq \beta^1 < 1$, where $\beta_t^2 = 1 - c\eta_t$ for some $c > 0$, where $\odot 2$ denotes entrywise squaring of a vector, and where $\varepsilon > 0$ is some regularizing constant.

Then this algorithm is locally convergent.

To obtain this result, the square gradient statistics ψ collected to compute the adaptive preconditioner are treated as a part of the parameter to be estimated: namely, the general convergence result is applied to $\theta^+ := (\theta, \psi)$. At each step, ψ is updated by incorporating a value observed on the current sample. The update of θ^+ is not a gradient step of a loss function, hence the interest of considering the generalized update operators $\mathcal{U}_t$ and the non-symmetric generalized Hessians. Momentum is incorporated by treating it as part of the state s_t of the dynamical system; then the momentum variable J_t coincides with the RTRL Jacobian J_t (Section 3.1.3).

Recurrent models: backpropagation through time, RTRL approximations. RTRL cannot be used directly with large-dimensional recurrent systems due to the impossibility to store J_t, whose size is $(\dim s_t) \times (\dim \theta)$. For such systems, *backpropagation through time* on time intervals $[T_k; T_{k+1}]$ allows for gradients to be computed efficiently on each such interval [Jaeger, 2002; Pearlmutter, 1995]. Alternatively, low-dimensional approximations of RTRL have been introduced, such as NoBackTrack, UORO, or Kronecker-factored RTRL [Ollivier et al., 2015; Tallec and Ollivier, 2018; Mujika et al., 2018]. We now describe results for these situations.

Truncated backpropagation through time using time intervals $[T_k; T_{k+1}]$ of fixed length $T_{k+1} - T_k$ produces a biased algorithm: dynamical effects exceeding the length of these intervals are ignored (see, e.g., the simple "influence balancing" example of divergence in Tallec and Ollivier [2018]). Thus we let the truncation length $L(T)$ grow to ∞ at a slow rate t^A for some exponent $A < 1$. There is a sweet spot for A, related to the learning rates. If $L(T)$ is too small, gradients are biased. If $L(T)$ is too large, then the gradients computed on the time interval $[T; T + L(T)]$ will be large, and the gradient step on θ at the end of each interval will be too large for

convergence. This is described by the relationship between the various exponents in the following result; remember that $\gamma = 0$ if gradients and Hessians of losses are bounded over time close to θ^*, and that a encodes the speed at which empirical averages along the trajectory converge to their limit over time.

Theorem 2.4 (informal, see Definition 3.13 and Theorem 3.14)**.** *Consider a parameterized dynamical system $s_t = \mathbf{T}_t(s_{t-1}, \theta)$ with loss function $\mathcal{L}_t$ as above, satisfying all the assumptions above. Let θ^* be a local optimum of the empirical loss, in the sense above.*

Let $(\eta_t)_{t\geq 0}$ be a non-increasing stepsize sequence satisfying $\eta_t = \overline{\eta}\, t^{-b}\ (1 + o\,(1/t^\gamma))$ where $\overline{\eta} > 0$ is the overall learning rate and b is any exponent such that $\max(a, \gamma) + 2\gamma < b \leq 1$, with a and γ the exponents in the technical assumptions above.

Consider the truncated backpropation through time algorithm using a sequence of time intervals $[T_k; T_{k+1}]$: the system is run with a constant parameter during each such interval, and at time T_{k+1} the cumulated gradient of all losses on $[T_k; T_{k+1})$ is computed via backpropagation through time, and the parameter θ is updated by a gradient step with stepsize $\eta_{T_{k+1}}$ (Definition 3.13).

Assume $T_{k+1} - T_k$ grows like ${T_k}^A$ for some $\max(a, \gamma) < A < b - 2\gamma$.

Then truncated backpropagation through time on the intervals $[T_k; T_{k+1}]$ converges locally to θ^*.

As far as we know, this type of result for truncated backpropagation through time is new.

Finally, let us turn to RTRL approximations such as UORO and NoBackTrack. In such "imperfect" RTRL algorithms (Definition 2.10), instead of maintaining the Jacobian J_t, a smaller-dimensional approximation $\tilde{J}_t$ is used. The computation of $\tilde{J}_t$ follows the RTRL equation, but an additional error E_t is incurred at each time step:

$$\tilde{J}_t = \frac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial s}\, \tilde{J}_{t-1} + \frac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial \theta} + E_t.$$

In NoBackTrack and UORO,[5] the approximation J_t is built in a random way so that the expectation of E_t is 0 at every step. Since the equation on J_t is affine, all subsequent gradients are unbiased, which allows us to prove convergence.

We give a formal mathematical description of NoBackTrack and UORO in Section 3.3.1. Convergence is proved by a single result for imperfect RTRL algorithms (Theorem 2.28) via general properties of the error E_t. Namely, convergence holds as soon as the expectation of E_t knowning previous errors $(E_s)_{s\leq t}$ is 0 at every step (Assumption 2.18), and that the size of the error E_t is sublinear in $\tilde{J}_t$ at every step (Assumption 2.21). In NoBackTrack and UORO, the latter property is ensured by the "variance-reduction" factors originally introduced in the algorithm [Ollivier et al., 2015; Tallec and Ollivier, 2018] (see Section 3.3.1 for details): they play a major role for convergence by ensuring that the error E_t scales at most like $\sqrt{\left\|\tilde{J}_t\right\|}$ at each step.

For this situation we obtain a convergence result similar to RTRL, but with stricter constraints on the learning rates, and with probability tending to 1 as the overall learning rate tends to 0.

[5] Although not formally covered in this text, we believe our results also hold for the more recently introduced Kronecker-factored RTRL [Mujika et al., 2018], which is derived from UORO. Indeed it is enough to check that the assumptions on E_t hold, in a way similar to Section 7.6.

Theorem 2.5 (informal, see Theorem 2.28)**.** *Consider a parameterized dynamical system $s_t = \mathbf{T}_t(s_{t-1}, \theta)$ with loss function $\mathcal{L}_t$ as above, satisfying all the assumptions above. Let θ^* be a local optimum of the empirical loss, in the sense above.*

Consider an imperfect RTRL algorithm with random errors E_t satisfying the unbiasedness and sublinearity assumptions above (which hold for NoBackTrack and for UORO).

Let $(\eta_t)_{t\geq 0}$ be a non-increasing stepsize sequence satisfying $\eta_t = \overline{\eta}\, t^{-b}\ (1 + o\,(1/t^\gamma))$ where $\overline{\eta} > 0$ is the overall learning rate and b is any exponent such that $\max(a, 1/2+\gamma) + 2\gamma < b \leq 1$, where a and γ are the exponents from the assumptions above, respectively about convergence of time averages and growth of losses.

Then there exists a neighborhood $\mathcal{N}_{\theta^*}$ of θ^* and a neighborhood $\mathcal{N}_0^J$ of 0 such that for any $\varepsilon > 0$, there exists $\overline{\eta}_{\text{conv}} > 0$ such that for any overall learning rate $\overline{\eta} < \overline{\eta}_{\text{conv}}$, with probability greater than $1 - \varepsilon$, the following convergence holds:

For any initial parameter $\theta_0 \in \mathcal{N}_{\theta^*}$ and any initial differential $J_0 \in \mathcal{N}_0^J$, the imperfect RTRL learning trajectory

$$\begin{cases} s_t = \mathbf{T}_t\left(s_{t-1},\, \theta_{t-1}\right), \\ \tilde{J}_t = \dfrac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial s}\, \tilde{J}_{t-1} + \dfrac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial \theta} + E_t, \\ \theta_t = \theta_{t-1} - \eta_t \left(\dfrac{\partial \mathcal{L}_t(s_t)}{\partial s} \cdot \tilde{J}_t \right) \end{cases}$$

satisfies $\theta_t \to \theta^*$ as $t \to \infty$.

This result can also be combined with the extended update operators $\mathcal{U}_t$ and Φ, see Theorem 2.28. As far as we know, this is the first theoretical analysis of NoBackTrack and UORO.

2.3 Formal Definitions: Parameterized Dynamical System, RTRL, Extended RTRL Algorithms

We now turn to fully formal definitions and technical assumptions for the convergence theorem. Alternatively, the reader may go directly to the applications and corollaries presented in Section 3.

Linear algebra notation. We have tried to make the text readable under two alternative conventions for linear algebra, with minimal notational fuss. With the programmer's notation, the parameter and state θ and s are tuples of real numbers. This convention makes no difference between row or column tuples or vectors, so that we write simple stochastic gradient descent as $\theta \leftarrow \theta - \eta\, \partial\mathcal{L}/\partial\theta$, ignoring the fact that the tuple $\partial\mathcal{L}/\partial\theta$ is formally a linear form (row vector), that is mapped to a vector using the canonical quadratic form on $\mathbb{R}^{\dim(\theta)}$. This is the convention most relevant for the applications (Section 3).

For the bulk of the mathematical proof, we treat the state and parameter of the system as elements of some finite-dimensional vector spaces and use standard differential geometry notation. Given vector spaces E and F, we denote $\mathrm{L}(E, F)$ the set of linear maps from E to F. Given a smooth map $f\colon E \to F$ and $x \in E$, the differential $\frac{\partial f}{\partial x}(x)$ of f at x is an element of $\mathrm{L}(E, F)$, which can be represented by the Jacobian matrix $\partial f_i(x)/\partial x_j$ in a basis. We write indifferently $\partial_x f$ or $\frac{\partial f}{\partial x}$, depending on typography.

In particular, derivatives of the loss are linear forms $\partial_\theta \mathcal{L} \in \mathrm{L}(\Theta, \mathbb{R})$, not vectors: this is necessary for consistency of the chain rule. This double convention occasionally leads to a few inconsistencies: notably, the Hessian $\partial^2_\theta \mathcal{L}(\theta) \in \mathrm{L}(\mathrm{L}(\Theta, \mathbb{R}), \mathbb{R})$ is formally a $(0,2)$-tensor (a row vector of row vectors), but we sometimes abuse notation and treat it as a matrix. The same occurs for the Lyapunov matrix B of Sections 6.1 and 7.3.

If the vector spaces E and F are equipped with some norms, we always equip $\mathrm{L}(E,F)$ with the operator norm $\|f\|_{\mathrm{op}} := \sup_{x \neq 0} \|f(x)\| / \|x\|$. We follow this convention for compound spaces: for example, the proofs involve spaces of the type $\mathrm{L}(\mathrm{L}(E,F),G)$, which is equipped with the operator norm coming from the operator norm on $\mathrm{L}(E,F)$ and the norm of G.

For pairs, such as the state-parameter pair (s,θ) appearing in some assumptions below, we use the supremum norm; for instance, $\|(s,\theta)\| := \max(\|s\|, \|\theta\|)$.

Parameterized dynamical systems, RTRL. We now provide the formal definitions for RTRL on a parameterized dynamical system.

Definition 2.6 (Parameterized dynamical system)**.** *We consider a dynamical system parameterized by $\theta \in \Theta$, whose state $s_t \in \mathcal{S}_t$ at time $t \geq 1$ is subjected to the evolution equation*

$$s_t = \mathbf{T}_t(s_{t-1}, \theta),$$

where $\Theta \simeq \mathbb{R}^{\dim(\theta)}$ and $\mathcal{S}_t \simeq \mathbb{R}^{\dim(s_t)}$ are some finite-dimensional Euclidean vector spaces (not necessarily of constant dimension with time t), and, for each $t \geq 1$,

$$\mathbf{T}_t \colon \mathcal{S}_{t-1} \times \Theta \to \mathcal{S}_t$$

is a (time-dependent) C^2 map, the transition operator*.*

Such data will be called a parameterized dynamical system*. A sequence of states $(s_t)_{t \geq 0}$ satisfying the evolution equation will be called a* trajectory *with parameter θ.*

We denote by $\mathbf{s}_t \colon \mathcal{S}_0 \times \Theta \to \mathcal{S}_t$ the function that to $s_0 \in \mathcal{S}_0$ and $\theta \in \Theta$, associates the value s_t at time t of the trajectory starting at s_0 with parameter θ.

As usual in statistical learning, the quality of the parameter is assessed through loss functions.

Definition 2.7 (Loss function)**.** *A* loss function *along a parameterized dynamical system, is a family of functions*

$$\mathcal{L}_t \colon \mathcal{S}_t \to \mathbb{R}$$

for each integer $t \geq 1$. We assume that $\mathcal{L}_t$ is C^2 for all t. Given $t \geq 1$, $\theta \in \Theta$ and $s_0 \in \mathcal{S}_0$ we denote

$$\mathcal{L}_{\rightsquigarrow t}(s_0, \theta) := \mathcal{L}_t(\mathbf{s}_t(s_0, \theta))$$

the loss function at the state obtained at time t from $\theta \in \Theta$ and $s_0 \in \mathcal{S}_0$.

Our smoothness assumptions on the $\mathbf{T}_t$'s and the $\mathcal{L}_t$'s imply that $\mathcal{L}_{\rightsquigarrow t}$ is C^2 for all $t \geq 1$.

Here, we have assumed that the state s_t of the system at time t contains all the information necessary to compute the loss. In some applications, the loss has an additional explicit dependency on θ; this can be dealt with by including the

current parameter as part of the state, namely, working on the augmented state $s_t^+ := (s_t, \theta)$.[6]

The goal of training is to find a parameter θ^* such that the asymptotic average loss

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_{\rightsquigarrow t}(s_0, \theta^*) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_t(s_t, \theta^*)$$

is as small as possible, where (s_t) is the trajectory defined by θ^* and s_0.[7]

Let us now define the RTRL algorithm presented informally in Section 2.1.

Definition 2.8 (RTRL algorithm)**.** *The RTRL algorithm with step sizes $(\eta_t)_{t\geq 1}$, starting at $s_0 \in \mathcal{S}_0$ and $\theta_0 \in \Theta$, maintains a state $s_t \in \mathcal{S}_t$, a parameter $\theta_t \in \Theta$, and a Jacobian estimate $J_t \in \mathrm{L}(\Theta, \mathcal{S}_t)$, subjected to the evolution equations*

$$\begin{cases} s_t = \mathbf{T}_t\left(s_{t-1},\, \theta_{t-1}\right), \\ J_t = \dfrac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial s}\, J_{t-1} + \dfrac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial \theta}, \qquad J_0 = 0, \\ v_t = \dfrac{\partial \mathcal{L}_t(s_t)}{\partial s} \cdot J_t, \\ \theta_t = \theta_{t-1} - \eta_t\, v_t \end{cases}$$

for $t \geq 1$.

We will also deal with more general algorithms that perform more complicated updates on the parameter: preconditioning, adaptive per-parameter learning rates, additional error terms... These will be obtained by applying transformations $\mathcal{U}_t$ and Φ_t to the gradient directions computed by RTRL; we will specify assumptions on $\mathcal{U}_t$ and Φ_t later (Section 2.4.3).

Definition 2.9 (Extended RTRL algorithm)**.** *An extended RTRL algorithm with step sizes $(\eta_t)_{t\geq 1}$, starting at $s_0 \in \mathcal{S}_0$ and $\theta_0 \in \Theta$, maintains a state $s_t \in \mathcal{S}_t$, a parameter $\theta_t \in \Theta$, and a Jacobian estimate $J_t \in \mathrm{L}(\Theta, \mathcal{S}_t)$, subjected to the evolution equations*

$$\begin{cases} s_t = \mathbf{T}_t\left(s_{t-1},\, \theta_{t-1}\right), \\ J_t = \dfrac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial s}\, J_{t-1} + \dfrac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial \theta}, \qquad J_0 = 0, \\ v_t = \mathcal{U}_t\left(\dfrac{\partial \mathcal{L}_t(s_t)}{\partial s} \cdot J_t,\, s_t,\, \theta_{t-1}\right), \\ \theta_t = \Phi_t(\theta_{t-1}, \eta_t\, v_t) \end{cases}$$

for $t \geq 1$, for some choice of update functions $\mathcal{U}_t$ and Φ_t.

[6] More precisely, use the extended state space $\mathcal{S}_t^+ := \mathcal{S}_t \times \Theta$ together with the extended transition operators $\mathbf{T}_t^+((s_{t-1}, \theta_{t-1}), \theta_t) := (\mathbf{T}_T(s_{t-1}, \theta_t), \theta_t)$, thus, always storing the latest parameter value in the state. Notably, this does not affect the spectral radius of the operators in Definition 2.12, so that the stability assumption 2.13 is satisfied for the extended system if and only if it is satisfied for the basic system.

[7] A priori this may depend on s_0. We can either decide that s_0 is fixed once and for all by the algorithm, or formally let s_0 be part of the parameter to be optimized. But in the end, under our ergodicity assumptions, the state s_0 will be forgotten and the asymptotic average loss will not depend on s_0.

Imperfect RTRL Algorithms. The RTRL algorithm is unreasonably heavy in most situations, because J_t is an object of dimension $\dim(s_t) \times \dim(\theta)$. Several approximation algorithms are in existence, such as NoBackTrack or UORO. They usually store a smaller-dimensional approximation $\tilde{J}_t$ of J_t (such as a small rank approximation). Since computing J_t from J_{t-1} tends to break this smaller-dimensional structure, the approximation has to be performed after every step. Thus, these algorithms introduce an additional error E_t at each step in the computation of J_t:

$$\tilde{J}_t = \frac{\partial \mathbf{T}_t}{\partial s}\left(s_{t-1},\, \theta_{t-1}\right) \cdot \tilde{J}_{t-1} + \frac{\partial \mathbf{T}_t}{\partial \theta}\left(s_{t-1},\, \theta_{t-1}\right) + E_t.$$

In NoBackTrack, UORO, and Kronecker-factored RTRL, these errors are built in a random way to be centered on average.

For now, we just define an imperfect RTRL algorithm to be one that incurs some error E_t on J_t; Assumptions 2.18 and 2.21 below will require this noise to be not too large (sublinear in J_t) and centered on average.

Definition 2.10 (Imperfect RTRL algorithm)**.** *An* imperfect RTRL algorithm *with step sizes $(\eta_t)_{t\geq 1}$, starting at $s_0 \in \mathcal{S}_0$ and $\theta_0 \in \Theta$, is any algorithm that maintains a state $s_t \in \mathcal{S}_t$, a parameter $\theta_t \in \Theta$, and a Jacobian estimate $\tilde{J}_t \in \mathrm{L}\left(\Theta,\, \mathcal{S}_t\right)$, subjected to the evolution equations for $t \geq 1$*

$$\begin{cases} s_t = \mathbf{T}_t\left(s_{t-1},\, \theta_{t-1}\right), \\ \tilde{J}_t = \dfrac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial s}\, \tilde{J}_{t-1} + \dfrac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial \theta} + E_t, \qquad \tilde{J}_0 = 0, \\ v_t = \mathcal{U}_t\left(\dfrac{\partial \mathcal{L}_t(s_t)}{\partial s} \cdot \tilde{J}_t,\, s_t,\, \theta_{t-1}\right), \\ \theta_t = \Phi_t(\theta_{t-1}, \eta_t\, v_t) \end{cases}$$

for some error term $E_t \in \mathrm{L}\left(\Theta,\, \mathcal{S}_t\right)$.

The errors E_t can be seen as noise on the computation of J performed by the RTRL algorithm. They play a somewhat different role from usual SGD gradient noise (which is encoded by the dependency on t in $\mathcal{L}_t$, usually depending on output data y_t at time t): first, E_t is transmitted from one step to the next in the recurrent computation of J_t; second, these errors are introduced by the optimization algorithm while $\mathcal{L}_t$ is part of the specification of the initial problem.

2.4 Assumptions for Local Convergence

We will prove local convergence of RTRL and extended RTRL algorithms under several assumptions (Theorem 2.28). The various other algorithms described in the introduction are obtained as corollaries by a suitable choice of the update operator $\mathcal{U}_t$ and a suitable definition of the system state s_t encompassing the internal state of an algorithm; this is done in Section 3.

We subdivide the assumptions of our local convergence theorem into "non-technical" assumptions (properties of a strict local optimum, stability of the target dynamical system, centered errors E_t for imperfect RTRL), and "technical" assumptions (grouped in Section 2.4.5). The latter are "technical" in the sense that they would always be satisfied on a finite dataset if every function involved is smooth, for the standard parameter update operators $\mathcal{U}_t$ and Φ_t.

Let us start with the non-technical assumptions. To prove convergence of the learning algorithm towards a local optimum θ^*, we need two key assumptions: first, that θ^* is indeed a local optimum of the loss function. Second, that the system with fixed parameter θ^* is *stable* in the classical sense of dynamical systems, namely: if the parameter is fixed to $\theta = \theta^*$, and if the inputs of the system are fixed (here the inputs are implicit in the definition of the transition operators $\mathbf{T}_t$), then the system eventually forgets its initial state.

Moreover, for extended RTRL algorithms, we assume that applying $\mathcal{U}_t$ and Φ_t behaves reasonably like a gradient step. For imperfect RTRL algorithms ($E_t \neq 0$), we will assume that the errors are centered and sublinear in J_t.

So, let $\theta^* \in \Theta$ and $s_0^* \in \mathcal{S}_0$. Let $(s_t^*)_{t\geq 0}$ be the trajectory starting at s_0^* with parameter θ^*. Provided the assumptions below are satisfied, we will refer to θ^* as the *local optimum*, and to the trajectory $s_t^* := \mathbf{s}_t(\theta^*, s_0^*)$ obtained from θ^* as the *target trajectory*.

The assumptions below are all required to hold locally: either at the target trajectory itself, or only in some neighborhood of the target trajectory. Thus, we fix some radii $r_\Theta > 0$ and $r_\mathcal{S} > 0$, and we will require these assumptions to hold in the balls $B_\Theta(\theta^*, r_\Theta)$ in Θ and $B_{\mathcal{S}_t}(s_t^*, r_\mathcal{S})$ in $\mathcal{S}_t$.

2.4.1 Local Optima of the Loss for a Dynamical System

Defining a local optimum notion for a dynamical system is not straightfoward. In stochastic optimization, the global loss $\bar{\mathcal{L}}$ associated to the parameter is the expectation, over some random variable i, of a loss $\mathcal{L}_i$ which depends on i. Often, i is the random choice of a training sample among a set of data, and $\mathcal{L}_i$ is the loss computed on this sample. In this setting, a local extremum is a parameter θ^* such that, the derivative of the loss evaluated at this parameter vanishes on average:

$$\mathbb{E}\left[\frac{\partial \mathcal{L}_i}{\partial \theta}(\theta^*)\right] = 0.$$

For a dynamical system, we will replace the expectation with respect to i by a temporal average, and we define a local extremum as a point where the temporal averages of gradients converge to 0:

$$\frac{1}{T}\sum_{t=0}^{T} \frac{\partial \mathcal{L}_{\rightsquigarrow t}}{\partial \theta}(s_0^*,\, \theta^*) \to 0,$$

as T tends to infinity. Note that the gradient is computed through the whole dynamics, thanks to the use of $\mathcal{L}_{\rightsquigarrow t}$, which encodes the dependency of s_t on θ. Here no probabilistic assumption is made on the inputs or outputs to the system: instead we work under this "ergodic" assumption of time averages. The classical case (Equation (2)) of a non-recurrent situation corresponds to i.i.d. losses $\mathcal{L}_t$, so that the ergodic assumption is satisfied with probability 1 by the law of large numbers.

For the extremum θ^* to be a minimum, and in order to guarantee the convergence of the gradient descent, we also need a second order condition. We will assume that temporal averages of the Hessians, evaluated at the local extremum, end up being positive definite: the smallest eigenvalue of

$$\frac{1}{T}\sum_{t=0}^{T} \frac{\partial^2 \mathcal{L}_{\rightsquigarrow t}}{\partial \theta^2}(s_0^*,\, \theta^*)$$

when $T \to \infty$, should be positive.

Actually, the rate of convergence (with respect to T) of these limits will affect the range of possible learning rates. For instance, in the non-recurrent case, cycling over a finite dataset results in a convergence $O(1/T)$ of gradients to their average, as opposed to the usual statistical rate $O(1/\sqrt{T})$. This will allow for learning rates $\eta_t = t^{-b}$ with any $b > 0$, instead of the classical $b > 1/2$ for i.i.d. samples. Cycling over samples acts as a variance reduction method; the ergodic viewpoint makes these distinctions clear.

This is why we introduce an exponent a in the next assumption, controlling the rate at which the gradients at θ^* tend to their average.

For a simpler exposition, we first state a version of the assumption corresponding to non-extended algorithms, $\mathcal{U}_t(v, s, \theta) = v$. Then the assumption corresponds to θ^* being a strict local minimum in the traditional sense: gradients at θ^* average to 0, and the Hessian of the expected loss at θ^* is definite positive.

Remember that the loss function $\mathcal{L}_{\rightsquigarrow t}$ (Def. 2.7) encodes the loss at time t of a parameter θ when the system is run with that parameter from time 0 to time t, and is C^2 for all t.

Assumption 2.11.a (θ^* is a local optimum of the average loss function)**.** *We assume the existence of a parameter $\theta^* \in \Theta$, an initial state $s_0^* \in \mathcal{S}_0$, and an exponent $0 < a < 1$ such that:*

1. *Gradients of the loss at θ^* average to 0, at rate t^a/t:*

$$\frac{1}{T}\sum_{t=1}^{T} \frac{\partial}{\partial \theta}\mathcal{L}_{\rightsquigarrow t}(s_0^*, \theta^*) = O(T^a/T).$$

2. *On average, Hessians of the loss at θ^* converge to a positive definite matrix, at rate t^a/t: there is a positive definite matrix H such that*

$$\frac{1}{T}\sum_{t=1}^{T} \frac{\partial^2}{\partial \theta^2}\mathcal{L}_{\rightsquigarrow t}(s_0^*, \theta^*) = H + O(T^a/T).$$

Next we express the corresponding assumption for extended algorithms $\mathcal{U}$: Assumption 2.11.b reduces to Assumption 2.11.a when $\mathcal{U}_t(v, s, \theta) = v$. With extended algorithms (non-trivial $\mathcal{U}$), the optimality assumption works out as follows. Note that it becomes a joint property of the dynamical system and the optimization algorithm: this expresses a condition on θ^* to be a fixed point of the algorithm[8].

Assumption 2.11.b (θ^* is a local optimum of the extended algorithm)**.** *We assume the existence of a parameter $\theta^* \in \Theta$, an initial state $s_0^* \in \mathcal{S}_0$, and an exponent $0 < a < 1$ such that:*

1. *Updates of the open-loop algorithm at θ^* average to 0, at rate t^a/t:*

$$\frac{1}{T}\sum_{t=1}^{T} \mathcal{U}_t\left(\frac{\partial}{\partial \theta}\mathcal{L}_{\rightsquigarrow t}(s_0^*, \theta^*), \mathbf{s}_t(s_0^*, \theta^*), \theta^*\right) = O(T^a/T).$$

[8] The examples of $\mathcal{U}_t$ in Section 3 will still converge to the same local optima. But for instance, with one-dimensional data, by letting $\mathcal{U}_t(v, s, \theta)$ interpolate between v and $\text{sign}(v)$, we could interpolate between computing a mean or a median, so that θ^* depends on the algorithm.

2. *On average, Jacobians of the update at θ^* converge to a positive-stable matrix, at rate t^a/t. Namely, denoting*

$$\mathcal{H}_t(\theta) := \frac{\partial}{\partial \theta}\left(\theta \mapsto \mathcal{U}_t\left(\frac{\partial}{\partial \theta}\mathcal{L}_{\rightsquigarrow t}(s_0^*, \theta), \mathbf{s}_t(s_0^*, \theta), \theta\right)\right), \tag{7}$$

we assume there is a matrix $\Lambda \in \mathrm{L}(\Theta, \Theta)$ such that

$$\frac{1}{T}\sum_{t=1}^{T} \mathcal{H}_t(\theta^*) = \Lambda + O(T^a/T)$$

and all eigenvalues of Λ have positive real part.

Dealing with positive-stable matrices, instead of just symmetric definite positive matrices, is crucially needed for adaptive algorithms such as RMSProp and Adam (see Section 3): the associated updates do not correspond to a gradient direction. Technically this does not pose added difficulties. A reminder on positive-stable matrices is included in Appendix A.

2.4.2 Stability of the Target Trajectory

The next non-technical assumption deals with stability: if the target trajectory is numerically unstable as a dynamical system, then it is unlikely that an RTRL-like algorithm could learn it online. (Besides, the interest of learning unstable models is debatable.) Thus, we will assume that the target trajectory defined by the local optimum θ^* is stable.

Linear systems $s_t = A(\theta)s_{t-1} + B(\theta) + C(\theta)x_t$ with inputs x_t are stable if the spectral radius of A is less than 1 [Willems, 1970], namely, if there exists $k \geq 1$ such that $\left\|A^k\right\|_{\mathrm{op}} < 1$. Since we are going to consider time-inhomogeneous, nonlinear systems we need a slightly extended definition.

Definition 2.12 (Spectral radius of a sequence of linear operators)**.** *A sequence of linear operators $(A_t)_{t\geq 0}$ on a normed vector space is said to have* spectral radius less than 1 *if there exists $\alpha > 0$ and an integer $h \geq 1$ (called* horizon*) such that for any t, the product $A_{t+h-1} \ldots ... A_{t+1}A_t$ has operator norm less than $1 - \alpha$.*

For a constant sequence $A_t \equiv A$ on a finite-dimensional space, this is equivalent to A having spectral radius less than 1.

Assumption 2.13 (The system with parameter θ^* is stable around s^*)**.** *Let*

$$A_t := \frac{\partial \mathbf{T}_t}{\partial s}(s_{t-1}^*, \theta^*).$$

Then the sequence $(A_t)_{t\geq 1}$ has spectral radius less than 1*.*

For a linear system $s_t = A(\theta)s_{t-1} + B(\theta)$ this boils down to classical stability for the parameter $\theta = \theta^*$, namely, $A(\theta^*)$ has spectral radius less than 1.

In the non-recurrent case (2), this is always satisfied, since $\frac{\partial \mathbf{T}_t}{\partial s} = 0$.

A sufficient condition for this criterion is that every A_t has operator norm less than $1 - \alpha$. So, for a simple RNN given by (4), a sufficient condition would be that the matrix W has operator norm less than $4(1 - \alpha)$ (because the sigmoid is 1/4-Lipschitz). But this sufficient condition is far from necessary. Stability can also be checked empirically on a learned model by adding small perturbations.

For advanced recurrent models such as LSTMs, this criterion might be too restrictive because it imposes a time horizon k for contractivity, thus requiring the trained model to have finite effective memory, while LSTMs are specifically designed to have arbitrarily long memory. Allowing the learned model to have infinite memory in our framework would require allowing the spectral radius of the sequence to tend to 1 over time, but this is beyond the scope of the present work.

2.4.3 Extended RTRL Algorithms: Assumptions on $\mathcal{U}_t$ and Φ_t

The next assumptions deal with with extended RTRL algorithms, namely, with the kind of update operators $\mathcal{U}_t$ and Φ_t that can be used instead of directly adding the gradient, $\theta \leftarrow \theta - \eta\, \partial_\theta \mathcal{L}$ as in simple SGD.

The standard RTRL algorithm corresponds to $\mathcal{U}_t(v, s, \theta) = v$. The assumption on $\mathcal{U}_t$ states that $\mathcal{U}_t$ is smooth and behaves (sub)linearly with respect to its first argument v. In extended RTRL algorithms, $\mathcal{U}_t$ is applied to $v = \partial_s \mathcal{L}_t \cdot J$ which is a linear form $v \in \mathrm{L}(\Theta, \mathbb{R})$ that encodes the RTRL estimated gradient $\partial_\theta \mathcal{L}_{\rightsquigarrow t}$.

Assumption 2.14 (Extended update rules $\mathcal{U}_t$)**.** *The extended update rules used in the extended RTRL algorithm are C^1 functions $\mathcal{U}_t \colon \mathrm{L}(\Theta, \mathbb{R}) \times \mathcal{S}_t \times \Theta \to \Theta$.*

We assume that, in a neighborhood of the target trajectory (s_t^*, θ^*), the first derivative of $\mathcal{U}_t$ with respect to v is bounded, and its first derivative with respect to $(s,\, \theta)$ is at most linear in v. Namely, we assume that there exists a constant $\kappa_{\mathcal{U}} > 0$ such that, for any $t \geq 1$, for any $v \in \mathrm{L}(\Theta, \mathbb{R})$, $s \in B_{\mathcal{S}_t}(s_t^*, r_{\mathcal{S}})$ and $\theta \in B_\Theta(\theta^*, r_\Theta)$, one has

$$\left\| \frac{\partial \mathcal{U}_t}{\partial v}\,(v,\, s,\, \theta) \right\|_{\mathrm{op}} < \kappa_{\mathcal{U}} \qquad \text{and} \qquad \left\| \frac{\partial \mathcal{U}_t}{\partial\,(s,\, \theta)}\,(v,\, s,\, \theta) \right\|_{\mathrm{op}} \leq \kappa_{\mathcal{U}}\,(1 + \|v\|).$$

Finally, we assume that

$$\mathcal{U}_t\,(0,\, s_t^*,\, \theta^*) = O\,(t^\gamma)$$

when $t \to \infty$, for some exponent $0 \leq \gamma < 1$ (also used in Assumptions 2.24 and 2.26).

Remark 2.15. *This covers notably the case of preconditioned SGD or RTRL, namely, $\mathcal{U}_t\,(v,\, s,\, \theta) = P(\theta)\, v$ for any smooth matrix-valued P. More generally this covers the case where $\mathcal{U}_t$ depends on v in an affine way, namely, $\mathcal{U}_t\,(v,\, s,\, \theta) = P_t(\theta)\, v + Q_t(\theta)$ where P_t and Q_t are bounded and C^1 close to θ^*, uniformly in time.*

Remark 2.16. *$\mathcal{U}_t$ can also be used to encode small error terms in the algorithm. For instance, $\mathcal{U}_t(v, s, \theta) = v + \varepsilon_t$ where ε_t is some time-dependent error term, corresponds to a perturbed RTRL update $\theta_t = \theta_{t-1} - \eta_t\, v_t - \eta_t\, \varepsilon_t$. The size of ε_t is limited by Assumption 2.11.b which requires that $\frac{1}{T} \sum_{t \leq T} \varepsilon_t = O(T^a/T)$.*

Finally, we assume that the update operators $\Phi_t(\theta, v)$ are equal to $\theta - v$ up to a second-order error in $\|v\|$. This covers simple SGD (no second-order term), as well as, for instance, the exponential map $\exp_\theta(-v)$ in a Riemannian manifold, when expressed in coordinates, and clipped updates such as $\theta - \frac{v}{1+\|v\|}$ (since the algorithm applies Φ to $\eta_t\, v$ not v, this amounts to clipping the update $\eta_t\, v$, not the gradient direction v).

Assumption 2.17 (Parameter update operators)**.** *We assume that the parameter update operators $\Phi_t \colon \Theta \times \Theta \to \Theta$ can be written as*

$$\Phi_t(\theta, v) = \theta - v + \|v\|^2\, \Phi_t^{(2)}(\theta, v)$$

where the second-order term $\Phi_t^{(2)}(\theta, v)$ is bounded and Lipschitz with respect to (θ, v) in some ball $B_\Theta(\theta^*, r_\Theta) \times B_\Theta(0, \tilde{r}_{\mathcal{V}})$, for some $\tilde{r}_{\mathcal{V}} > 0$, uniformly in t.

2.4.4 Assumptions on Errors for Imperfect RTRL Algorithms

Imperfect RTRL algorithms such as NoBackTrack, UORO and Kronecker-factored RTRL introduce an additional error E_t in the definition of J_t (Def. 2.10). This error has been built to be centered on average; since the evolution equation for J_t is affine, this property is preserved through time, a key point in the theoretical analysis.

Thus, we will assume that the errors E_t are random, and centered on average, knowing everything that has happened up to time t.

Assumption 2.18 (Unbiased errors E_t for imperfect RTRL)**.** *We assume that the errors E_t are random variables that satisfy, for every $t \geq 1$,*

$$\mathbb{E}\left[E_t \middle| E_1, \dots, E_{t-1}, \mathcal{F}_0\right] = 0,$$

where $\mathcal{F}_0$ is the σ–algebra generated by the initial parameter θ_0, the initial state s_0, the initial Jacobian estimate $\tilde{J}_0$, and all the algorithm operators, namely $(\mathbf{T}_t)_{t\geq 1}$, $(\mathcal{L}_t)_{t\geq 1}$, $(\mathcal{U}_t)_{t\geq 1}$ and $(\Phi_t)_{t\geq 1}$.

Note that $\mathcal{F}_0$ contains all future algorithm *operators* (the mathematical operations defining the transitions of the dynamical system), not the values of the states themselves.

In RTRL approximations such as NoBackTrack or UORO, the noise E_t is not imposed by the problem, but user-chosen to simplify computation of $\tilde{J}$. The assumption states that this noise should be uncorrelated from all other sources of randomness of the problem, past and future, contained in $\mathcal{F}_0$. Notably, the data samples are implicitly contained in $\mathcal{F}_0$ via the algorithm operators $(\mathbf{T}_t)_{t\geq 1}$ and $(\mathcal{L}_t)_{t\geq 1}$ (see Section 2.1), though the states themselves are not. Thus, this assumption precludes using a recurrent noise E_t that would be correlated to future random choices of data samples (this would obviously produce biases). Since E_t is user-chosen in NoBackTrack or UORO, this is not a problem: just build E_t from random numbers independent from other random choices made by the user. This assumption also precludes "adversarial" recurrent settings in which the universe would send future data that are correlated to the user-chosen noise E_t.

Moreover, we assume that the error E_t is almost surely sublinear with respect to J. For this, let us first define *error gauge* functions which capture this sublinearity.

Definition 2.19 (Gauge for the error)**.** *We call* error gauge *a function $\phi\colon \mathbb{R}_+^2 \to \mathbb{R}_+$ such that*

1. *ϕ is bounded on any compact;*
2. *$\phi(x, y)$ is negligible in front of x, when x goes to infinity, uniformly for y in compact sets: for any compact set $\mathcal{K} \subset \mathbb{R}_+$, when $x \to \infty$ we have*

$$\sup_{y\in\mathcal{K}} \phi(x, y) = o(x).$$

Remark 2.20 (Example of error gauge)**.** *For instance, any function $\phi(x, y) = C(1 + x^\beta)(1 + y)$ with $C > 0$ and $\beta < 1$, is an error gauge. With $\beta = 1/2$, this is the error gauge for the NoBackTrack and UORO algorithms (see Section 7.6).*

Assumption 2.21 (Control of the error by the gauge)**.** *We assume that there exists a gauge function ϕ such that the error E_t of the imperfect RTRL algorithm satisfies, for all $t \geq 1$,*

$$\|E_t\|_{\text{op}} \leq \phi\left(\left\|\tilde{J}_{t-1}\right\|_{\text{op}}, \left\|\frac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial(s, \theta)}\right\|_{\text{op}}\right).$$

Different imperfect RTRL algorithms may admit the same error gauge. This way, the bounds developed below will be satisfied for all these algorithms simultaneously.

Finally, the noise E_t on $\tilde{J}$ needs to stay centered after computing the update direction via $\mathcal{U}(\partial_s \mathcal{L} \cdot \tilde{J}, s, \theta)$, so we assume that $\mathcal{U}$ is linear with respect to its first argument.

Assumption 2.22 (Linearity of the extended updates with respect to the first argument for imperfect RTRL algorithms)**.** *For imperfect RTRL algorithms, we assume that the functions $\mathcal{U}_t$ are linear with respect to their first argument. Namely, we assume that for each $t \geq 1$, for each $s \in \mathcal{S}_t$ and $\theta \in \Theta$, there exists a linear operator $P_t(s, \theta) \colon \mathrm{L}(\Theta, \mathbb{R}) \to \Theta$ such that for any $v \in \mathrm{L}(\Theta, \mathbb{R})$ one has*

$$\mathcal{U}_t(v, s, \theta) = P_t(s, \theta) \cdot v$$

in addition to Assumption 2.14.

This covers, notably, preconditioned SGD algorithms such as those in Section 3.

2.4.5 Technical Assumptions

The following three assumptions are "technical" in the sense that they would be automatically satisfied for smooth functions in the non-recurrent case with a finite dataset (because the sup over t would become a max over the dataset). However, they still encode important properties:

- Uniformity of the dynamical system around the target trajectory,
- The output noise should not grow too fast. In the i.i.d. case, this corresponds to a property of moments of the output noise, see Section 3.1.6.
- The (extended) Hessians of the loss should be uniformly continuous over time in some neighborhood of θ^*.

Assumption 2.23 (The transition functions are uniformly smooth around the target trajectory)**.** *We assume that the derivatives of $\mathbf{T}_t$ are uniformly bounded over time around the target trajectory, namely:*

$$\sup_{t \geq 1} \left\| \frac{\partial \mathbf{T}_t}{\partial (s, \theta)} (s_{t-1}^*, \theta^*) \right\|_{\mathrm{op}} < \infty,$$

and that the second derivatives of $\mathbf{T}_t$ are bounded around θ^* and s_t^*:

$$\sup_{t \geq 1} \sup_{\substack{\theta \in B_\Theta(\theta^*, r_\Theta) \\ s \in B_{\mathcal{S}_{t-1}}(s_{t-1}^*, r_\mathcal{S})}} \left\| \frac{\partial^2 \mathbf{T}_t}{\partial (s, \theta)^2} (s, \theta) \right\|_{\mathrm{op}} < \infty.$$

The next assumption deals with the growth derivatives of the loss $\mathcal{L}_t$ along the trajectory. In the simplest, non-recurrent case (2), $\mathcal{L}_t$ encodes the error between the predicted value and the actual observation; therefore, at θ^*, the difference is equal to the output noise of the model. So in that case, the assumption on derivatives of $\mathcal{L}_t$ implicitly encodes an assumption on the law of the output noise of the model. This point is developed in Section 3.1.6.

Assumption 2.24 (Derivatives of the loss functions have controlled growth along the target trajectory)**.** *We assume that the derivatives of $\mathcal{L}_t$ along the target trajectory grow at most in a controlled way over time, namely, that there exists an exponent $0 \leq \gamma < 1$ such that*

$$\left\| \frac{\partial \mathcal{L}_t}{\partial s}(s_t^*) \right\|_{\mathrm{op}} = O(t^\gamma).$$

Moreover we assume that the second derivative of $\mathcal{L}_t$ is controlled around s_t^*:

$$\sup_{s \in B_{\mathcal{S}_t}(s_t^*, r_{\mathcal{S}})} \left\| \frac{\partial^2 \mathcal{L}_t}{\partial s^2}(s) \right\|_{\mathrm{op}} = O\left(t^\gamma\right).$$

The final technical assumption requires the Hessians (or extended Hessians) of the loss to be uniformly continuous in time in some neighborhood of the local optimum. We first state it for the simple algorithm without $\mathcal{U}_t$ or Φ_t.

Assumption 2.25.a (The Hessians of the loss are uniformly continous close to θ^*.)**.** *We assume that Hessians of the loss are continuous close to θ^*, uniformly in t: there exists a continuous function $\rho\colon \mathbb{R}^+ \to \mathbb{R}^+$ with $\rho(0) = 0$ such that for all t, for all $\theta \in B_\Theta(\theta^*, r_\Theta)$,*

$$\left\| \frac{\partial^2}{\partial \theta^2} \mathcal{L}_{\rightsquigarrow t}(s_0^*, \theta) - \frac{\partial^2}{\partial \theta^2} \mathcal{L}_{\rightsquigarrow t}(s_0^*, \theta^*) \right\|_{\mathrm{op}} \leq \rho(\|\theta - \theta^*\|).$$

For extended algorithms with non-trivial $\mathcal{U}_t$, this rewrites as follows using the Jacobians $\mathcal{H}_t$ of the update direction, defined by (7): these play the role of the Hessian of the loss when the update $\mathcal{U}_t$ is not the gradient of a loss. By construction, Assumption 2.25.b reduces to Assumption 2.25.a in the basic case $\mathcal{U}_t(v, s, \theta) = v$.

Assumption 2.25.b (Jacobians of the updates at θ^* are uniformly continous close to θ^*.)**.** *We assume that the Jacobians $\mathcal{H}_t(\theta)$ of the updates at θ^*, defined by (7), are continuous close to θ^*, uniformly in t: there exists a continuous function $\rho\colon \mathbb{R}^+ \to \mathbb{R}^+$ with $\rho(0) = 0$ such that for all t, for all $\theta \in B_\Theta(\theta^*, r_\Theta)$,*

$$\|\mathcal{H}_t(\theta) - \mathcal{H}_t(\theta^*)\|_{\mathrm{op}} \leq \rho(\|\theta - \theta^*\|).$$

This equicontinuity assumption is arguably the most technical. However, we prove in Appendix B that Assumption 2.25.a is automatically satisfied if the transition and loss operators are C^3 with uniformly bounded first, second and third derivatives. For non-trivial $\mathcal{U}_t$, Assumption 2.25.b is satisfied if in addition, $\mathcal{U}_t$ is C^2 with second derivatives controlled in a certain way (satisfied notably when $\mathcal{U}_t(v, s, \theta) = P_t(s, \theta) \cdot v$ with P_t regular enough).

2.5 A Convergence Theorem for Extended RTRL Algorithms

We now introduce constraints on the stepsize sequence $(\eta_t)_{t \geq 1}$. As explained in the introduction, in non-i.i.d. settings, the step sizes of the gradient descent must satisfy time-homogeneity conditions stricter than the classical Robbins–Monro criterion, in order to avoid correlations between the step sizes and the internal state of the dynamical system. Otherwise, this could bias the gradient descent: for instance, having a step size 0 at every odd step will produce bad results if the underlying

recurrent system exhibits period-2 phenomena. In i.i.d. settings, this is not necessary. Besides, in some applications, we will need the step sizes to be constant for a few steps (for instance, truncated BPTT corresponds to a constant learning rate on each truncation interval): for this we introduce a "wiggle room" factor $1 + o(1/t^\gamma)$.

Assumption 2.26 (Stepsize sequence)**.** *We assume the stepsize sequence $\eta_t \geq 0$ is non-increasing and satisfies*

$$\eta_t = \overline{\eta}\, t^{-b}\, (1 + o\,(1/t^\gamma))$$

for some $b > 0$, where $\overline{\eta} \geq 0$ is the overall learning rate*, and where γ is the exponent in Assumptions 2.14 and 2.24.*

For simple or extended RTRL algorithms (Defs. 2.8–2.9), we assume that $\max(a, \gamma) + 2\gamma < b \leq 1$*, where a is the exponent in Assumptions 2.11.a and 2.11.b.*

For imperfect RTRL algorithms (Def. 2.10), we assume that $\max(a, 1/2 + \gamma) + 2\gamma < b \leq 1$.

In the sequel, we will prove convergence provided the overall learning rate $\overline{\eta}$ is small enough. Thus, in the whole text, η_t is implicitly a function of $\overline{\eta}$.

The conditions on the exponent b deserve some comment. The exponent a encodes the speed at which empirical averages of gradients at θ^* converge to 0, and likewise for Hessians. In typical situations, this holds for any $a > 1/2$ (the standard statistical rate for empirical averages). But on a finite dataset, cycling over the samples in the dataset makes it possible to go down to $a = 0$ (see Section 3). Meanwhile, on a finite dataset, $\gamma = 0$; in general, as discussed above, γ encodes a bound on the growth of the gradients at θ^* over time (see also Section 3.1.6). Thus, for simple and extended RTRL algorithms, it is possible for b to range from 0 to 1 in some cases. But for imperfect RTRL algorithms (NoBackTrack, UORO), the conditions impose $b > 1/2$ as in the standard Robbins–Monro criterion: this is due to inherent added stochasticity in imperfect RTRL algorithms.

We now state the local convergence result for RTRL and extended and imperfect RTRL algorithms.

Definition 2.27 (Local convergence)**.** *Given a parameterized dynamical system as above, we say that an algorithm producing a sequence (θ_t)* converges locally *around θ^* if the following holds: There exists an overall learning rate $\overline{\eta}_{\text{conv}} > 0$ such that, if the parameter θ_0 is initialized close enough to θ^* and the initial state s_0 is close enough to s_0^*, then for any overall learning $\overline{\eta} < \overline{\eta}_{\text{conv}}$, the sequence θ_t computed by the algorithm converges to θ^*. For imperfect RTRL algorithms (NoBackTrack, UORO...), which make random choices for E_t, convergence is meant with probability tending to 1 as the overall learning rate tends to 0.*

Local convergence is a relatively weak requirement for an algorithm; still, as far as we know, no such statement was available for any of the algorithms considered here. Local convergence rules out the kind of bad surprise identified for Adam in Reddi et al. [2018].

The assumptions themselves are only local: we consider that one step in a "wrong" zone of the parameter space may be impossible to recover from. With only local assumptions, the maximal learning rate $\overline{\eta}_{\text{conv}}$ will usually depend on the data: in the non-recurrent case, this means that $\overline{\eta}_{\text{conv}}$ may depend on the random choice of input-output pairs (x_t, y_t), namely, on the dataset and SGD choices. This

is different from global convergence under global convexity assumptions. This is unavoidable with only local assumptions: if noise is unbounded, one single random large step could take the algorithm out of the safe zone where the assumptions hold. So with only local assumptions, the quantifiers need to be reversed: given the dataset (or the sequence of operators $\mathbf{T}_t$ and $\mathcal{L}_t$, encoding a sequence of observations (x_t, y_t)), some learning rate will work. If noise is bounded (e.g., if the dataset is finite) there is no such problem.

For imperfect RTRL algorithms, convergence occurs only with probability close to 1 if the learning rate is small enough. Indeed, these algorithms introduce added stochasticity in the gradient computation. The same remark applies to stochasticity coming from the data (which we consider fixed in the whole text): the assumptions will be satisfied with probability 1, but there will be a data-dependent maximal learning rate. This implies convergence with probability tending to 1 as the overall learning rate $\overline{\eta}$ tends to 0.

Indeed, with only local assumptions, depending on the noise, the computed gradients may deviate from the true gradient during an arbitrarily long time; this is only compensated if the learning rate is small enough compared to the deviation produced by the noise, while larger learning rates may bring the trajectory outside of the safe zone on which the assumptions hold. Thus, in general, with stochasticity and without any assumptions outside of a safe zone, convergence with only occur with probability tending to 1 as $\overline{\eta}$ tends to 0. This contrasts with global convexity assumptions.

Theorem 2.28 (Local convergence of RTRL, extended RTRL, and imperfect RTRL algorithms)**.** *Let $(\mathbf{T}_t)$ be a parameterized dynamical sytem (Def. 2.6) with loss functions $\mathcal{L}_t$ (Def. 2.7). Consider an extended or imperfect RTRL algorithm (Defs. 2.8 or 2.9 or 2.10) on this system.*

Let θ^* be a local optimum for this system (Assumption 2.11.b, which reduces to Assumption 2.11.a if $\mathcal{U}_t$ and Φ_t are not used), with initial state s_0^*. Assume that the system with parameter θ^* starting at s_0^* is stable (Assumption 2.13).

For extended RTRL algorithms, assume the update operators $\mathcal{U}_t$ and Φ_t satisfy Assumptions 2.14 and 2.17.

For imperfect RTRL algorithms, assume moreover that the random RTRL errors E_t are unbiased and controlled by some error gauge (Assumptions 2.18 and 2.21), and that the update operators $\mathcal{U}_t$ are linear with respect to the estimated gradient direction (Assumption 2.22).

Assume that the first and second derivatives of $\mathbf{T}_t$ and $\mathcal{L}_t$ are controlled around the target trajectory (Assumptions 2.23 and 2.24). Assume that the extended Hessians or Jacobians of $\mathcal{U}_t$ are uniformly continuous close to θ^* (Assumption 2.25.b).

Let $\boldsymbol{\eta} = (\eta_t)$ be a stepsize sequence satisfying Assumption 2.26, with overall learning rate $\overline{\eta}$.

Then the algorithm converges locally around θ^*; for imperfect RTRL algorithms, this convergence occurs with probability tending to 1 as the overall learning rate tends to 0.

More precisely,

- *For RTRL or an extended RTRL algorithm, there exists a neighborhood $\mathcal{N}_{\theta^*}$ of θ^*, a neighbourhood $\mathcal{N}_{s_0^*}$ of s_0^*, a neighborhood $\mathcal{N}_0^J$ of 0 in $\mathrm{L}(\Theta, \mathcal{S}_0)$, and an overall learning rate $\overline{\eta}_{\mathrm{conv}} > 0$ such that for any overall learning rate $\overline{\eta} < \overline{\eta}_{\mathrm{conv}}$, the following convergence holds:*

- *For an imperfect RTRL algorithm, there exists a neighborhood $\mathcal{N}_{\theta^*}$ of θ^*, a neighbourhood $\mathcal{N}_{s_0^*}$ of s_0^*, a neighborhood $\mathcal{N}_0^J$ of 0 in $\mathrm{L}(\Theta, \mathcal{S}_0)$ such that for any $\varepsilon > 0$, there exists $\overline{\eta}_{\mathrm{conv}} > 0$ such that for any overall learning rate $\overline{\eta} < \overline{\eta}_{\mathrm{conv}}$, with probability greater than $1 - \varepsilon$, the following convergence holds:*

For any initial parameter $\theta_0 \in \mathcal{N}_{\theta^*}$, any initial state $s_0 \in \mathcal{N}_{s_0^*}$ and any initial differential $\tilde{J}_0 \in \mathcal{N}_0^J$, the trajectory given by

$$\begin{cases} s_t = \mathbf{T}_t\left(s_{t-1},\, \theta_{t-1}\right), \\ \tilde{J}_t = \dfrac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial s}\, \tilde{J}_{t-1} + \dfrac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial \theta} + E_t, \\ v_t = \mathcal{U}_t\left(\dfrac{\partial \mathcal{L}_t(s_t)}{\partial s} \cdot \tilde{J}_t,\, s_t,\, \theta_{t-1}\right), \\ \theta_t = \Phi(\theta_{t-1}, \eta_t\, v_t), \end{cases}$$

(with $E_t = 0$ for non-imperfect RTRL algorithms) satisfies $\theta_t \to \theta^*$ as $t \to \infty$.

2.6 Discussion: How Local is Local Convergence?

The convergence in Theorem 2.28 assumes that θ_0 is initialized close enough to the optimal parameter θ^*. We believe this is not a fundamental limitation of the approach, and that similar results can be extended to initializations θ_0 in the whole basin of attraction θ^* under the "idealized" (non-noisy, infinitesimal-learning-rate) dynamics of the underlying ODE.

Indeed, convergence is obtained by proving a contractivity property for some well-chosen distance (Assumption 4.18.2, proven via Lemma 7.11). In practice, a suitable distance function is obtained by expanding the dynamics at second order around θ^*, so that locally the idealized dynamics behaves like the ODE $\theta' = -\Lambda\theta$ for some matrix Λ whose eigenvalues have positive real part (Λ being the Hessian of the loss in the simplest case). This holds in the zone where the dynamics is close enough to its Taylor expansion.

By defining the distance via a suitable Lyapunov function, a similar contractivity property can be obtained over the whole basin of attraction of a stable fixed point θ^* of an *arbitrary* ODE $\theta' = -U(\theta)$, not only in a neighborhood of θ^*. Indeed, consider an ODE

$$\theta' = -U(\theta)$$

and assume that θ^* is a stable fixed point of the ODE. (In our setting, U is the update operator $\mathcal{U}_t$ of the RTRL gradient descent, averaged over t, and the ODE represents the behavior of the system when the learning rates tend to 0, so that noise averages out.)

Let $\mathcal{B}$ be basin of attraction of θ^*: the set of those θ_0 such that the ODE starting at θ_0 converges to θ^*. On $\mathcal{B}$, define the *Lyapunov distance* by

$$d_L(\theta, \theta') := \sqrt{\int_{t=0}^{\infty} \|\theta_t - \theta'_t\|^2}$$

where θ_t and θ'_t are the value at time t of the ODE starting at θ and θ', respectively. This distance is finite on $\mathcal{B}$ because both trajectories converge exponentially to θ^*. This is indeed a distance, because it is the L^2 distance between the trajectories $(\theta_t)_{t\geq 0}$ defined by θ and θ'. In the linear case $U(\theta) = \Lambda\theta$ with positive-stable Λ, this

distance d_L is exactly the Euclidean metric associated to the positive-definite matrix B providing the Lyapunov function in our proof (Lemma 6.3 and Appendix A).

By construction, the Lyapunov distance decreases along the flow. More precisely, if (θ_t) and (θ'_t) are trajectories of the ODE starting at θ_0 and θ'_0, respectively, then $d_L(\theta_t, \theta'_t) \leq d_L(\theta_0, \theta'_0)$. This is because the integral defining $d_L(\theta_t, \theta'_t)^2$ is the same as the integral defining $d_L(\theta_0, \theta'_0)^2$, minus the time segment $[0; t]$. A more precise short-time contractivity is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}_{|t=0} d_L(\theta_t, \theta'_t)^2 = -\left\|\theta_0 - \theta'_0\right\|^2$$

by construction of d_L.

In particular

$$d_L(\theta_t, \theta^*) \leq d_L(\theta_0, \theta^*)$$

On compact subsets of the basin of attraction $\mathcal{B}$, by bounding d_L as a function of $\|\cdot\|$ (i.e., writing $\|\theta - \theta^*\|^2 \geq \mu\, d_L(\theta, \theta^*)^2$ for some constant $\mu > 0$, obtained by compactness), this can be strengthened to strict contractivity $\frac{\mathrm{d}}{\mathrm{d}t}_{|t=0} d_L(\theta_t, \theta^*)^2 \leq -\mu\, d_L(\theta_0, \theta^*)^2$, resulting in strict contractivity $d_L(\theta_t, \theta^*) \leq e^{-\mu t/2} d_L(\theta_0, \theta^*)$.

Then one can use the Lyapunov distance d_L as the distance that gets contracted for Assumption 4.18 below. This way, proving contractivity of the idealized dynamics would not require comparing the dynamics to its Taylor expansion in a small neighborhood of θ^*.

At the same time, such an approach would require all the other assumptions above to hold in the whole basin of attraction $\mathcal{B}$, not only at or in a neighborhood of θ^*. Notably, for Assumption 2.11.a we would have to assume that for *any* parameter θ in $\mathcal{B}$, the average gradients $\frac{1}{T}\sum_{t=1}^{T} \mathcal{U}_t\left(\frac{\partial}{\partial\theta}\mathcal{L}_{\rightsquigarrow t}(s_0^*, \theta^*), \mathbf{s}_t(s_0^*, \theta^*), \theta^*\right)$ converge to an asymptotic dynamics $U(\theta)$ at rate $O(T^a/T)$, uniformly in (compact subsets of) $\mathcal{B}$. Likewise, Assumption 2.13 (that the dynamical system with a fixed parameter θ is stable) would have to hold not only at θ^*, but over all the zone in which we expect to prove convergence.

We believe these would be interesting topics for future research. In any case, the existence of the Lyapunov distance d_L proves that the ODE method is not intrinsically limited to convergence results in a small neighborhood of θ^*.

3 Examples and Applications

The theorem above establishes local convergence of the basic RTRL algorithm. (Local convergence is defined in Def. 2.27.) We now show how other algorithms can be obtained as particular cases of extended RTRL algorithms for suitable systems.

3.1 Non-Recurrent Situations

We first illustrate how these results play out in the ordinary, non-recurrent case.

In this whole section we consider a finite dataset $D = (x_n, y_n)_{n\in[1;N]}$ of inputs and labels (with values in any sets). The "streaming" setting in which infinitely many independent samples are available and pure online SGD is performed, leads to a different analysis (Section 3.1.6).

We consider a loss function $\ell(x, y, \theta)$ depending on θ. We assume that ℓ is C^3 with respect to θ for each pair (x, y) in the dataset (so all "technical" assumptions of Section 2.4.5 are automatically satisfied, notably the equicontinuity Assumption 2.25.b).

We call *strict local optimum* a local optimum of the average loss with positive definite Hessian, namely, a parameter θ^* such that

$$\frac{1}{N}\sum_{n=1}^{N}\partial_\theta \ell(x_n, y_n, \theta^*) = 0$$

and

$$H := \frac{1}{N}\sum_{n=1}^{N}\partial_\theta^2 \ell(x_n, y_n, \theta^*) \succ 0.$$

We consider three variants of SGD depending on how samples are selected at each step. These will lead to different possible learning rates.

Definition 3.1 (I.i.d. sampling, cycling, random reshuffling)**.** *We call respectively* cycling, i.i.d. sampling, and random reshuffling, *a sequence of integers $i_t \in [1; N]$ where for each time t, i_t is chosen by*

$$i_t = \begin{cases} t \mod N & \textit{(cycling over } D\textit{);} \\ \mathrm{Unif}([1; N]) & \textit{(i.i.d. sampling);} \\ \pi_k(t \mod N) & \textit{(random reshuffling)} \end{cases}$$

where $k = \lceil t/N \rceil$ and for each k, π_k is a random permutation of $[1; N]$, and where for convenience purposes, we define $t \mod N$ to take values in $[1; N]$ instead of $[0; N-1]$.

We will abbreviate

$$\ell_t(\theta) := \ell(x_{i_t}, y_{i_t}, \theta).$$

In this setting, the "technical" assumptions of Section 2.4.5 are automatically satisfied, with exponent $\gamma = 0$, because the dataset is finite (so a supremum over t becomes a maximum over D) and the functions involved are smooth. (Contrast with Section 3.1.6 on infinite datasets.) But the non-technical assumptions still have to be checked and lead to interesting phenomena.

3.1.1 Ordinary SGD on a Finite Dataset

Ordinary SGD is cast as follows in our formalism; we consider two cases depending on whether a random sample is taken from the dataset at each step, or whether we cycle through all samples.

Example 3.2 (Non-recurrent case)**.** *We call* non-recurrent case *the following choice of evolution operators $\mathbf{T}_t$ and loss functions $\mathcal{L}_t$: the state is just the parameter θ itself, namely*

$$\mathbf{T}_t(s_{t-1}, \theta) := \theta, \qquad s_t = \theta$$

and the loss is

$$\mathcal{L}_t(s) := \ell(x_{i_t}, y_{i_t}, s)$$

where for each time t, the sample i_t is chosen by i.i.d. sampling, random reshuffling, or cycling over the dataset.

Then the RTRL algorithm for this non-recurrent case corresponds to ordinary stochastic gradient descent.

The choice of i.i.d. sampling for i_t, versus cycling over D or random reshuffling, influences the speed at which $\frac{1}{T}\sum_{t=1}^{T} \partial_\theta \ell_t(\theta^*)$ converges to 0 (Assumption 2.11.a). Indeed, for cycling and random reshuffling, each sample in the dataset is sampled exactly once in every interval $(kN; (k+1)N]$. Therefore, when summing $\partial_\theta \ell_t(\theta^*)$ from time 1 to T, full swipes over the dataset ("epochs") exactly cancel to 0, and the average $\frac{1}{T}\sum_{t=1}^{T} \partial_\theta \ell_t(\theta^*)$ tends to 0 at rate $O(1/T)$.

On the other hand, with i.i.d. sampling, the averages $\frac{1}{T}\sum_{t=1}^{T} \partial_\theta \ell_t(\theta^*)$ converge to 0 almost surely at a rate $O(\sqrt{(\ln \ln T)/T})$ by the law of the iterated logarithm. The same applies to the Hessians. Therefore, we find:

Proposition 3.3. *Assume that θ^* is a strict local optimum of the average loss over the dataset D. Then Assumption 2.11.a is satisfied*

- *for any $a \geq 0$ for the case of cycling over D or random reshuffling;*
- *for any $a > 1/2$ for the i.i.d. case, with probability 1 over the choice of random samples i_t.*

As a consequence, larger learning rates can be used when cycling over the data than when using i.i.d. samples. Indeed, remember that the exponent a directly constrains the set of possible learning rates (Assumption 2.26) in the convergence theorem. Here we have $\gamma = 0$ so Assumption 2.26 is satisfied with rates $\eta_t = t^{-b}$ for any $a < b \leq 1$. Also note that the spectral radius assumption 2.13 is trivially satisfied because $\partial_s \mathbf{T}_t = 0$. Therefore, Theorem 2.28 yields the following:

Corollary 3.4. *Consider ordinary stochastic gradient descent*

$$\theta_t = \theta_{t-1} - \eta_t \, \partial_\theta \ell(x_{i_t}, y_{i_t}, \theta)$$

over a finite dataset D with loss ℓ as above. Assume the learning rates satisfy $\eta_t \propto t^{-b}$ with

$$\begin{cases} 0 < b \leq 1 & \text{for cycling over } D \text{ or random reshuffling;} \\ 1/2 < b \leq 1 & \text{for i.i.d. sampling of } i_t. \end{cases}$$

Then this algorithm is locally convergent.

Thus, cycling over D or random reshuffling allow for larger learning rates than the traditional range of exponents $b \in (1/2; 1]$ for i.i.d. sampling. Cycling acts as a basic form of variance reduction in SGD, ensuring that every data is sampled exactly once within each cycle of N steps. (See discussion in the introduction and related work section.)

The case of a genuine online SGD with infinitely many distinct samples is different and is treated in Section 3.1.6.

Thus, for the rest of Section 3.1 we set learning rates $\eta_t = \overline{\eta}\, t^{-b}$ with $0 < b \leq 1$ for random reshuffling or cycling over D, or $1/2 < b \leq 1$ for i.i.d. sampling.

3.1.2 SGD with Known Preconditioning Matrix

Let us now illustrate the case of gradient descent preconditioned by a matrix $P(\theta)$, of the form

$$\theta \leftarrow \theta - \eta_t P(\theta) \, \partial_\theta \ell_t$$

thus using a non-trivial update operator $\mathcal{U}_t$. This illustrates how Assumption 2.11.b plays out with the extended Hessians $\mathcal{H}_t$.

We first assume that we can compute $P(\theta)$ explicitly given θ. (The case where P is estimated online is treated below, in the section on adaptive algorithms.) This covers, for instance, the natural gradient with $P(\theta)$ the inverse of the Fisher matrix at θ.

(For the case of Riemannian metrics, $\theta \leftarrow \theta - \eta_t P(\theta)\, \partial_\theta \ell_t$ directly applies the update by addition in some coordinate system. For true "manifold" Riemannian gradients with an added exponential map, the exponential map can be put in the update operator Φ_t.)

Example 3.5 (Preconditioned SGD)**.** *Consider again the non-recurrent setting of Example 3.2. Let $\theta \mapsto P(\theta)$ be a C^1 map from Θ to the set of square matrices of size $\dim(\Theta)$. We call* preconditioned SGD *the RTRL algorithm resulting from the update operator $\mathcal{U}_t\,(v,\, s,\, \theta) := P(\theta)v$.*

By Remark 2.15, this choice of $\mathcal{U}_t$ is covered by our assumptions.

Thus, a corollary of our main theorem is the following.

Corollary 3.6 (Convergence of preconditioned SGD)**.** *Assume that θ^* is a strict local optimum of the average loss over the dataset D. Assume moreover that $P(\theta^*) + P(\theta^*)^\top$ is positive definite. Take learning rates as in Corollary 3.4.*

Then Assumption 2.11.b is satisfied. Therefore, preconditioned SGD converges locally.

Moreover, the matrix Λ in Assumption 2.11.b is

$$\Lambda = P(\theta^*)H$$

with $H = \frac{1}{N}\sum_{n=1}^{N} \partial_\theta^2 \ell(x_n, y_n, \theta^*)$ the Hessian of the average loss at θ^*.

Proof. Let us check the first point of the assumption, namely, that the average of the updates is 0. Indeed, we have

$$\mathcal{U}_t\,(\partial_\theta \mathcal{L}_{\rightsquigarrow t}, \mathbf{s}_t, \theta) = P(\theta)\partial_\theta \mathcal{L}_{\rightsquigarrow t} = P(\theta)\partial_\theta \ell_t$$

since $\mathcal{L}_{\rightsquigarrow t} = \ell_t$ in the non-recurrent case. We have to take the average over time at $\theta = \theta^*$. Since $P(\theta^*)$ does not depend on t, we just have to check that the average of $\partial_\theta \ell_t$ at $\theta = \theta^*$ vanishes, which is the case by assumption.

The next point of the assumption deals with the average extended Hessians

$$\mathcal{H}_t(\theta) = \partial_\theta\, \mathcal{U}_t\,(\partial_\theta \mathcal{L}_{\rightsquigarrow t}, \mathbf{s}_t, \theta)$$

and here with $\mathcal{U}_t\,(v,\, s,\, \theta) = P(\theta)v$, and using again that $\mathcal{L}_{\rightsquigarrow t} = \ell_t$, we find

$$\mathcal{H}_t(\theta) = \partial_\theta\,(P(\theta)\partial_\theta \ell_t) = (\partial_\theta P(\theta))\,\partial_\theta \ell_t + P(\theta)\partial_\theta^2 \ell_t$$

which we have to average at $\theta = \theta^*$. Since $\partial_\theta \ell_t$ averages to 0 at θ^*, the first term averages to 0 and we find

$$\frac{1}{T}\sum_{t=1}^{T} \mathcal{H}_t(\theta^*) \to P(\theta^*)H$$

where H is the (ordinary) Hessian at θ^* of the average loss over the dataset. So the matrix Λ is

$$\Lambda = P(\theta^*)H$$

By the assumptions on $P(\theta^*)$ and H, and by one of the criteria for positive-stability (Proposition A.2), this is positive-stable. □

3.1.3 Adding Momentum

SGD with momentum appears naturally as an RTRL algorithm with a suitable recurrent state, as follows.

Corollary 3.7 (SGD with momentum)**.** *Consider a recurrent system with real-valued state s subject to the evolution equation*

$$s_t = \mathbf{T}_t(s_{t-1}, \theta) := \beta s_{t-1} + (1-\beta)\ell(x_{i_t}, y_{i_t}, \theta)$$

for some $0 \leq \beta < 1$, where each sample index i_t is chosen as in Example 3.2. Define the loss functions $\mathcal{L}_t(s_t) := s_t$.

Then RTRL on this recurrent system is equivalent to SGD with momentum:

$$\theta_t = \theta_{t-1} - \eta_t J_t, \qquad J_t = \beta J_{t-1} + (1-\beta)\partial_\theta \ell(x_{i_t}, y_{i_t}, \theta_{t-1}) \tag{8}$$

Moreover, Assumption 2.11.a is satisfied, with the same exponents as in Proposition 3.3. Therefore, SGD with momentum β converges locally.

Proof. First, the evolution operators $\mathbf{T}_t$ are obviously β-contracting on s. Since $\beta < 1$, the stability assumption 2.13 on the system is satisfied.

Second, by Definition 2.8, the variable J_t of RTRL for this system exactly follows the right-hand side of (8). This proves that RTRL for this system is equivalent to SGD with momentum.

Finally, let us check Assumption 2.11.a: we have to compute the time averages of gradients and Hessians of $\mathcal{L}_{\rightsquigarrow t}$ with respect to θ. Let us prove that those behave asymptotically as in the momentum-less case. Indeed, for this system we have (dropping s_0 for simplicity)

$$\mathcal{L}_{\rightsquigarrow t}(\theta) = (1-\beta)\sum_{j \leq t} \beta^{t-j} \ell(x_{i_j}, y_{i_j}, \theta)$$

by induction. Therefore,

$$\begin{aligned}
\sum_{t=1}^{T} \mathcal{L}_{\rightsquigarrow t}(\theta) &= \sum_{j \leq T} \ell(x_{i_j}, y_{i_j}, \theta)(1-\beta)\sum_{t=j}^{T} \beta^{t-j} \\
&= \sum_{j \leq T} \ell(x_{i_j}, y_{i_j}, \theta)(1-\beta^{T-j}) \\
&= \sum_{j \leq T} \ell(x_{i_j}, y_{i_j}, \theta) - \sum_{j \leq T} \beta^{T-j} \ell(x_{i_j}, y_{i_j}, \theta) \\
&= \sum_{j \leq T} \ell(x_{i_j}, y_{i_j}, \theta) + O(1)
\end{aligned}$$

as $\sum_{j \leq T} \beta^{T-j}$ is finite and $\ell(x_{i_j}, y_{i_j}, \theta)$ is bounded (since we deal with a finite dataset). Therefore, the time averages $\frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_{\rightsquigarrow t}(\theta)$ coincide up to $O(1/T)$ with the momentum-less case. The same argument applies to gradients $\partial_\theta \mathcal{L}_{\rightsquigarrow t}$ and Hessians $\partial_\theta^2 \mathcal{L}_{\rightsquigarrow t}$. Therefore, at θ^*, we have $\frac{1}{T}\sum_{t=1}^{T} \partial_\theta \mathcal{L}_{\rightsquigarrow t}(\theta^*) \to 0$ and $\frac{1}{T}\sum_{t=1}^{T} \partial_\theta^2 \mathcal{L}_{\rightsquigarrow t}(\theta^*) \to H$ with the same rates as in the momentum-less case. ∎

3.1.4 Adaptive Algorithms: Collecting Statistics Online

Adaptive algorithms such as RMSProp, Adam, or the online natural gradient, estimate a preconditioning matrix $P(\theta)$ via a moving average along the optimization trajectory. Thus it is not possible to apply Corollary 3.6, since the latter assumes direct access to $P(\theta)$ for each θ.

So let us consider an algorithm that maintains an auxiliary variable ψ, computed by aggregating past values of some statistic $\Psi(x_t, y_t, \theta_{t-1})$ depending on past observations. Namely, consider an algorithm of the type

$$\begin{aligned}\theta_t &= \theta_{t-1} - \eta_t P(\theta_{t-1}, \psi_{t-1})\, \partial_\theta \ell_t \\ \psi_t &= \beta_t \psi_{t-1} + (1-\beta_t)\, \Psi(x_t, y_t, \theta_{t-1})\end{aligned}$$

where we let the "inertia" parameter β_t tend to 1 at the same rate as the learning rates, namely,

$$\beta_t = 1 - c\,\eta_t$$

for some constant $c > 0$. (This choice will be discussed later.)

For example, RMSProp and Adam are based on collecting statistics about square gradients, namely, letting Ψ_t be the vector

$$\Psi_t(\theta) := (\partial_\theta \ell_t)^{\odot 2}$$

and then RMSProp uses the preconditioner

$$P(\theta, \psi) := \operatorname{diag}(\psi + \varepsilon)^{-1}$$

for some regularizing constant ε. Adam uses momentum in addition and is treated in Section 3.1.5 below.

The online natural gradient corresponds to a full matrix-valued ψ with[9]

$$\Psi_t(\theta) := (\partial_\theta \ell_t)^{\otimes 2}, \qquad P(\theta, \psi) := \psi^{-1}.$$

The extended Kalman filter in the "static" case (for estimating the state of a fixed system from noisy nonlinear measurements) has been shown to be equivalent to a particular case of the online natural gradient [Ollivier, 2018]: it is an online natural gradient with stepsize $\eta_t = 1/(t+1)$ and Gaussian noise model. Therefore, the results here apply to the static extended Kalman filter as well.

A key idea to treat such algorithms is to view ψ as part of the parameter to be estimated. Indeed, the update of ψ can be seen as a gradient descent for the loss $\|\psi - \Psi_t\|^2$. However, this idea does not work directly: incorporating $\|\psi - \Psi_t\|^2$ into the loss changes the gradients for θ, because Ψ_t typically depends on θ, so that extraneous gradient terms on θ appear.

Instead, here we will take full advantage of the generalized Hessians $\mathcal{H}_t$ for non-gradient updates, and of the fact that the matrix Λ in Assumption 2.11.b does not need to be positive definite, only to have eigenvalues with positive real part. This works out as follows.

[9] This corresponds to the "Gauss–Newton" or "outer product" version of the natural gradient [Martens, 2014; Ollivier, 2015]. The other version has an expectation over predicted values of y_t instead of the actual data y_t, corresponding to a choice of Ψ_t that depends only on x_t and θ_{t-1}, and can be treated similarly.

Corollary 3.8 (Local convergence of adaptive preconditioning)**.** *Consider a finite dataset $D = (x_n, y_n)$ as above. Take learning rates as in Corollary 3.4.*

Let $(x, y, \theta) \mapsto \Psi(x, y, \theta) \in \mathbb{R}^{\dim(\Psi)}$ be any C^1 map. Let P be any C^1 map sending $(\theta, \psi \in \mathbb{R}^{\dim(\Psi)})$ to a square matrix of size $\dim(\theta)$. Consider the following algorithm: the average of Ψ is estimated online via

$$\psi_t = \beta_t \psi_{t-1} + (1 - \beta_t)\, \Psi(x_t, y_t, \theta_{t-1}) \tag{9}$$

with $\beta_t = 1 - c\,\eta_t$ for some $c > 0$, and the parameter is updated using the preconditioning matrix P computed either from ψ_t or ψ_{t-1},

$$\theta_t = \theta_{t-1} - \eta_t P(\theta_{t-1}, \psi_{t-1})\, \partial_\theta \ell_t \tag{10}$$

or

$$\theta_t = \theta_{t-1} - \eta_t P(\theta_{t-1}, \psi_t)\, \partial_\theta \ell_t. \tag{11}$$

Let θ^* be a strict local optimum for the dataset. Let

$$\psi^* := \frac{1}{N} \sum_{n=1}^{N} \Psi(x_n, y_n, \theta^*)$$

be the average value of the statistic at θ^*. Assume that $P(\theta^*, \psi^*) + P(\theta^*, \psi^*)^\top$ is positive definite.

Then the adaptive algorithms (9)–(10) and (9)–(11) converge locally.

Proof. Define the augmented parameter $\theta^+ := (\theta, \psi) \in \Theta \times \mathbb{R}^{\dim(\Psi)}$. Define a system via

$$s_t = \mathbf{T}_t(s_{t-1}, \theta^+) := \theta^+$$

with the loss ℓ_t on $s_t = \theta^+ = (\theta, \psi)$ given by

$$\ell_t(\theta, \psi) := \ell(x_{i_t}, y_{i_t}, \theta)$$

as in Example 3.2. Thus $\partial_{\theta^+} \mathcal{L}_{\rightsquigarrow t}(\theta^+) = \partial_{(\theta,\psi)} \ell(x_{i_t}, y_{i_t}, \theta) = (\partial_\theta \ell(x_{i_t}, y_{i_t}, \theta), 0)$.

Define the extended RTRL algorithm on $\theta^+ = (\theta, \psi)$ with update

$$\mathcal{U}_t((v_\theta, v_\psi), s, (\theta, \psi)) := \begin{pmatrix} P(\theta, \psi) v_\theta \\ c\psi - c\Psi(x_{i_t}, y_{i_t}, \theta) \end{pmatrix}$$

where the first row describes the update of θ and the second row, the update of ψ.

Then by construction, the extended RTRL update on $\theta^+ = (\theta, \psi)$ coincides with (9)–(10) with $\beta_t = 1 - c\eta_t$. (The case of (11) is treated later.)

This choice of $\mathcal{U}_t$ is affine in v, so by Remark 2.15 it is covered by Assumption 2.14 on $\mathcal{U}_t$.

Let us check Assumption 2.11.b. For this system, the initial state s_0 plays no role, so for simplicity we drop it from the notation. The gradients computed by this algorithm are

$$\begin{aligned} u_t(\theta, \psi) &:= \mathcal{U}_t \left(\partial_{\theta^+} \mathcal{L}_{\rightsquigarrow t}(\theta^+), \mathbf{s}_t(\theta^+), \theta^+ \right) \\ &= \begin{pmatrix} P(\theta, \psi)\, \partial_\theta \ell(x_{i_t}, y_{i_t}, \theta) \\ c\psi - c\Psi(x_{i_t}, y_{i_t}, \theta) \end{pmatrix} \end{aligned}$$

and the extended Hessians are $\mathcal{H}_t = \partial_{(\theta,\psi)} u_t$. We find

$$\mathcal{H}_t = \begin{pmatrix} P\,\partial_\theta^2 \ell_t + (\partial_\theta P)\partial_\theta \ell_t & (\partial_\psi P)\partial_\theta \ell_t \\ -c\,\partial_\theta \Psi_t & c\,\mathrm{Id} \end{pmatrix}$$

where we have abbreviated ℓ_t for $\ell(x_{i_t}, y_{i_t}, \theta)$ and likewise for Ψ_t.

We have to prove that the average of $u_t(\theta^*, \psi^*)$ over time is 0, and that the average of $\mathcal{H}_t(\theta^*, \psi^*)$ over time is a positive-stable matrix Λ.

We have $u_t(\theta^*, \psi^*) = (P(\theta^*, \psi^*)\,\partial_\theta \ell_t,\ c\psi^* - c\Psi_t)$. By definition of θ^*, the average of $\partial_\theta \ell_t$ over time is 0. Since $P(\theta^*, \psi^*)$ does not depend on t, this proves that the first component of $u_t(\theta^*, \psi^*)$ averages to 0. The second component of u_t averages to 0 by definition of ψ^*.

The rates of this convergence are $O(1/t)$ if cycling over the data, or $O(\sqrt{(\log\log t)/t})$ for the i.i.d. case, as in Proposition 3.3.

To compute the matrix Λ, let us average $\mathcal{H}_t(\theta^*, \psi^*)$ over time. The quantities $\partial_\theta P(\theta^*, \psi^*)$ and $\partial_\psi P(\theta^*, \psi^*)$ do not depend on time, and $\partial_\theta \ell_t$ averages to 0 at θ^*, so both terms involving $\partial_\theta \ell_t$ average to 0. By the assumption at the start of Section 3.1, the Hessians $\partial_\theta^2 \ell_t$ at θ^* average to some positive definite average Hessian H. Moreover, the time averages of $\partial_\theta \Psi_t$ also converge to their average over the dataset, which is some matrix C. Therefore we find

$$\Lambda = \begin{pmatrix} P(\theta^*, \psi^*)H & 0 \\ -c\,C & c\,\mathrm{Id} \end{pmatrix}$$

for some matrix C.

Since Λ is block-triangular, its eigenvalues are those of PH and of $c\,\mathrm{Id}$. As in Section 3.1.2, PH is positive-stable, and so is $c\,\mathrm{Id}$, so the matrix Λ is positive-stable.

This proves that Assumption 2.11.b is satisfied, for the variant (10) where θ and ψ are updated simultaneously.

The variant (11) where ψ is updated before θ can be treated via a simple trick: update ψ at odd steps and update θ at even steps. More precisely, for $t \geq 1$ define

$$\mathcal{U}_{2t-1}((v_\theta, v_\psi), s, (\theta, \psi)) := \begin{pmatrix} 0 \\ c\psi - c\Psi(x_{i_t}, y_{i_t}, \theta) \end{pmatrix}$$

at odd times, and

$$\mathcal{U}_{2t}((v_\theta, v_\psi), s, (\theta, \psi)) := \begin{pmatrix} P(\theta, \psi)v_\theta \\ 0 \end{pmatrix}$$

at even times. Redefine the step sizes η_t accordingly, and the losses via $\mathcal{L}_{2t-1} := 0$ and $\mathcal{L}_{2t} := \mathcal{L}(x_{i_t}, y_{i_t}, \theta)$. Then the RTRL algorithm for this choice of $\mathcal{U}_t$ coincides with (9)–(11). Crucially, the time averages of the new $\mathcal{U}_t$ (and thus of $\mathcal{H}_t$) from time 1 to $2t$ at θ^* are exactly half the time averages from 1 to t in the previous case. So u_t still averages to 0, and $\mathcal{H}_t$ averages to $\Lambda/2$ for the same matrix Λ. This deals with the case of the update (11). □

Remark 3.9 (Splitting an update into sub-updates)**.** *The trick above, of updating part of the parameter at even steps and the rest at odd steps, works more generally: it is always possible to split an update $\mathcal{U}_t$ into as many sub-updates as needed, in any order. Indeed, our assumptions only deal with an "open-loop" system using a fixed parameter ($\eta_t = 0$), and such split updates do not change the dynamics of the fixed-parameter system, so the assumptions and temporal averages will be unchanged.*

3.1.5 Adam with Fixed β^1 and $\beta^2 \to 1$

Algorithms like Adam can be treated by a direct combination of the arguments of Section 3.1.3 (for momentum) and Section 3.1.4 (adaptive preconditioning)[10].

Corollary 3.10 (Local convergence of adaptive preconditioning with momentum)**.** *Consider a finite dataset $D = (x_n, y_n)$ as above. Take learning rates as in Corollary 3.4.*

Consider an algorithm that maintains a momentum variable J together with statistics Ψ, updated via

$$J_t = \beta^1 J_{t-1} + (1-\beta^1)\partial_\theta \ell(x_{i_t}, y_{i_t}, \theta_{t-1}) \tag{12}$$

$$\psi_t = \beta_t^2\, \psi_{t-1} + (1-\beta_t^2)\, \Psi(x_t, y_t, \theta_{t-1}) \tag{13}$$

with fixed $0 \le \beta^1 < 1$ and with $\beta_t^2 = 1 - c\eta_t$. Here $\psi\colon (x, y, \theta) \mapsto \Psi(x, y, \theta) \in \mathbb{R}^{\dim(\Psi)}$ is any C^1 map.

Let the algorithm update the parameter θ via

$$\theta_t = \theta_{t-1} - \eta_t P(\theta_t, \psi_{t-1}) J_t \qquad \text{or} \qquad \theta_t = \theta_{t-1} - \eta_t P(\theta_t, \psi_t) J_t$$

where P is any C^1 map sending $(\theta, \psi \in \mathbb{R}^{\dim(\Psi)})$ to a square matrix of size $\dim(\theta)$.

Let θ^* be a strict local optimum for the dataset. Let

$$\psi^* := \frac{1}{N} \sum_{n=1}^{N} \Psi(x_n, y_n, \theta^*)$$

be the average value of the statistic at θ^*. Assume that $P(\theta^*, \psi^*) + P(\theta^*, \psi^*)^\top$ is positive definite.

Then this algorithm converges locally.

As in the previous section, Adam is recovered by letting the statistic Ψ and preconditioner P be

$$\Psi(x, y, \theta) := (\partial_\theta \ell(x, y, \theta))^{\odot 2}, \qquad P(\theta, \psi) := \operatorname{diag}(\psi + \varepsilon)^{-1}$$

for some regularizing constant ε.

The main difference with the counterexample in Reddi et al. [2018] is that we take $\beta^2 \to 1$ while the counterexample uses a fixed β^2. Fundamentally, with a fixed β^2, Adam introduces a non-vanishing correlation between the gradient and the preconditioner applied to this gradient; thus the step size becomes correlated with the gradients, which introduces a bias in the average step. With $\beta^2 \to 1$, the preconditioner is computed from an average over more and more past gradients, so this bias disappears asymptotically.

Proof. As in Section 3.1.4, define the augmented parameter $\theta^+ := (\theta, \psi) \in \Theta \times \mathbb{R}^{\dim(\Psi)}$. As in Section 3.1.3, consider a recurrent system with real-valued state s subject to the evolution equation

$$s_t = \mathbf{T}_t(s_{t-1}, (\theta, \psi)) := \beta^1 s_{t-1} + (1-\beta^1)\ell(x_{i_t}, y_{i_t}, \theta)$$

[10] We ignore the so-called bias correction factors of Adam, which tend to 1 exponentially fast and modify the learning rates in the first few iterations, but do not affect asymptotic convergence.

and to the loss function $\mathcal{L}_t(s_t) := s_t$. By the same computation as in Section 3.1.3, we have (dropping s_0 again for simplicity)

$$\mathcal{L}_{\rightsquigarrow t}(\theta, \psi) = (1-\beta) \sum_{j \leq t} \beta^{t-j} \ell(x_{i_j}, y_{i_j}, \theta)$$

and in particular $\partial_\psi \mathcal{L}_{\rightsquigarrow t}(\theta, \psi) = 0$.

By Definition 2.9, the variable J_t of RTRL for this system exactly follows (12).

Define the same update operator as in Section 3.1.4,

$$\mathcal{U}_t((v_\theta, v_\psi), s, (\theta, \psi)) := \begin{pmatrix} P(\theta, \psi) v_\theta \\ c\psi - c\Psi(x_{i_t}, y_{i_t}, \theta) \end{pmatrix}$$

and let us check Assumption 2.11.b.

The gradients computed by this algorithm are

$$\begin{aligned} u_t(\theta, \psi) &:= \mathcal{U}_t \left(\partial_{\theta^+} \mathcal{L}_{\rightsquigarrow t}(\theta^+), \mathbf{s}_t(\theta^+), \theta^+ \right) \\ &= \begin{pmatrix} P(\theta, \psi)\, \partial_\theta \mathcal{L}_{\rightsquigarrow t}(\theta, \psi) \\ c\psi - c\Psi_t \end{pmatrix} \end{aligned}$$

where we have abbreviated Ψ_t for $\Psi(x_{i_t}, y_{i_t}, \theta)$. The extended Hessians are $\mathcal{H}_t = \partial_{(\theta,\psi)} u_t$. We find

$$\mathcal{H}_t = \begin{pmatrix} P\, \partial_\theta^2 \mathcal{L}_{\rightsquigarrow t} + (\partial_\theta P) \partial_\theta \mathcal{L}_{\rightsquigarrow t} & (\partial_\psi P) \partial_\theta \mathcal{L}_{\rightsquigarrow t} \\ -c\, \partial_\theta \Psi_t & c\, \mathrm{Id} \end{pmatrix}.$$

The difference with Section 3.1.4 is that we get the recurrent loss $\mathcal{L}_{\rightsquigarrow t}$ instead of the instantaneous loss ℓ_t. However, as in the proof of Corollary 3.7, this does not change temporal averages: indeed the system is the same as in Corollary 3.7 and by the same computation we have

$$\sum_{t=1}^{T} \partial_\theta \mathcal{L}_{\rightsquigarrow t} = \sum_{t=1}^{T} \partial_\theta \ell(x_{i_t}, y_{i_t}, \theta) + O(1)$$

and likewise for the Hessians. Consequently, time averages of the gradients and Hessians coincide up to $O(1/T)$ with the momentum-less case, so the time averages are the same as in Section 3.1.4, and we can conclude in the same way. In particular we still have

$$\Lambda = \begin{pmatrix} P(\theta^*, \psi^*) H & 0 \\ -c\, C & c\, \mathrm{Id} \end{pmatrix}$$

for some matrix C. □

The theory of two-timescale algorithms (e.g., Tadic [2004]) can also be used to deal with preconditioned SGD, where one timescale is used to estimate the preconditioner at the current value θ_t, and the parameter is updated using this preconditioner. However, with such two-timescale algorithms, it is necessary to update the main parameter at a slower rate than the auxiliary statistic ψ, so that ψ_t has time to converge to its average value at θ_t before θ_t is updated.

Our treatment here allows the parameter θ_t to be updated as fast as the statistic ψ_t. This results in the off-diagonal block in the Λ matrix of the system; but this block has no influence on its eigenvalues hence no bearing on local convergence.

3.1.6 Non-Recurrent Case, Online Stochastic Setting with Infinite Dataset

Here we assume a pure "online SGD" setting with an infinite dataset (x_t, y_t) obtained from some unknown probability distribution, and where each sample is used for exactly one gradient step. One difference with the standard treatment of SGD on convex functions is that we work only under local assumptions: nothing is assumed outside of some ball around θ^*, so the assumptions have to ensure that the learning trajectory never ventures there. Another difference is that we get more constraints on possible learning rates, depending on which moments of the noise are finite (Proposition 3.11). We believe this may be because the empirical law of large numbers (Assumption 2.11.a) does not capture all properties of a random i.i.d. sequence.

Thus, in this section we assume that $(x_t, y_t)_{t\geq 1}$ is a sequence of i.i.d. samples from some probability distribution over some set of input-output pairs. We consider a loss function $\ell(x, y, \theta)$ depending on θ. We assume that ℓ is C^3 with respect to θ for each pair (x, y) in the dataset (as before, this guarantees that the smoothness and equicontinuity assumptions are satisfied). We will consider ordinary stochastic gradient descent

$$\theta_t = \theta_{t-1} - \eta_t \, \partial_\theta \ell_t$$

where as before we abbreviate $\ell_t(\theta) := \ell(x_t, y_t, \theta)$.

We call *strict local optimum* a local optimum of the average loss with positive definite Hessian, namely, a parameter θ^* such that

$$\mathbb{E}_{(x,y)} \partial_\theta \ell(x, y, \theta^*) = 0$$

and

$$H := \mathbb{E}_{(x,y)} \partial_\theta^2 \ell(x, y, \theta^*) \succ 0.$$

These assumptions have no consequences on what happens far from θ^*, where the loss function could be badly behaved. This impacts the behavior of stochastic gradient descent: to ensure local convergence to θ^*, the consecutive steps should never venture out of some ball around θ^*. This implies that the steps $\eta_t \partial_\theta \ell_t$ should be bounded. This has consequences for the possible learning rates depending on the noise on the gradients $\partial_\theta \ell_t$.

In this situation, the "technical" assumptions from Section 2.4.5 are not automatically satisfied. Let us examine all assumptions in turn.

We define the dynamical system as in Example 3.2, by identifying the state s with θ; namely, $\mathbf{T}_t(s, \theta) := \theta$, and $\mathcal{L}_t(s) := \ell(x_t, y_t, s)$. In particular, $\mathcal{L}_{\rightsquigarrow t}(\theta) = \ell(x_t, y_t, \theta)$.

We consider the simple RTRL algorithm (no $\mathcal{U}_t$ and $E_t = 0$). Assumption 2.13 is satisfied because $\partial_s \mathbf{T}_t = 0$. We have to check Assumptions 2.11.a, 2.23, 2.24 and 2.25.a.

Assumption 2.11.a encodes the speed at which empirical averages of gradients and Hessians converge to their expectation. It is satisfied with an exponent a depending on the moments of the noise, as follows.

Proposition 3.11. *Given a random sample (x, y), assume the random variables $\partial_\theta \ell(x, y, \theta^*)$ and $\partial_\theta^2 \ell(x, y, \theta^*)$ have finite moments of order h for some $2 \leq h < 4$. Then with probability 1 over the choice of samples (x_t, y_t), Assumption 2.11.a is satisfied, with exponent $a = 2/h$.*

In particular, if the gradients and Hessians of the loss at θ^* have moments of order 4, then Assumption 2.11.a is satisfied for any exponent $a > 1/2$.

Due to the statistical fluctuations in $1/\sqrt{t}$, Assumption 2.11.a is never satisfied for $a < 1/2$ unless gradients and Hessians are independent of (x, y).

Proof. By Theorem 3.b in Baum and Katz [1965] applied with $r = 2$, we know that i.i.d. centered variables X_t have a moment of order $2 \leq h < 4$ if and only if for every $\varepsilon > 0$, the series $\sum_{t=1}^{\infty} \Pr\left(\left|\sum_{i=1}^{t} X_i\right| > t^{2/h}\varepsilon\right)$ is finite. By Borel–Cantelli, this implies that $\sum_{i=1}^{t} X_i$ is $o(t^{2/h})$ with probability 1. This is what we need for Assumption 2.11.a with exponent $2/h$. □

Assumption 2.23 is trivially satisfied for our choice of $\mathbf{T}_t$.

The strongest assumption is Assumption 2.25.a: it states that $\partial_\theta^2 \ell(x, y, \theta)$ is equicontinuous in θ, uniformly in (x, y) for θ in some neighborhood of θ^*. This happens, for instance, if there exists a constant k such that the third derivatives of $\ell(x, y, \theta)$ are bounded by k for any (x, y) in a neighborhood of θ^*.

Assumption 2.24 amounts to an almost sure bound on the growth of $\partial_\theta \mathcal{L}_t$ and $\partial_\theta^2 \mathcal{L}_t$. For $\partial_\theta^2 \mathcal{L}_t$ this has to be uniform in a neighborhood of θ^*. If Assumption 2.25.a is satisfied, it is enough to check this for a dense (denumerable) set of values of θ in that neighborhood. By the Markov inequality and the Borell–Cantelli lemma, if a sequence of i.i.d. random variables (X_t) has finite moments of order h, then for any $\gamma > 1/h$, we have $X_t = O(t^\gamma)$ with probability 1. Therefore, under Assumption 2.25.a, Assumption 2.24 is again an assumption on moments of gradients and Hessians.

Putting everything together, we obtain the following.

Corollary 3.12 (SGD under local assumptions)**.** *Assume that $\partial_\theta \ell(x, y, \theta^*)$ and $\partial_\theta^2 \ell(x, y, \theta)$ have moments of order $h \geq 2$. Assume the third derivatives of $\ell(x, y, \theta)$ with respect to θ are bounded in a neighborhood of θ^*, uniformly over (x, y). Then the assumptions of Theorem 2.28 are satisfied for $a > \max(1/2, 2/h)$ and $\gamma > 1/h$.*

Consequently, for learning rates $\eta_t = \overline{\eta}\, t^{-b}$ with $\max(1/2, 2/h) + 2/h < b \leq 1$, the stochastic gradient descent $\theta_t = \theta_{t-1} - \eta_t\, \partial_\theta \ell(x_t, y_t, \theta_{t-1})$ converges locally to θ^*.

For instance, for linear regression $y_t = x_t \cdot \theta + \varepsilon_t$ with noise ε_t and bounded x_t, these constraints encode the moments of ε_t.

For $h \to \infty$, we recover the classical constraint $1/2 < b \leq 1$. However, when not all moments of the gradients and Hessians are finite, the learning rates are more constrained. This is due to working only under local assumptions, and reflects the need for the gradient descent to stay in a finite ball a priori.

We do not know if these bounds are optimal: the constraint $b > \max(1/2, 2/h) + 2/h$ may be spurious and due to analyzing the non-recurrent case from a recurrent viewpoint.

3.2 Truncated Backpropagation Through Time

We now consider the truncated backpropagation through time (TBPTT) algorithm, We refer to Jaeger [2002]; Pearlmutter [1995] for a discussion of TBPTT. We assume the truncation length slowly grows to ∞: with fixed truncation length, the gradient computation is biased and there is no reason for the algorithm to converge to a local minimum (see, e.g., the simple "influence balancing" example of divergence in Tallec and Ollivier [2018]). Thus, we let the truncation length grow to ∞ at a slow rate t^A for some exponent $A < 1$, related to the learning rate.

Truncated backpropagation through time comes in several variants [Williams and Peng, 1990; Williams and Zipser, 1995]: an "overlapping" variant in which, at each step t, backpropagation is run for L backwards step and the parameter is updated using the approximate gradient of $\mathcal{L}_t$ (running time $O(Lt)$: every state is visited once forward and L times backward); and a "non-overlapping" or "epochwise" variant in which the input sequence is split into segments of size L and the parameter is updated every L steps using the gradients of losses $\mathcal{L}_{t-L+1}, \dots, \mathcal{L}_t$ computed on this interval (running time $O(t)$: every state is visited once forward and once backward). As a compromise, partly overlapping settings exist [Williams and Peng, 1990; Williams and Zipser, 1995]. We treat the non-overlapping variant here.

We denote $\mathbf{s}_{t_0:t_1}(s_{t_0}, \theta)$ the state at time t_1 of the dynamical system starting at time t_0 in state s_{t_0}, with constant parameter θ. We denote $\mathcal{L}_{t_0 \rightsquigarrow t_1}(s_{t_0}, \theta) := \mathcal{L}_{t_1}(\mathbf{s}_{t_0:t_1}(s_{t_0}, \theta))$ the resulting loss at time t_1 as a function of θ.

Definition 3.13 (Truncated Backpropagation Through Time algorithm)**.** *The Truncated Backpropagation Through Time algorithm (TBPTT), with step sizes $(\eta_t)_{t \geq 1}$, truncation times T_k (an increasing integer sequence starting at $T_0 = 0$), starting at $s_0 \in \mathcal{S}_0$ and $\theta_0 \in \Theta$, maintains a state $s_t \in \mathcal{S}_t$, and a parameter $\theta_t \in \Theta$, subjected to the following evolution equations. For every $k \geq 0$ and every $T_k \leq t \leq T_{k+1}$, the states are computed using parameter θ_{T_k}, and the parameter is updated using the gradient of the loss on that time interval; more precisely,*

$$\begin{cases} s_t = \mathbf{T}_t(s_{t-1}, \theta_{T_k}), & T_k < t \leq T_{k+1} \\ \theta_{T_{k+1}} = \theta_{T_k} - \eta_{T_{k+1}} \displaystyle\sum_{t=T_k+1}^{T_{k+1}} \frac{\partial \mathcal{L}_{T_k \rightsquigarrow t}(s_{T_k}, \theta_{T_k})}{\partial \theta}. \end{cases}$$

Of course, we could also use update functions $\mathcal{U}_t$ and Φ_t as before.

It is well-known [Pearlmutter, 1995] that on a fixed time interval $(T_k; T_{k+1}]$, backpropagation through time computes the same quantity as RTRL with a fixed parameter ("open-loop" RTRL), namely, both compute the gradient of the total loss over the time interval. Thus, we can see TBPTT as a form of RTRL which resets the Jacobian J to 0 at the start of every time interval, and updates the parameter only once at the end (Proposition 7.22).

Thus, by a slight change of our analysis of RTRL, we obtain the following result, proved in Section 7.5.

We find that there is a "sweet spot" for the growth of truncation length: with short intervals, gradient are biased due to truncation. With too long intervals, the parameter is updated very rarely using a large number of gradients, and the steps may be large and diverge. This is the meaning of the constraint $a < A < b - 2\gamma$ relating truncation length t^A to the stepsize sequence exponents of Assumption 2.26.

Theorem 3.14 (Convergence of TBPTT)**.** *Let $(\mathbf{T}_t)$ be a parameterized dynamical sytem (Def. 2.6) with loss functions $\mathcal{L}_t$ (Def. 2.7).*

Let θ^* be a local optimum for this system (Assumption 2.11.a), with initial state s_0^*. Assume that the system with parameter θ^* starting at s_0^* is stable (Assumption 2.13).

Assume that the first and second derivatives of $\mathbf{T}_t$ and $\mathcal{L}_t$ are controlled around the target trajectory (Assumptions 2.23 and 2.24). Assume that the Hessians of the losses are uniformly continuous close to θ^* (Assumption 2.25.a).

Let $\boldsymbol{\eta} = (\eta_t)$ be a stepsize sequence with overall learning rate $\overline{\eta}$ and exponents b, a, γ satisfying Assumption 2.26.

We assume that the truncation intervals for TBPTT have lengths $T_{k+1} - T_k = {T_k}^A$*, for some* $\max(a, \gamma) < A < b - 2\gamma$.

Then truncated backpropagation through time on the intervals $[T_k; T_{k+1}]$ *converges locally around* θ^*.

Explicitly, there exists $\overline{\eta}_{\text{conv}} > 0$ *such that for any* $\overline{\eta} < \overline{\eta}_{\text{conv}}$*, the following holds: There is a neighborhood* $\mathcal{N}_{\theta^*}$ *of* θ^* *and a neighborhood* $\mathcal{N}_{s_0^*}$ *of* s_0^* *such that, for any initial parameter* $\theta_0 \in \mathcal{N}_{\theta^*}$ *and any initial state* $s_0 \in \mathcal{N}_{s_0^*}$ *the TBPTT trajectory given by, for every* $k \geq 0$

$$\begin{cases} s_t = \mathbf{T}_t(s_{t-1}, \theta_{T_k}), & T_k < t \leq T_{k+1} \\ \theta_{T_{k+1}} = \theta_{T_k} - \eta_{T_{k+1}} \displaystyle\sum_{t=T_k+1}^{T_{k+1}} \frac{\partial \mathcal{L}_{T_k \rightsquigarrow t}(s_{T_k}, \theta_{T_k})}{\partial \theta}, \end{cases}$$

produces a sequence $\theta_{T_k} \to \theta^*$ *as* $k \to \infty$.

3.3 Approximations of RTRL: NoBackTrack and UORO

RTRL is computationally too heavy for large-dimensional systems, hence the need for approximations such as NoBackTrack and UORO (see above).

Here we explain how the NoBackTrack and UORO algorithms fit into the framework for imperfect RTRL algorithms (Def. 2.10). These algorithms maintain a rank-one approximation $\tilde{J}_t$ of J_t at each step. Their key feature is that this approximation is unbiased: namely, Assumption 2.18 is satisfied.

As corollaries, we will obtain local convergence of NoBackTrack and UORO under the same assumptions as RTRL (Corollary 3.23); note however that the learning rates are more constrained if the output noise exponent γ is nonzero.

We believe that an identical argument also covers more recent extensions of UORO, such as Kronecker-factored RTRL [Mujika et al., 2018], although we do not include it explicitly. Indeed, the only properties needed are the assumptions from Section 2.4.4, namely, that the noise E_t with respect to true RTRL is unbiased and almost surely bounded by some sublinear function of J_t.

Our analysis also emphasizes the crucial role of the variance reduction scaling factors used in NoBackTrack and UORO (via norm equalization). These ensure that the error is sublinear in J at each step, namely, Assumption 2.21 is satisfied. In the convergence proof, this assumption ensures that errors do not accumulate over time.

3.3.1 The NoBackTrack and UORO Algorithms

For completeness, we include a formal definition of the NoBackTrack and UORO algorithms. We refer to the original publications Ollivier et al. [2015]; Tallec and Ollivier [2018] for ready-to-use formulas.

In NoBackTrack and UORO, the unbiasedness property is achieved by a "rank-one trick" [Ollivier et al., 2015] involving random signs at every step, which randomly reduces J_t to a rank-one approximation $\tilde{J}_t$ with the correct expectation. A variance reduction step is performed thanks to a careful rescaling (the norm equalizing operator), which guarantees that the approximation error on $\tilde{J}_t$ scales like $\sqrt{\left\|\tilde{J}_t\right\|}$.

Definition 3.15 (Norm equalizing operator)**.** *Given two normed vector spaces $\mathcal{E}_1$ and $\mathcal{E}_2$, we define the operator $\odot\colon \mathcal{E}_1 \times \mathcal{E}_2 \to \mathcal{E}_1 \times \mathcal{E}_2$ by*

$$v_1 \odot v_2 := \begin{cases} \left(\sqrt{\dfrac{\|v_2\|}{\|v_1\|}}\, v_1,\ \sqrt{\dfrac{\|v_1\|}{\|v_2\|}}\, v_2 \right) & \text{if} \quad v_1 \neq 0 \quad \text{and} \quad v_2 \neq 0, \\ (0,\, 0) & \text{otherwise.} \end{cases}$$

Note that $v_1 \odot v_2$ is invariant by multiplying v_1 and dividing v_2 by the same factor $\lambda > 0$.

Definition 3.16 (Random signs)**.** *We consider independent, identically distributed Bernoulli random variables*

$$\varepsilon_i\,(t)\,, \quad t \geq 1, \quad 1 \leq i \leq \dim \mathcal{S}_t,$$

which equal 1 or -1 both with probability $1/2$.

For every $t \geq 1$, we write $\varepsilon(t)$ the vector of the $\varepsilon_i\,(t)$'s, for $1 \leq i \leq \dim \mathcal{S}_t$.

The NoBackTrack and UORO reduction operators $\mathcal{R}_t$ are defined so that, if $\tilde{J}_t = v_t^{\mathcal{S}} \otimes v_t^{\Theta}$ is a rank-one approximation of J_t at time t, then $\mathcal{R}_t$ returns a pair $(v_{t+1}^{\mathcal{S}}, v_{t+1}^{\Theta})$ such that, on average over $\varepsilon(t)$,

$$\mathbb{E}_{\varepsilon(t)} \left[v_{t+1}^{\mathcal{S}} \otimes v_{t+1}^{\Theta} \right] = \frac{\partial \mathbf{T}_{t+1}}{\partial s} \tilde{J}_t + \frac{\partial \mathbf{T}_{t+1}}{\partial \theta}$$

namely, on average, the RTRL equation is satisfied.

Definition 3.17 (NoBackTrack reduction operators)**.** *Let $t \geq 1$. We define the NoBackTrack reduction operator at time t,*

$$\begin{aligned} \mathcal{R}_t :\ \ & \mathcal{S}_{t-1} \times \mathrm{L}(\Theta, \mathbb{R}) \times \mathcal{S}_{t-1} \times \Theta \to \mathcal{S}_t \times \mathrm{L}(\Theta, \mathbb{R}) \\ & v^{\mathcal{S}},\, v^{\Theta},\, s,\, \theta \mapsto \mathcal{R}_t \left(v^{\mathcal{S}},\, v^{\Theta},\, s,\, \theta \right), \end{aligned}$$

by

$$\mathcal{R}_t \left(v^{\mathcal{S}},\, v^{\Theta},\, s,\, \theta \right) := \left(\frac{\partial \mathbf{T}_t}{\partial e}\,(s,\, \theta)\ v^{\mathcal{S}} \right) \odot v^{\Theta} + \sum_{i=1}^{\dim \mathcal{S}_t} \varepsilon_i\,(t) \left(\mathfrak{e}_i \odot \frac{\partial \mathbf{T}_t^i}{\partial \theta}\,(s,\, \theta) \right),$$

where at each step t, the $\mathfrak{e}_i$'s are a (deterministic) orthonormal basis of vectors of the state space $\mathcal{S}_t$ (for brevity we omit the time dependency in the notation $\mathfrak{e}_i$).

Note that this operator is invariant by multiplying $v^{\mathcal{S}}$ and dividing v^{Θ} by the same factor $\lambda > 0$.

Definition 3.18 (UORO reduction operators)**.** *Let $t \geq 1$. We define the UORO reduction operator at time t,*

$$\begin{aligned} \mathcal{R}_t :\ \ & \mathcal{S}_{t-1} \times \mathrm{L}(\Theta, \mathbb{R}) \times \mathcal{S}_{t-1} \times \Theta \to \mathcal{S}_t \times \mathrm{L}(\Theta, \mathbb{R}) \\ & v^{\mathcal{S}},\, v^{\Theta},\, s,\, \theta \mapsto \mathcal{R}_t \left(v^{\mathcal{S}},\, v^{\Theta},\, s,\, \theta \right), \end{aligned}$$

by

$$\begin{aligned} \mathcal{R}_t \left(v^{\mathcal{S}},\, v^{\Theta},\, s,\, \theta \right) := & \left(\frac{\partial \mathbf{T}_t}{\partial e}\,(s,\, \theta)\ v^{\mathcal{S}} \right) \odot v^{\Theta} \\ & + \left(\sum_{i=1}^{\dim \mathcal{S}_t} \varepsilon_i\,(t)\ \mathfrak{e}_i \right) \odot \left(\sum_{i=1}^{\dim \mathcal{S}_t} \varepsilon_i\,(t)\ \frac{\partial \mathbf{T}_t^i}{\partial \theta}\,(s,\, \theta) \right), \end{aligned}$$

where the $\mathfrak{e}_i$'s form an orthonormal basis of vectors of $\mathcal{S}_t$.

The difference between NoBackTrack and UORO lies in the order of the sum and norm equalization in the second part. Notably, UORO only has two norm equalization operations, which leads to substantial algorithmic gains [Tallec and Ollivier, 2018].

These operators lead to the formal definition of the NoBackTrack and UORO algorithms, as follows.

Definition 3.19 (NoBackTrack and UORO operators)**.** *Let $t \geq 1$. Let $\mathrm{Rk}_1(\Theta, \mathcal{S}_t)$ denote the space of rank-one linear operators from Θ to $\mathcal{S}_t$. The NoBackTrack (respectively UORO) operator*

$$\begin{aligned} \mathcal{T}_t : \quad \Theta \times \mathcal{S}_{t-1} \times \mathrm{Rk}_1(\Theta, \mathcal{S}_{t-1}) &\to \mathrm{Rk}_1(\Theta, \mathcal{S}_t) \\ \theta,\, s,\, J &\mapsto \mathcal{T}_t(\theta,\, s,\, J) \end{aligned}$$

is defined as follows.

1. *Write $J = v^{\mathcal{S}} \otimes v^{\Theta}$, for some pair $\left(v^{\mathcal{S}}, v^{\Theta}\right)$ belonging to $\mathcal{S}_{t-1} \times \mathrm{L}(\Theta, \mathbb{R})$ (uniquely defined up to multiplying $v^{\mathcal{S}}$ and dividing v^{Θ} by some factor $\lambda \neq 0$).*
2. *Apply the reduction step, that is, set*

$$\left(w^{\mathcal{S}}, w^{\Theta}\right) = \mathcal{R}_t\left(v^{\mathcal{S}}, v^{\Theta}, s, \theta\right),$$

   with $\mathcal{R}_t$ the NoBackTrack (respectively UORO) reduction operator.
3. *Finally, define $\mathcal{T}_t(\theta, s, J) := w^{\mathcal{S}} \otimes w^{\Theta}$.*

This defines a random operator depending on the choice of the random signs $\varepsilon(t)$. In law, the value of $\mathcal{T}_t(\theta, s, J)$ is independent of the choice of decomposition $J = v^{\mathcal{S}} \otimes v^{\Theta}$: indeed, the operator $\mathcal{R}_t$ is invariant by multiplying $v^{\mathcal{S}}$ and dividing v^{Θ} by some factor $\lambda > 0$, so we just have to compare the decompositions $v^{\mathcal{S}} \otimes v^{\Theta}$ and $(-v^{\mathcal{S}}) \otimes (-v^{\Theta})$. Thanks to the random signs, these two choices lead to the same law for $\mathcal{T}_t(\theta, s, J)$.

Definition 3.20 (NoBackTrack and UORO)**.** *We define NoBackTrack and UORO as imperfect RTRL algorithms (Def. 2.10) by setting*

$$\tilde{J}_t := \mathcal{T}_t\left(\theta_{t-1}, s_{t-1}, \tilde{J}_{t-1}\right)$$

and tautologically defining E_t as the difference with respect to RTRL,

$$E_t := \tilde{J}_t - \frac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial s}\, \tilde{J}_{t-1} - \frac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial \theta}.$$

3.3.2 Convergence of NoBackTrack and UORO

We are now ready to state convergence of NoBackTrack and UORO. Compared to RTRL, the only thing to check is that the assumptions of Section 2.4.4 for imperfect RTRL algorithms are satisfied, namely, unbiasedness and boundedness of E_t. For this, we have to add the assumption that the dimension of states $\mathcal{S}_t$ does not increase to infinity over time.

Assumption 3.21 (Bounded state space dimension.)**.** $\sup_{t \geq 0} \dim \mathcal{S}_t < \infty$.

Lemma 3.22 (NoBackTrack and UORO as imperfect RTRL algorithms)**.** *Under Assumption 3.21 of bounded state space dimension, the errors E_t of NoBackTrack and UORO satisfy Assumptions 2.18 (unbiasedness) and 2.21 (sublinearity in J).*

This lemma is proved in Section 7.6. Then Theorem 2.28 provides the following conclusion.

Corollary 3.23 (Convergence of NoBackTrack and UORO)**.** *Under Assumption 3.21 of bounded state space dimension, and under the additional constraint on stepsizes for imperfect RTRL algorithms (Assumption 2.26), NoBackTrack and UORO converge locally under the same general assumptions as RTRL, with probability tending to* 1 *when the overall learning rate tends to* 0*.*

We refer to Section 2.5 for a discussion of why the convergence only occurs with probability tending to 1 as the learning rate tends to 0. In short, with only local instead of global assumptions, with some positive probability, the noise introduced by NoBackTrack and UORO may bring the trajectory outside of the safe zone where our assumptions apply. This gets less and less likely as the learning rate decreases, because the noise is more averaged out.

4 Abstract Online Training Algorithm for Dynamical Systems

In this section, we define an abstract model of a learning algorithm applied to a dynamical system, and formulate assumptions about its behaviour. Notably, for RTRL, the abstract state $\mathfrak{m}_t$ will encompass both the state s_t of the dynamical system, and the internal variable J_t maintained by RTRL.

4.1 Model

We start by defining the spaces the quantities used by the algorithm live on.

Definition 4.1 (Parameter, maintained quantities and tangent vectors spaces)**.** *Let Θ be some metric space, the* parameter space*. Let $(\mathcal{M}_t)_{t\geq 0}$ be a sequence of metric spaces, which represents the objects maintained in memory by an algorithm at time t. Let $(\mathcal{V}_t)_{t\geq 0}$ be a sequence of normed vector spaces, containing the gradients computed at time t.*

The transition operator of the algorithm which we introduce below symbolises all the updates performed on the objects the algorithm maintains in memory.

Definition 4.2 (Transition operator)**.** *We call* transition operator *a family $(\mathcal{A})_{t\geq 1}$ of functions*

$$\begin{aligned}\mathcal{A}_t\colon \quad \Theta \times \mathcal{M}_{t-1} &\to \mathcal{M}_t\\ \theta, \mathfrak{m} &\mapsto \mathcal{A}_t(\theta, \mathfrak{m}).\end{aligned}$$

Given such a transition operator, we call open-loop system *the family of operators $\mathcal{A}_{T_i:T_f}$ for $T_i < T_f$,*

$$\begin{aligned}\mathcal{A}_{T_i:T_f}\colon \quad \Theta \times \mathcal{M}_{T_i} &\to \mathcal{M}_{T_f}\\ \theta, \mathfrak{m} &\mapsto \mathfrak{m}_{T_f},\end{aligned}$$

where the sequence $(\mathfrak{m}_t)$ is defined inductively by $\mathfrak{m}_{T_i} := \mathfrak{m}$ and, for $T_i + 1 \leq t \leq T_f$,

$$\mathfrak{m}_t = \mathcal{A}_t\left(\theta,\, \mathfrak{m}_{t-1}\right).$$

Informations about how to modify the parameter are computed thanks to the following gradient computation operators.

Definition 4.3 (Gradient computation operators)**.** *We call* gradient computation operators *a sequence* $(\boldsymbol{V}_t)_{t\geq 1}$ *where each* $\boldsymbol{V}_t$ *is a function from* $\Theta \times \mathcal{M}_t$ *to* $\mathcal{V}_t$*.*

We call trajectory the sequence of objects maintained in memory, together with the gradients collected along it.

Definition 4.4 (Trajectories)**.** *Let* $(\theta_t)_{t\geq 0}$ *be a sequence of elements of* Θ*, and let* $\mathfrak{m}_0 \in \mathcal{M}_0$*. We call* trajectory with parameter (θ_t) starting at $\mathfrak{m}_0$*, the sequence* $(\mathfrak{m}_t)$ *defined inductively by*

$$\mathfrak{m}_t = \mathcal{A}_t(\theta_{t-1}, \mathfrak{m}_{t-1})$$

together with the sequence $(v_t)_{t\geq 1}$ *of gradients*

$$v_t = \boldsymbol{V}_t(\theta_{t-1}, \mathfrak{m}_t).$$

If θ *is a single parameter value, we extend this definition to the constant sequence* $\theta_t \equiv \theta$*.*

Later, we will call optimal trajectory *the trajectory associated to the optimal parameter* θ^* *starting at* $\mathfrak{m}_0^*$*, once these are introduced.*

Finally, we call parameter update operator the update rule of the parameter.

Definition 4.5 (Parameter update operators)**.** *We call* parameter update operators *a sequence* $(\Phi_t)_{t\geq 1}$ *where each* Φ_t *is a function from* $\Theta \times \mathcal{V}_t$ *to* Θ*.*

Let $(\Phi_t)_{t\geq 1}$ *be parameter update operators. Given a sequence of gradients* $v_t \in \mathcal{V}_t$ *and two integers* $0 \leq t_1 \leq t_2$*, we denote* $\Phi_{t_1:t_2}(\theta, (v_t))$ *the consecutive application of* $\Phi_t(\cdot, v_t)$ *to* θ *from time* $t_1 + 1$ *to time* t_2*, namely*

$$\Phi_{t_1:t_2}(\theta, (v_t)) := \theta_{t_2},$$

where θ_t *is defined inductively via* $\theta_{t_1} := \theta$ *and* $\theta_t = \Phi_t(\theta_{t-1}, v_t)$ *for* $t_1 < t \leq t_2$*.*

Thanks to the operators we have just defined, we model a gradient descent trajectory by the update equations, for $t \geq 1$,

$$\begin{cases} \mathfrak{m}_t = \mathcal{A}_t\left(\theta_{t-1},\, \mathfrak{m}_{t-1}\right) \\ v_t = \boldsymbol{V}_t\left(\theta_{t-1},\, \mathfrak{m}_t\right) \\ \theta_t = \Phi_t\left(\theta_{t-1},\, \eta_t\, v_t\right), \end{cases}$$

for some initial parameter θ_0, some initial $\mathfrak{m}_0 \in \mathcal{M}_0$ and some sequence of step sizes (η_t). Our formalism also encompasses algorithms updating the parameter only after a batch of consecutive steps, as can be seen in Theorem 4.29. We prove convergence of these procedures in Theorem 4.27 and Theorem 4.29.

The next definition will be used instead of "the average of gradients at $\theta \in \Theta$". It replaces the average with the sum of gradients over a number of steps, with θ kept fixed.

Definition 4.6 (Open-loop update from t_1 to t_2)**.** *Let* $\theta \in \Theta$*, let* $0 \leq t_1 \leq t_2$*, let* $\mathfrak{m}_{t_1} \in \mathcal{M}_{t_1}$*, and let* $(\eta_t)_{t\geq 1}$ *be a sequence of step sizes.*

We call open-loop update *of $\theta \in \Theta$ from time t_1 to time t_2, denoted $\Phi_{t_1:t_2}(\theta, \mathfrak{m}_{t_1}, (\eta_t))$, the value of the parameter obtained by first computing the trajectory with fixed parameter θ starting at $\mathfrak{m}_{t_1}$, then collecting the gradients along this trajectory and applying them to θ. More precisely, define by induction for $t > t_1$*

$$\mathfrak{m}_t = \mathcal{A}_t(\theta, \mathfrak{m}_{t-1}) = \mathcal{A}_{t_1:t}(\theta, \mathfrak{m}_{t_1}), \qquad v_t = \boldsymbol{V}_t(\theta, \mathfrak{m}_t)$$

and then set

$$\Phi_{t_1:t_2}(\theta, \mathfrak{m}_{t_1}, (\eta_t)) := \Phi_{t_1:t_2}(\theta, (\eta_t v_t)).$$

If $\eta_t \equiv \eta$ is a constant sequence, we will abbreviate this to $\Phi_{t_1:t_2}(\theta, \mathfrak{m}_{t_1}, \eta)$.

Definition 4.7 (Stepsize sequence: overall learning rate and stepsize schedule)**.** *The sequence of step sizes $\boldsymbol{\eta} = (\eta_t)_{t\geq 1}$ of the algorithm will be parameterized as*

$$\eta_t := \overline{\eta}\, \rho_t,$$

where $\overline{\eta} \geq 0$ is the overall learning rate *and $(\rho_t)_{t\geq 0}$ is a sequence with values in $[0; 1]$, the* stepsize schedule*.*

In the sequel we will assume that a stepsize schedule is given, and prove convergence provided the overall learning rate is small enough. Thus, in the whole text η_t *is implicitly a function of* $\overline{\eta}$. The assumptions on η_t below are actually assumptions on ρ_t.

The following definition will be useful to keep track of orders of magnitude of the quantities involved.

Definition 4.8 (Scale function)**.** *We call scale function a non-negative function f, defined on the non-negative real axis, which satisfies the following properties.*

1. *$f(t)$ tends to infinity, when t tends to infinity.*
2. *f preserves asymptotic equivalence at infinity: if $x_t \to \infty$ and $y_t \sim x_t$ then $f(y_t) \sim f(x_t)$ as $t \to \infty$.*
3. *f is non-decreasing, and $f(1) \geq 1$.*

Remark 4.9. *For instance, for every $a > 0$, $t \mapsto t^a$ is a scale function.*

4.2 Assumptions on the Model

4.2.1 Assumptions about the Transition Operators

Let $\theta^* \in \Theta$ be a target parameter and let $(\mathfrak{m}_t^*)_{t\geq 0}$ be the corresponding trajectory, initialized at $\mathfrak{m}_0^* \in \mathcal{M}_0$. In order to control the trajectory of the quantities maintained by the algorithm, we assume they are contained in some stable tube, defined below. Under some conditions of contractivity for the transition operator, the tubes reduce to sequence of balls of common radius. However, when the transition operator satisfies weak contractivity assumptions, which is our working assumption in the second part, we must use the more general definition below.

Definition 4.10 (Stable tube)**.** *A* stable tube *around a target trajectory $(\mathfrak{m}_t^*)$ with parameter θ^*, is a ball $B_\Theta^* \subset \Theta$ of positive radius r_Θ^* centered* [11] *at θ^*, together with sets $\mathbb{T}_{\mathcal{M}_t} \subset \mathcal{M}_t$ for each $t \geq 0$, such that:*

[11] We have to assume that B_Θ^* is centered at θ^*, for the proof of Lemma 5.9 (we control $d(\theta_t, \theta^*)$ to show that $\theta_t \in B_\Theta^*$).

1. *Stability: If* $t \geq 1$, $\theta \in B_\Theta^*$ *and* $\mathfrak{m}_{t-1} \in \mathbb{T}_{\mathcal{M}_{t-1}}$, *then* $\mathcal{A}_t(\theta, \mathfrak{m}_{t-1}) \in \mathbb{T}_{\mathcal{M}_t}$;

2. *Upper and lower boundedness: There exist* $0 < r \leq R$ *such that, for all* $t \geq 0$,

$$B_{\mathcal{M}_t}(\mathfrak{m}_t^*, r) \subset \mathbb{T}_{\mathcal{M}_t} \subset B_{\mathcal{M}_t}(\mathfrak{m}_t^*, R);$$

3. *Initial value:* $\mathfrak{m}_0^* \in \mathbb{T}_{\mathcal{M}_0}$.

Assumption 4.11 (Stable tube around the target trajectory)**.** *There exists a stable tube around the target trajectory* $(\mathfrak{m}_t^*)$ *defined by* θ^*.

This assumption concerns the fixed-parameter system we start with, not the system where the parameter is learned via the algorithm. A priori, learning might make the trajectories diverge.

Remark 4.12. *We can always decrease the size of* B_Θ^* *while preserving the stable tube condition. This will be useful when introducing additional assumptions below, which may hold over a smaller set of parameters. For simplicity we will just express these assumptions for* $\theta \in B_\Theta^*$, *implicitly taking a smaller* B_Θ^* *if necessary.*

On the other hand, the stable tube condition is generally not stable by reducing $\mathbb{T}_{\mathcal{M}_t}$, *because* $\mathbb{T}_{\mathcal{M}_t}$ *must contain* $\mathcal{A}_t(B_\Theta^*, \mathbb{T}_{\mathcal{M}_{t-1}})$.

Finally, we assume the following behaviour for the transition operator on the stable tube.

Assumption 4.13 (The state update is Lipschitz w.r.t. the parameter.)**.** *We assume that* $\mathcal{A}_t$ *is Lipschitz over* θ *in* $B_\Theta^* \times \mathbb{T}_{\mathcal{M}_{t-1}}$, *uniformly in* t. *Namely, there exists a constant* $\kappa_{\text{lip}\theta}$ *such that for all* $t \geq 1$, *for all* $\theta, \theta' \in B_\Theta^*$ *and* $\mathfrak{m} \in \mathbb{T}_{\mathcal{M}_{t-1}}$,

$$d\left(\mathcal{A}_t(\theta, \mathfrak{m}),\, \mathcal{A}_t(\theta', \mathfrak{m})\right) \leq \kappa_{\text{lip}\theta}\, d\left(\theta,\, \theta'\right).$$

Assumption 4.14 (Exponential forgetting of initialization for fixed θ)**.** *We assume that there exist* $0 < \alpha \leq 1$ *and a constant* κ_1 *such that, for any parameter* $\theta \in B_\Theta^*$, *for any* $t_0 \geq 0$, *for any states* $\mathfrak{m}_{t_0}, \mathfrak{m}'_{t_0} \in \mathbb{T}_{\mathcal{M}_{t_0}}$, *the trajectories* $(\mathfrak{m}_t)$ *and* $(\mathfrak{m}'_t)$ *with parameter* θ *starting at* $\mathfrak{m}_{t_0}$ *and* $\mathfrak{m}'_{t_0}$ *at time* t_0, *respectively, satisfy*

$$d\left(\mathfrak{m}_t,\, \mathfrak{m}'_t\right) \leq \kappa_1 (1-\alpha)^{t-t_0}\, d\left(\mathfrak{m}_{t_0},\, \mathfrak{m}'_{t_0}\right)$$

for all $t \geq t_0$.

Note that these trajectories stay in $\mathbb{T}_{\mathcal{M}_t}$, so this is a "local" assumption.

4.2.2 Assumptions on Gradients

The following assumption aims at controlling the magnitude of the gradients.

Assumption 4.15 (Locally bounded instantaneous gradients)**.** *We suppose that there exists a function* $m(t)$ *which is either a scale function* $m(t) \ll t$ *or* $m(t) \equiv 1$, *such that, when* $t \to \infty$,

$$\sup_{\theta \in B_\Theta^*} \sup_{\mathfrak{m} \in \mathbb{T}_{\mathcal{M}_t}} \|\boldsymbol{V}_t(\theta, \mathfrak{m})\| = O(m(t)).$$

We denote $B_{\mathcal{V}_t}$ *the ball of* $\mathcal{V}_t$ *with radius* $\sup_{\theta \in B_\Theta^*} \sup_{\mathfrak{m} \in \mathbb{T}_{\mathcal{M}_t}} \|\boldsymbol{V}_t(\theta, \mathfrak{m})\|$.

The gradients are Lipschitz with respect to the parameter, and the quantities in memory.

Assumption 4.16 (Instantaneous gradients are locally Lipschitz.)**.** *We assume that there exist κ_5 and a scale function $m_{\mathrm{H}}(t) \ll t$ (or $m_{\mathrm{H}}(t) \equiv 1$) such that for any parameters $\theta, \theta' \in B^*_\Theta$, for any $t \geq 1$, for any $\mathfrak{m}, \mathfrak{m}' \in \mathbb{T}_{\mathcal{M}_t}$, the instantaneous gradients satisfy*

$$d\left(\boldsymbol{V}_t\left(\theta,\, \mathfrak{m}\right),\, \boldsymbol{V}_t\left(\theta',\, \mathfrak{m}'\right)\right) \leq \kappa_5 \left(d\left(\theta,\, \theta'\right) + d\left(\mathfrak{m},\, \mathfrak{m}'\right)\right) m_{\mathrm{H}}(t).$$

The reason we put a separate scale function $m_{\mathrm{H}}(t)$ instead of reusing $m(t)$ is because these can be different even in very basic examples. Indeed, $m(t)$ controls the size of gradients whereas $m_{\mathrm{H}}(t)$ controls the Lipschitz constant of gradients, i.e., essentially the Hessian. For instance, for a linear model with loss $\frac{1}{2}(y_t - \theta \cdot x)^2$, the gradients are $\theta \cdot x - y_t$ and if y_t is unbounded, then $m(t)$ might be large. But the *differences* of gradients are $\theta \cdot x - \theta' \cdot x'$ which do not depend on the norm of y_t: in this example the Hessian is bounded so that $m_{\mathrm{H}}(t)$ is constant.

4.2.3 Parameter Updates

We would like the update operators to cover parameter-dependent updates such as $\Phi(\theta, v) = \theta - P(\theta)\, v + O(\|v\|^2)$ where P is a θ-dependent linear operator. This is covered by the following assumptions, which basically state that $\Phi(\theta, v) - \theta$ is Lipschitz w.r.t. both θ and v, and the Lipschitz constant w.r.t. θ is controlled by $\|v\|$.

Assumption 4.17 (Parameter update operators)**.** *We assume that the parameter update operators Φ_t satisfy:*

1. *for any $\theta \in \Theta$ and $t \geq 1$, $\Phi_t(\theta, 0) = \theta$;*
2. *there exists $r_{\mathcal{V}} > 0$ and $c_\Phi > 0$ such that for any parameters $\theta, \theta' \in B^*_\Theta$, for any $t \geq 1$, for any $v, v' \in \mathcal{V}_t$ with $\|v\| \leq r_{\mathcal{V}}$ and $\|v'\| \leq r_{\mathcal{V}}$, then*

$$d\left(\Phi_t\left(\theta, v\right),\, \Phi_t\left(\theta', v'\right)\right) \leq d\left(\theta,\, \theta'\right) + c_\Phi \left(d\left(v,\, v'\right) + \left(\|v\| + \|v'\|\right) d\left(\theta,\, \theta'\right)\right).$$

For instance, the assumption is satisfied with $\theta - P(\theta)\, v + \|v\|^2 f_2(\theta, v)$ with locally Lipschitz second-order term f_2. It also works with $\Phi_t(\theta, v) = \arg\min_{\theta'}\{v \cdot (\theta' - \theta) + \frac{1}{2} D_t(\theta|\theta')\}$ where D_t is a suitable second-order penalty between θ and θ', such as a KL divergence (the trivial case being $D = \|\theta - \theta'\|^2$).

4.2.4 Local Optimality of θ^*

We now formulate the crucial assumption of local optimality for a parameter: it mimicks classical second-order optimality conditions, in a way adapted to our sequential setting.

Assumption 4.18 (Local optimality of θ^*)**.** *We assume there exists a scale function L, negligible with respect to the identity function near infinity, and $\overline{\eta}_{\mathrm{op}} > 0$ such that the following properties are satisfied.*

1. *First-order stability condition. For $\overline{\eta} \leq \overline{\eta}_{\mathrm{op}}$,*

$$d\left(\Phi_{t:t+L(t)}\left(\theta^*, \mathfrak{m}^*_t, \eta_t\right),\, \theta^*\right) = o\left(\eta_t\, L(t)\right)$$

when $t \to \infty$, uniformly in $\overline{\eta} \leq \overline{\eta}_{\mathrm{op}}$.

2. *Second-order or contractivity condition. There exists* $\lambda_{\min} > 0$ *with the following property. For any* $\overline{\eta} \leq \overline{\eta}_{\mathrm{op}}$*, for any parameter* $\theta \in B^*_\Theta$ *then*

$$d\left(\Phi_{t:t+L(t)}\left(\theta, \mathfrak{m}^*_t, \eta_t\right), \Phi_{t:t+L(t)}\left(\theta^*, \mathfrak{m}^*_t, \eta_t\right)\right)$$

is at most

$$\left(1 - \lambda_{\min}\, \eta_t\, L(t)\right)\, d\left(\theta,\, \theta^*\right) + o\left(\eta_t\, L(t)\right)$$

when $t \to \infty$*. Moreover, this* $o()$ *is uniform over* $0 \leq \overline{\eta} \leq \overline{\eta}_{\mathrm{op}}$ *and over* $\theta \in B^*_\Theta$*.*

4.2.5 Constraints on the Stepsize Sequence

The various assumptions above constrain the admissible step size sequences we may use for the descent. Remember (Def. 4.7) that the step sizes η_t are defined via $\eta_t := \overline{\eta}\, \rho_t$ with overall learning rate $\overline{\eta}$ and stepsize schedule (ρ_t).

Assumption 4.19 (Stepsize sequence: behaviour of the stepsize schedule)**.** *The stepsize sequence* $\boldsymbol{\eta} = (\eta_t)$ *and the scale functions satisfy:*

1. η_t *is positive, and* $\sum_{t=1}^{\infty} \eta_t = +\infty$*.*
2. $\eta_t\, L(t)\, m(t)\, m_{\mathrm{H}}(t) \to 0$ *when* $t \to \infty$*.*
3. $m_{\mathrm{H}}(t) \ll L(t)$ *when* $t \to \infty$*.*
4. *When* $t \to \infty$*,*

$$\frac{\sup_{t<s\leq t+L(t)} \eta_s}{\inf_{t<s\leq t+L(t)} \eta_s} = 1 + o(1/m(t)).$$

Remark 4.20. *These assumptions are invariant by scaling all* η_s *by the same factor. So they are a property of the stepsize schedule and not of* $\overline{\eta}$*.*

The sup/inf assumption is automatically satisfied if η_t is the inverse of a scale function and $m(t) \equiv 1$ (because Assumption 4.18 asks that we have $L(t) = o\,(t)$, as $t \to \infty$, so that $t + L(t) \sim t$).

4.2.6 Timescale Adapted to the Optimality Criterion

To study the algorithm, we consider it in the following time-scale, which is adapted to the dynamics of the optimality assumptions.

Definition 4.21 (Timescale associated with a scale function.)**.** *Let* $L()$ *be the scale function appearing in Assumption 4.18. We define the integer sequence* $(T_k)_{k\geq 0}$ *by induction via* $T_0 = 0$*,* $T_1 = 1$ *and, for* $k \geq 1$*,*

$$T_{k+1} = T_k + L(T_k).$$

We denote the integer intervals defined by this timescale by

$$I_k := (T_k; T_{k+1}].$$

Lemma 4.22 (Asymptotic behavior of (T_k))**.** (T_k) *is strictly increasing, and tends to infinity. Moreover,* $T_{k+1} \sim T_k$ *when* $k \to \infty$*.*

Proof. The first statements follow from $L(T) \geq 1$ for $T \geq 1$. The last statement follows from $L(T) = o(T)$ when $T \to \infty$. □

4.3 Noisy Updates

We now introduce definitions to deal with sequences of states $(\mathfrak{m}_t)$ that do not follow the algorithm $\mathcal{A}_t$, but nevertheless stay close to it in some sense. We refer to this as the *noisy* case. This will be useful to deal with stochastic approximations of RTRL such as NoBackTrack or UORO. Thus, the states $\mathfrak{m}_t$ may be random, but the parameter updates θ_t are still computed normally from $\mathfrak{m}_t$ via v_t.

Definition 4.23 (Random trajectory)**.** *A* random trajectory *is, for every stepsize sequence (η_t), a probability distribution over trajectories $(\mathfrak{m}_t, v_t, \theta_t)_{t\geq 0}$, such that*

$$\begin{cases} v_t = \boldsymbol{V}_t\left(\theta_{t-1},\, \mathfrak{m}_t\right) \\ \theta_t = \Phi_t\left(\theta_{t-1},\, \eta_t\, v_t\right) \end{cases}$$

holds for all $t \geq 1$, with probability 1*.*

We say that a random trajectory respects the stable tube *if, for any stepsize sequence, for every $t \geq 1$, $\mathfrak{m}_{t-1} \in \mathbb{T}_{\mathcal{M}_{t-1}}$ and $\theta_{t-1} \in B^*_\Theta$ imply $\mathfrak{m}_t \in \mathbb{T}_{\mathcal{M}_t}$ with probability* 1*.*

The next notion captures the noise produced on θ_t by a sequence of states $(\mathfrak{m}_t)$ that does not follow the algorithm $\mathcal{A}_t$. This definition compares the sequence $(\mathfrak{m}_t)$ to a sequence $(\bar{\mathfrak{m}}_t)$ starting at the same state but that follows $\mathcal{A}_t$. This extends to our setting the noise on J_t in imperfect RTRL algorithms compared to exact RTRL.

Definition 4.24 (Deviation of a random trajectory from an algorithm)**.** *Let $(\mathfrak{m}_t)$ be any sequence of states. Let $t_0 \geq 0$ and let θ_{t_0} be any parameter. Consider the following two sequences (θ_t) and $(\bar{\theta}_t)$ defined from $(\mathfrak{m}_t)$ and θ_{t_0} as follows.*

Define (θ_t) as the parameter trajectory defined from the sequence of states $(\mathfrak{m}_t)$ starting at θ_{t_0} at time t_0, namely,

$$v_t = \boldsymbol{V}_t\left(\theta_{t-1},\, \mathfrak{m}_t\right), \qquad \theta_t = \Phi_t(\theta_{t-1},\, \eta_t v_t)$$

for $t > t_0$.

Let $(\bar{\mathfrak{m}}_t)$ be the algorithm trajectory starting at $\mathfrak{m}_{t_0}$ with parameters (θ_t) ("regularized" trajectory), and $\left(\bar{\theta}_t\right)$ be the sequence of parameters obtained with it. Namely, define $\bar{\mathfrak{m}}_{t_0} = \mathfrak{m}_{t_0}$, $\bar{\theta}_{t_0} = \theta_{t_0}$ and

$$\bar{\mathfrak{m}}_t = \mathcal{A}_t(\theta_{t-1}, \bar{\mathfrak{m}}_{t-1}), \qquad \bar{v}_t = \boldsymbol{V}_t\left(\theta_{t-1},\, \bar{\mathfrak{m}}_t\right), \qquad \bar{\theta}_t = \Phi_t(\bar{\theta}_{t-1},\, \eta_t \bar{v}_t).$$

for $t > t_0$. (Beware $\bar{\mathfrak{m}}_t$ and $\bar{v}_t$ use θ_{t-1} while $\bar{\theta}_t$ uses $\bar{\theta}_{t-1}$.)

We call deviation of $(\mathfrak{m}_t)$ from $(\mathcal{A}_t)$ at time t_1 *the quantity*

$$D_{t_0:t_1}\left(\theta_{t_0},\, (\mathfrak{m}_t),\, \boldsymbol{\eta}\right) := d\left(\theta_{t_1},\, \bar{\theta}_{t_1}\right).$$

Definition 4.25 (Negligible noise)**.** *Let $K \geq 0$, and let (δ_k) be a nonnegative sequence which tends towards* 0*.*

We say that a trajectory (θ_t), $(\mathfrak{m}_t)$ has negligible noise *in the timescale (T_k), from time K onwards, at speed (δ_k) if, for all $k \geq K$, we have*

$$\theta_{T_k} \in B^*_\Theta, d\left(\theta_{T_k},\, \theta^*\right) \leq \frac{r^*_\Theta}{3}, \mathfrak{m}_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}} \Rightarrow D_{T_k:T_{k+1}}\left(\theta_{T_k},\, (\mathfrak{m}_t),\, \boldsymbol{\eta}\right) \leq \delta_k\, \eta_{T_k}\, L(T_k).$$

4.4 Convergence Theorems

We now formulate the convergence theorem for our abstract model, in three cases: non-noisy, noisy (with random trajectories), and open-loop on intervals (which models truncated backpropagation through time). For the noisy case, this works under the assumption that the deviation from the non-noisy case is negligible in the sense above, with high probability.

We first gather all assumptions on the system.

Definition 4.26. *We call Assumption* $\boldsymbol{A}$ *the following setting.*

Let Θ *be some metric space, the* parameter space. *Let* $(\mathcal{M}_t)_{t\geq 0}$ *be a sequence of metric spaces, which represents the objects maintained in memory by an algorithm at time* t. *Let* $(\mathcal{V}_t)_{t\geq 1}$ *be a sequence of normed vector spaces, containing the gradients computed at time* t.

Let $(\mathcal{A}_t)$ *be a family of transition operators (Definition 4.2) admitting a stable tube (Assumption 4.11). Assume the state updates are Lipschitz with respect to the parameter (Assumption 4.13), and the family forgets exponentially fast the initialisation when the parameter is fixed (Assumption 4.14).*

Let $(\boldsymbol{V}_t)$ *be gradient computation operators (Definition 4.3), which are locally bounded (Assumption 4.15) and Lipschitz (Assumption 4.16).*

Let (Φ_t) *be parameter update operators (Definition 4.5), which are Lipschitz with respect to the parameter and the tangent vectors (Assumption 4.17).*

Let $\boldsymbol{\eta} = (\eta_t)_{t\geq 1}$ *be a stepsize sequence with overall learning rate* $\overline{\eta}$ *(Def. 4.7), whose asymptotic behavior satisfies Assumption 4.19.*

Let θ^* *be a parameter satisfying the local optimality conditions of Assumption 4.18.*

Theorem 4.27 (Convergence of the gradient descent algorithm)**.** *Consider a system satisfying Assumption* $\boldsymbol{A}$.

Then there exists $\overline{\eta}_{\text{conv}} > 0$ *such that, for any overall learning rate* $\overline{\eta} < \overline{\eta}_{\text{conv}}$, *the following convergence holds. For any parameter* θ_0 *and maintained quantity* $\mathfrak{m}_0$ *satisfying*

$$(\theta_0,\, \mathfrak{m}_0) \in \left\{ \theta \in \Theta \,\middle|\, d\,(\theta,\, \theta^*) \leq \frac{r^*_\Theta}{4} \right\} \times \mathbb{T}_{\mathcal{M}_0},$$

consider the gradient descent trajectory given by

$$\begin{cases} \mathfrak{m}_t = \mathcal{A}_t\,(\theta_{t-1},\, \mathfrak{m}_{t-1}) \\ v_t = \boldsymbol{V}_t\,(\theta_{t-1},\, \mathfrak{m}_t) \\ \theta_t = \Phi_t\,(\theta_{t-1},\, \eta_t\, v_t)\,, \end{cases}$$

for $t \geq 1$. *Then* θ_t *tends to* θ^* *as* $t \to \infty$.

Theorem 4.28 (Convergence of the gradient descent algorithm, noisy case)**.** *Consider a system satisfying Assumption* $\boldsymbol{A}$.

Assume that we are given a random trajectory $(\mathfrak{m}_t, v_t, \theta_t)_{t\geq 0}$ *which respects the stable tube* $(\mathbb{T}_{\mathcal{M}_t})_{t\geq 0}$, *in the sense of Definition 4.23 (namely,* $\mathfrak{m}_t$ *is random but lies in* $\mathbb{T}_{\mathcal{M}_t}$ *provided* $\mathfrak{m}_{t-1} \in \mathbb{T}_{\mathcal{M}_{t-1}}$ *and* $\theta_{t-1} \in B^*_\Theta$, *while* v_t *and* θ_t *are updated as in the non-noisy case).*

Assume that $\mathfrak{m}_0 \in \mathbb{T}_{\mathcal{M}_0}$ *and* $d\,(\theta_0,\, \theta^*) \leq \frac{r^*_\Theta}{4}$.

Assume there exists $\varepsilon > 0$, $K \geq 0$, *a non-negative sequence* (δ_k) *which tends to* 0, *and* $\overline{\eta}_{\text{noise}} > 0$ *such that, for all* $\bar{\eta} \leq \overline{\eta}_{\text{noise}}$, *with probability greater than*

$1-\varepsilon$, *the random trajectory* $(\mathfrak{m}_t, \theta_t)$ *has negligible noise starting at* K *at speed* (δ_k) *(Definition 4.25).*

Then there exists $\overline{\eta}_{\text{conv}} > 0$ *such that for any* $\overline{\eta} < \overline{\eta}_{\text{conv}}$*, with probability at least* $1-\varepsilon$*,* θ_t *tends to* θ^* *as* $t \to \infty$*.*

We now define the open-loop algorithm, which models truncated backpropagation through time (in the "non-overlapping" variant, see Section 3.2): the parameter is updated only once at the end of each time interval $(T_k, T_{k+1}]$, by collecting all gradients computed during that interval. In TBPTT using time intervals $(T_k, T_{k+1}]$, whenever the parameter is updated, gradients are not backpropagated through the boundary, but reset to 0: this corresponds to resetting the RTRL state derivative J to 0. Moreover, the state may or may not be reset to some default value at the start of each new TBPTT interval. So here, at the end of every time interval $(T_k, T_{k+1}]$, the running quantity $\mathfrak{m}_{T_k}$ is discarded and reset to some $\mathfrak{m}'_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$ (typically, $\mathfrak{m}'_{T_k} = 0$, which belongs to $\mathbb{T}_{\mathcal{M}_{T_k}}$ as shown in Corollary 6.24). Note that $\mathfrak{m}_{T_k}$ is still used to compute the gradient update for the last step of the previous time interval, hence our use of a substitution $\mathfrak{m}_{T_k} \leftarrow \mathfrak{m}'_{T_k}$ at the beginning of each new interval.

Theorem 4.29 (Convergence of the open-loop gradient descent algorithm)**.** *Consider a system satisfying Assumption* $\boldsymbol{A}$*.*

Let (T_k) *be the time-scale* (T_k) *of Definition 4.21, namely,* $T_{k+1} = T_k + L(T_k)$ *where* $L()$ *is from Assumption 4.18.*

There exists $\overline{\eta}_{\text{conv}} > 0$ *such that, for any overall learning rate* $\overline{\eta} < \overline{\eta}_{\text{conv}}$*, the following convergence holds. For any parameter* θ_0 *and maintained quantity* $\mathfrak{m}_0$ *satisfying*

$$(\theta_0, \mathfrak{m}_0) \in \left\{ \theta \in \Theta \,\middle|\, d(\theta, \theta^*) \leq \frac{r^*_\Theta}{4} \right\} \times \mathbb{T}_{\mathcal{M}_0},$$

for any sequence of reset states $(\mathfrak{m}'_{T_k})_{k\geq 0}$ *with* $\mathfrak{m}'_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$*, consider the open-loop gradient descent trajectory which resets the state* $\mathfrak{m}$ *to* $\mathfrak{m}'$ *at the start of every time interval* $(T_k, T_{k+1}]$*, and updates the parameter at the end of every time interval; namely, for each* $k \geq 0$*, the computation performed in the interval* $(T_k, T_{k+1}]$ *is*

$$\mathfrak{m}_{T_k} \leftarrow \mathfrak{m}'_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$$

and, for $T_k < t \leq T_{k+1}$*,*

$$\begin{cases} \mathfrak{m}_t = \mathcal{A}_t(\theta_{T_k}, \mathfrak{m}_{t-1}) \\ v_t = \boldsymbol{V}_t(\theta_{T_k}, \mathfrak{m}_t) \\ \theta_t = \Phi_t\left(\theta_{t-1}, \eta_{T_{k+1}} v_t\right). \end{cases}$$

Then θ_t *tends to* θ^* *as* $t \to \infty$*.*

The open-loop algorithm models TBPTT when $\Phi_t(\theta, v) = \theta - v$. For more complicated Φ, we would also have the option to update the parameter once with the sum of gradients via $\theta_{T_{k+1}} = \Phi_{T_{k+1}}(\theta_{T_k}, \eta_{T_{k+1}} \sum_{T_k+1}^{T_{k+1}} v_t)$ instead of applying Φ at every step. In general, the difference between these two options is of second-order, as shown in the course of the proofs.

5 Proof of Convergence for the Abstract Algorithm

We now turn to the proof of Theorems 4.27, 4.28 and 4.29. The notation and assumptions are as in Section 4.

The proof proceeds in three main stages. First, we derive a priori bounds on the trajectories. Then, we quantify the amount by which trajectories diverge from each other as time goes on. This divergence will be negated by the contractivity properties satisfied around the local optimum. Finally, we are able to prove convergence.

Remark 5.1. *All statements in the first two subsections can be made to start at an arbitrary time $t_0 \geq 0$ rather than at time 0, by applying them to the operators $\mathcal{A}_{t+t_0}$, $\boldsymbol{V}_{t+t_0}$, Φ_{t+t_0}, stepsizes η_{t+t_0}, etc., which satisfy the same assumptions as those using $t_0 = 0$.*

Remark 5.2. *In all proofs in the text, when we write $O()$, the constants implied in the $O()$ notation depend only on the constants explicitly appearing in the assumptions and on the constants implied in those $O()$ appearing in the assumptions. In particular, the $O()$ notation is always uniform over other quantities of interest such as θ, $\mathfrak{m}$, J, $\overline{\eta}$, etc.*

5.1 *A Priori* Bounds on Trajectories

5.1.1 Admissible Learning Rates

First, we define the maximum overall learning rate we will consider in the proofs. The bound depends on the magnitude of the gradientsi in the assumptions. (This is not yet the maximum learning rate allowed for the final convergence theorem, for which further constraints will be needed.)

Definition 5.3 (Bound on the learning rate)**.** *We define*

$$\overline{\eta}_{\mathcal{V}} := \min\left(1, \frac{1}{\sup_{s\geq 1} \rho_s\, m(s)}\right) \min\left(1, \frac{r_{\mathcal{V}}}{\sup_{t\geq 1} \sup_{\theta\in B_\Theta^*} \sup_{\mathfrak{m}\in\mathbb{T}_{\mathcal{M}_t}} \|\boldsymbol{V}_t(\theta,\mathfrak{m})\|\, m(t)^{-1}}\right)$$

where $r_{\mathcal{V}}$ is the value from Assumption 4.17.

Remark 5.4. *By the second point of Assumption 4.19, the sequence $(\rho_s\, m(s))$ is bounded, so that the supremum of the $\rho_s\, m(s)$'s is well-defined. By Assumption 4.15, the supremum of $\|\boldsymbol{V}_t(\theta,\mathfrak{m})\|\, m(t)^{-1}$ is finite. Therefore, $\overline{\eta}_{\mathcal{V}} > 0$.*

Corollary 5.5. *Let $t \geq 1$, and $v_t \in B_{\mathcal{V}_t}$. Then $\|v_t\| \leq m(t)\, r_{\mathcal{V}}/\overline{\eta}_{\mathcal{V}}$.*

Moreover, for any $0 \leq \overline{\eta} \leq \overline{\eta}_{\mathcal{V}}$, for any $t \geq 1$, for any $v \in B_{\mathcal{V}_t}$ we have $\eta_t\, v \in B_{\mathcal{V}_t}$ and $\|\eta_t\, v\| \leq r_{\mathcal{V}}$.

Proof. The first assertion is true by definition of $\overline{\eta}_{\mathcal{V}}$: indeed $\overline{\eta}_{\mathcal{V}} \leq r_{\mathcal{V}}/(\|v_t\|\, m(t)^{-1})$ for any $v_t \in B_{\mathcal{V}_t}$, by definition of $B_{\mathcal{V}_t}$.

If $\overline{\eta} \leq \overline{\eta}_{\mathcal{V}}$, then $\eta_t \leq 1$. Indeed, $\eta_t = \overline{\eta}\rho_t$ and $\overline{\eta}_{\mathcal{V}} \leq 1/(\rho_t m(t)) \leq 1/\rho_t$ because $m(t) \geq 1$ as a scale function. This proves that $\eta_t\, v \in B_{\mathcal{V}_t}$ if $v \in B_{\mathcal{V}_t}$.

For the last assertion, for any $t \geq 1$, we have

$$\begin{aligned}
\|\eta_t\, v\| &\leq \overline{\eta}_{\mathcal{V}}\, \rho_t\, \|v\| \\
&\leq \left(\sup_{s\geq 1} \rho_s\, m(s)\right)^{-1} \rho_t\, \frac{r_{\mathcal{V}}\, \|v\|}{\sup_{s\geq 1} \sup_{\theta\in B^*_\Theta} \sup_{\mathfrak{m}\in\mathbb{T}_{\mathcal{M}_s}} \|\boldsymbol{V}_s(\theta,\mathfrak{m})\|\, m(s)^{-1}} \\
&\leq \left(\sup_{s\geq 1} \rho_s\, m(s)\right)^{-1} \rho_t\, \frac{r_{\mathcal{V}}\, \|v\|}{\sup_{\theta\in B^*_\Theta} \sup_{\mathfrak{m}\in\mathbb{T}_{\mathcal{M}_t}} \|\boldsymbol{V}_t(\theta,\mathfrak{m})\|\, m(t)^{-1}} \\
&\leq \left(\sup_{s\geq 1} \rho_s\, m(s)\right)^{-1} \rho_t\, r_{\mathcal{V}}\, m(t) \\
&\leq r_{\mathcal{V}}.
\end{aligned}$$

□

5.1.2 Short-Time Stability

In this subsection, we do not assume that Assumption 4.13, Assumption 4.14 or Assumption 4.18 hold.

Corollary 5.6 (Stability of states)**.** *Let (θ_t) be a sequence of parameters in B^*_Θ, and let $\mathfrak{m}_0 \in \mathbb{T}_{\mathcal{M}_0}$. Let $(\mathfrak{m}_t)$ be the trajectory associated with (θ_t) starting at $\mathfrak{m}_0$. Then for all $t \geq 0$, $\mathfrak{m}_t \in \mathbb{T}_{\mathcal{M}_t}$.*

Proof. By induction from Assumption 4.11. □

Corollary 5.7 (Smoothness of parameter updates)**.** *There exists $\mathrm{M} > 0$ such that, for any $t \geq 1$, for any parameters $\theta, \theta' \in B^*_\Theta$, for any $v, v' \in B_{\mathcal{V}_t}$, for any $0 \leq \overline{\eta} \leq \overline{\eta}_{\mathcal{V}}$,*

$$d\left(\Phi_t\left(\theta, \eta_t\, v\right),\, \Phi_t\left(\theta', \eta_t\, v'\right)\right) \leq d\left(\theta,\, \theta'\right) + \mathrm{M}\, \eta_t\, m(t).$$

Proof. By Corollary 5.5, for $v \in B_{\mathcal{V}_t}$, we have $\|\eta_t v\| \leq \eta_t\, m(t)\, r_{\mathcal{V}}/\overline{\eta}_{\mathcal{V}}$, and likewise for v'. Then the assertion follows from Assumption 4.17 by setting $\mathrm{M} := c_\Phi(2r_{\mathcal{V}} + 4r_{\mathcal{V}} r^*_\Theta)/\overline{\eta}_{\mathcal{V}}$. □

Definition 5.8 (Safe time horizon for staying in B^*_Θ)**.** *Let $\boldsymbol{\eta} = (\eta_s)$ be a stepsize sequence, and let $t_0 \geq 0$. We define*

$$T_{t_0}^{r^*_\Theta}(\boldsymbol{\eta}) = \inf\left\{ t \geq t_0 + 1 \;\middle|\; \sum_{s=t_0+1}^{t} \eta_s\, m(s) > \frac{r^*_\Theta}{3\mathrm{M}} \right\},$$

or $T_{t_0}^{r^*_\Theta} = \infty$ if this set is empty.

The next lemma shows that a parameter trajectory $\theta_t = \Phi_t\left(\theta_{t-1},\, \eta_t\, v_t\right)$ stays in the stable tube for a time at least $T_0^{r^*_\Theta}(\boldsymbol{\eta})$, provided it is computed from states $\mathfrak{m}_t$ and parameters $\bar{\theta}_t$ within the stable tube.

Lemma 5.9 (Parameter trajectories stay in the stable tube for a time $T_0^{r^*_\Theta}(\boldsymbol{\eta})$.)**.** *Let $\boldsymbol{\eta} = (\eta_t)_{t\geq 1}$ be a stepsize sequence with $\overline{\eta} \leq \overline{\eta}_{\mathcal{V}}$.*

Let (θ_t), $(\bar{\theta}_t)$ be sequences of parameters, $(\mathfrak{m}_t)$ a sequence of states, and (v_t) a sequence of gradients, such that $\theta_0 \in B^*_\Theta$ with $d(\theta_0, \theta^*) \leq r^*_\Theta/3$, $\bar{\theta}_0 \in B^*_\Theta$, $\mathfrak{m}_0 \in \mathbb{T}_{\mathcal{M}_0}$, and for any $t \geq 1$,

$$\begin{cases} \mathfrak{m}_t = \mathcal{A}_t(\bar{\theta}_{t-1}, \mathfrak{m}_{t-1}) & \text{or } \mathfrak{m}_t \in \mathbb{T}_{\mathcal{M}_t} \\ v_t = \boldsymbol{V}_t\left(\bar{\theta}_{t-1},\, \mathfrak{m}_t\right) & \text{or } v_t \in B_{\mathcal{V}_t} \\ \theta_t = \Phi_t\left(\theta_{t-1},\, \eta_t\, v_t\right) & \\ \bar{\theta}_t = \theta_s \text{ for some } s \leq t, & \text{or } \bar{\theta}_t \in B^*_\Theta. \end{cases}$$

Then for all $1 \leq t < T_0^{r^*_\Theta}(\boldsymbol{\eta})$, the trajectory lies in the stable tube: $\theta_t, \bar{\theta}_t \in B^*_\Theta$, $v_t \in B_{\mathcal{V}_t}$, and $\mathfrak{m}_t \in \mathbb{T}_{\mathcal{M}_t}$.

Proof. Let us prove by induction on t that the conclusion holds and that moreover, for $t \geq 1$, we have

$$d(\theta_t, \theta^*) \leq \frac{r^*_\Theta}{3} + \mathrm{M} \sum_{s=1}^{t} \eta_s\, m(s).$$

This holds at time $t = 0$ by assumption.

By Assumption 4.11, $\mathfrak{m}_t = \mathcal{A}_t\left(\bar{\theta}_{t-1}, \mathfrak{m}_{t-1}\right)$ belongs to $\mathbb{T}_{\mathcal{M}_t}$ provided the conclusion holds at time $t-1$. By Assumption 4.15, v_t belongs to $B_{\mathcal{V}_t}$ provided the conclusion holds at time $t-1$.

Then from Corollary 5.7 applied to $(\theta_{t-1}, \eta_t v_t)$ and $(\theta^*, 0)$, we find

$$\begin{aligned} d(\theta_t, \theta^*) &= d\left(\Phi_t\left(\theta_{t-1},\, \eta_t\, v_t\right),\, \theta^*\right) \\ &\leq d(\theta_{t-1}, \theta^*) + \mathrm{M}\, \eta_t\, m(t) \\ &\leq \frac{r^*_\Theta}{3} + \mathrm{M} \sum_{s=1}^{t} \eta_s\, m(s) \end{aligned}$$

by the induction hypothesis at time $t-1$. For $t < T_0^{r^*_\Theta}$, by definition of $T_0^{r^*_\Theta}$, this is at most $\frac{2\, r^*_\Theta}{3}$. So θ_t belongs to B^*_Θ and the induction hypothesis holds at time t. □

Lemma 5.10 (Parameter trajectories stay in the stable tube for a time $T_0^{r^*_\Theta}(\boldsymbol{\eta})$ for trajectories which respect the stable tube)**.** *Let $\boldsymbol{\eta} = (\eta_t)_{t \geq 1}$ be a stepsize sequence with $\overline{\eta} \leq \overline{\eta}_{\mathcal{V}}$.*

Let $(\mathfrak{m}_t, v_t, \theta_t)_{t \geq 0}$ be a random trajectory which respects the stable tube, in the sense of Definition 4.23. Assume that $\theta_0 \in B^*_\Theta$ with $d(\theta_0, \theta^*) \leq r^*_\Theta/3$ and $\mathfrak{m}_0 \in \mathbb{T}_{\mathcal{M}_0}$. Then for all $1 \leq t < T_0^{r^*_\Theta}(\boldsymbol{\eta})$, the trajectory lies in the stable tube: with probability 1, $\theta_t \in B^*_\Theta$, $v_t \in B_{\mathcal{V}_t}$, and $\mathfrak{m}_t \in \mathbb{T}_{\mathcal{M}_t}$.

Proof. The proof is identical to that of Lemma 5.9. □

Note that by taking the overall learning rate $\overline{\eta}$ small enough, we can ensure that $T_0^{r^*_\Theta}(\boldsymbol{\eta})$ is arbitrarily large. We shall need this later, in case the contractivity property of gradient descent kicks in late, so we state the following.

Lemma 5.11 (Small learning rates for arbitrary control horizon)**.** *Let $T > 0$. Then there exists $\overline{\eta}^T > 0$ with the following property:*

For any $0 \leq \overline{\eta} \leq \overline{\eta}^T$, *for any* $\theta_0 \in B^*_\Theta$ *such that* $d(\theta_0, \theta^*) \leq \frac{r^*_\Theta}{4}$ *and any* $\mathfrak{m}_0 \in \mathbb{T}_{\mathcal{M}_0}$, *consider a trajectory* (θ_t), $(\mathfrak{m}_t)$, (v_t) *such that, for all* $t \geq 1$,

$$\begin{cases} \mathfrak{m}_t = \mathcal{A}_t(\theta_s, \mathfrak{m}_{t-1}) \text{ for some } s \leq t-1, & \text{or } \mathfrak{m}_t \in \mathbb{T}_{\mathcal{M}_t}, \\ v_t = \boldsymbol{V}_t(\theta_s, \mathfrak{m}_t) \text{ for some } s \leq t-1, & \text{or } v_t \in B_{\mathcal{V}_t}, \\ \theta_t = \Phi_t(\theta_{t-1}, \eta_t v_t). \end{cases}$$

Then $\theta_T \in B^*_\Theta$, $d(\theta_T, \theta^*) \leq \frac{r^*_\Theta}{3}$, *and* $\mathfrak{m}_T \in \mathbb{T}_{\mathcal{M}_T}$.

Proof. Remember that $\eta_t = \overline{\eta}\rho_t$ where ρ_t is the stepsize schedule. Define $\overline{\eta}^T$ such that $\overline{\eta}^T \sum_{t=1}^T \rho_t m(t) \leq \frac{r^*_\Theta}{12\mathrm{M}}$. Then proceed similarly to Lemma 5.9, as follows.

Let us prove by induction on t that, for $t \geq 0$, we have that $\mathfrak{m}_t \in \mathbb{T}_{\mathcal{M}_t}$, that $v_t \in B_{\mathcal{V}_t}$, that $\theta_s \in B^*_\Theta$ for every $s \leq t-1$, and that, for $t \geq 1$,

$$d(\theta_t, \theta^*) \leq \frac{r^*_\Theta}{4} + \mathrm{M} \sum_{s=1}^{t} \eta_s m(s).$$

This holds at time $t = 0$ by assumption.

By Assumption 4.11, $\mathfrak{m}_t = \mathcal{A}_t(\theta_{s-1}, \mathfrak{m}_{t-1})$ for some $s \leq t-1$ belongs to $\mathbb{T}_{\mathcal{M}_t}$ provided the conclusion holds at time $t-1$. By Assumption 4.15, $v_t = \boldsymbol{V}_t(\theta_s, \mathfrak{m}_t)$ for some $s \leq t-1$ belongs to $B_{\mathcal{V}_t}$ provided the conclusion holds at time $t-1$.

Then from Corollary 5.7 applied to $(\theta_{t-1}, \eta_t v_t)$ and $(\theta^*, 0)$ we find

$$\begin{aligned} d(\theta_t, \theta^*) &= d(\Phi_t(\theta_{t-1}, \eta_t v_t), \theta^*) \\ &\leq d(\theta_{t-1}, \theta^*) + \mathrm{M}\eta_t m(t) \\ &\leq \frac{r^*_\Theta}{4} + \mathrm{M} \sum_{s=1}^{t} \eta_s m(s) \end{aligned}$$

by the induction hypothesis at time $t-1$. For $t \leq T$, by definition of $\overline{\eta}^T$ this is at most $\frac{r^*_\Theta}{4} + \frac{r^*_\Theta}{12} = \frac{r^*_\Theta}{3}$. So θ_t belongs to B^*_Θ and the induction hypothesis holds at time t. □

We now state a version of this lemma for random trajectories in the sense of Definition 4.23.

Lemma 5.12 (Small learning rates for arbitrary control horizon, for random trajectories)**.** *Let* $T > 0$. *Then there exists* $\overline{\eta}^T > 0$ *with the following property.*

For any $0 \leq \overline{\eta} \leq \overline{\eta}^T$, *for any* $\theta_0 \in B^*_\Theta$ *such that* $d(\theta_0, \theta^*) \leq \frac{r^*_\Theta}{4}$ *and any* $\mathfrak{m}_0 \in \mathbb{T}_{\mathcal{M}_0}$, *for any random trajectory* (θ_t), $(\mathfrak{m}_t)$, (v_t) *which starts at* $(\theta_0, \mathfrak{m}_0)$ *and respects the stable tube (Def. 4.23), with probability* 1 *it holds that* $\theta_T \in B^*_\Theta$, $d(\theta_T, \theta^*) \leq \frac{r^*_\Theta}{3}$, *and* $\mathfrak{m}_T \in \mathbb{T}_{\mathcal{M}_T}$.

Proof. By Definition 4.23, a random trajectory which respects the stable tube satisfies, for all $t \geq 1$

$$\begin{cases} \mathfrak{m}_{t-1} \in \mathbb{T}_{\mathcal{M}_{t-1}} \text{ and } \theta_{t-1} \in B^*_\Theta \implies \mathfrak{m}_t \in \mathbb{T}_{\mathcal{M}_t} & \text{w.p. 1} \\ v_t = \boldsymbol{V}_t(\theta_{t-1}, \mathfrak{m}_t), \\ \theta_t = \Phi_t(\theta_{t-1}, \eta_t v_t). \end{cases}$$

Remember that $\eta_t = \overline{\eta}\rho_t$ where ρ_t is the stepsize schedule. Define $\overline{\eta}^T$ such that $\overline{\eta}^T \sum_{t=1}^T \rho_t m(t) \leq \frac{r^*_\Theta}{12\mathrm{M}}$. Then proceed similarly to Lemma 5.9, as follows.

Let us prove by induction on t that, with probability 1, for $t \geq 0$, we have that $\mathfrak{m}_t \in \mathbb{T}_{\mathcal{M}_t}$, that $v_t \in B_{\mathcal{V}_t}$, that $\theta_t \in B^*_\Theta$ and that, for $t \geq 1$,

$$d(\theta_t, \theta^*) \leq \frac{r^*_\Theta}{4} + \mathrm{M} \sum_{s=1}^{t} \eta_s \, m(s).$$

This holds at time $t = 0$ by assumption.

Since we assume that $\mathfrak{m}_{t-1} \in \mathbb{T}_{\mathcal{M}_{t-1}}$ and $\theta_{t-1} \in B^*_\Theta \implies \mathfrak{m}_t \in \mathbb{T}_{\mathcal{M}_t}$, if the conclusion holds at time $t-1$, then $\mathfrak{m}_t$ belongs to $\mathbb{T}_{\mathcal{M}_t}$ with probability one. Then, by Assumption 4.15, $v_t = \boldsymbol{V}_t(\theta_{t-1}, \mathfrak{m}_t)$ belongs to $B_{\mathcal{V}_t}$ provided the conclusion holds at time $t-1$.

Then from Corollary 5.7 applied to $(\theta_{t-1}, \eta_t v_t)$ and $(\theta^*, 0)$ we find

$$\begin{aligned} d(\theta_t, \theta^*) &= d(\Phi_t(\theta_{t-1}, \eta_t v_t), \theta^*) \\ &\leq d(\theta_{t-1}, \theta^*) + \mathrm{M}\, \eta_t \, m(t) \\ &\leq \frac{r^*_\Theta}{4} + \mathrm{M} \sum_{s=1}^{t} \eta_s \, m(s) \end{aligned}$$

with probability 1, by the induction hypothesis at time $t-1$. For $t \leq T$, by definition of $\overline{\eta}^T$ this is at most $\frac{r^*_\Theta}{4} + \frac{r^*_\Theta}{12} = \frac{r^*_\Theta}{3}$. So θ_t belongs to B^*_Θ with probability 1 and the induction hypothesis holds at time t. □

The proof of the next result, which controls the finite-time divergence between two sequences of parameters, is analogous to the control of the fixed point iterates in the proof of the Cauchy–Lipschitz theorem.

Lemma 5.13 (Parameter updates at first order in η)**.** *There exists a constant $\kappa_6 > 0$ with the following property.*

Let $\theta_0, \theta'_0 \in B^*_\Theta$ with $d(\theta_0, \theta^*)$ and $d(\theta'_0, \theta^*)$ at most $r^*_\Theta/3$, and let (v_t), (v'_t) be two gradient sequences with $v_t, v'_t \in B_{\mathcal{V}_t}$ for all $t \geq 1$. Let (η_t) be a stepsize sequence with $\overline{\eta} \leq \overline{\eta}_{\mathcal{V}}$. Define by induction for $t \geq 1$,

$$\theta_t = \Phi_t(\theta_{t-1}, \eta_t v_t), \qquad \theta'_t = \Phi_t(\theta'_{t-1}, \eta_t v'_t).$$

Then for any $0 \leq t < T_0^{r^*_\Theta}(\boldsymbol{\eta})$,

$$d(\theta_t, \theta'_t) \leq 2d(\theta_0, \theta'_0) + \kappa_6 \sum_{1 \leq s \leq t} \eta_s \, d(v_s, v'_s)$$

and

$$d(\theta_t, \theta'_t) \leq 2d(\theta_0, \theta'_0) + \kappa_6 \sum_{1 \leq s \leq t} \eta_s \, m(s).$$

In particular, taking $v'_t = 0$ and $\theta'_0 = \theta_0$, we have

$$d(\theta_t, \theta_0) \leq \kappa_6 \sum_{1 \leq s \leq t} \eta_s \, m(s).$$

Proof. First, by Lemma 5.9, the trajectories stay in the stable tube for $t < T_0^{r^*_\Theta}$, and so the various bounds and assumptions apply.

The second and third statements follow from the first up to increasing κ_6. Indeed, v_s and v'_s are bounded by $m(s)\, r_{\mathcal{V}}/\overline{\eta}_{\mathcal{V}}$ by Corollary 5.5. So we only have to prove the first statement.

By Corollary 5.5, we have $\|\eta_s v_s\| \leq \eta_s\, m(s)\, r_{\mathcal{V}}/\overline{\eta}_{\mathcal{V}}$ and likewise for v'_s. Let us denote this bound by $\tilde{\eta}_s$, namely

$$\tilde{\eta}_s := \eta_s\, m(s)\, r_{\mathcal{V}}/\overline{\eta}_{\mathcal{V}}.$$

By Assumption 4.17, for $t \geq 1$ we have

$$\begin{aligned} d\left(\theta_t,\, \theta'_t\right) &\leq d\left(\theta_{t-1},\, \theta'_{t-1}\right) + c_\Phi\left(d\left(\eta_t v_t,\, \eta_t v'_t\right) + 2\tilde{\eta}_t\, d\left(\theta_{t-1},\, \theta'_{t-1}\right)\right) \\ &= \left(1 + 2c_\Phi\, \tilde{\eta}_t\right) d\left(\theta_{t-1},\, \theta'_{t-1}\right) + c_\Phi\, \eta_t\, d\left(v_t,\, v'_t\right). \end{aligned}$$

Set $p_{s,t} := \prod_{j=s+1}^{t}(1 + 2c_\Phi\, \tilde{\eta}_j)$. By induction we obtain

$$d\left(\theta_t,\, \theta'_t\right) \leq p_{0,t}\, d\left(\theta_0,\, \theta'_0\right) + \sum_{s=1}^{t} p_{s,t}\, c_\Phi\, \eta_s\, d\left(v_s,\, v'_s\right)$$

and the conclusion will follow if we prove that the various factors $p_{s,t}$ are bounded.

Since $1 + 2c_\Phi\, \tilde{\eta}_j \leq \exp(2c_\Phi\, \tilde{\eta}_j)$ we have $p_{s,t} \leq \exp\left(\sum_{j=1}^{t} 2c_\Phi\, \tilde{\eta}_j\right)$. But by definition of $T_0^{r^*_\Theta}$, for $t < T_0^{r^*_\Theta}$ we have

$$\sum_{j=1}^{t} 2c_\Phi\, \tilde{\eta}_j = \frac{2c_\Phi\, r_{\mathcal{V}}}{\overline{\eta}_{\mathcal{V}}} \sum_{j=1}^{t} \eta_j\, m(j) \leq \frac{2c_\Phi\, r_{\mathcal{V}}\, r^*_\Theta}{3\overline{\eta}_{\mathcal{V}}\, \mathrm{M}}. \tag{14}$$

The value of M from Corollary 5.7 satisfies $\mathrm{M} \geq 4c_\Phi r_{\mathcal{V}} r^*_\Theta/\overline{\eta}_{\mathcal{V}}$, so the right-hand side of Equation (14) is bounded by $1/6$. (This happens precisely because we have taken $T_0^{r^*_\Theta}$ small enough to avoid exponential divergence of trajectories in time $t < T_0^{r^*_\Theta}$.)

Therefore, for $t < T_0^{r^*_\Theta}$ we have $p_{s,t} \leq \exp(1/6) \leq 2$. This ends the proof. □

5.1.3 Forgetting of Initial Conditions

Here, we investigate the consequences of Assumption 4.13 and Assumption 4.14.

Corollary 5.14 (Exponential forgetting of instantaneous gradients)**.** *Let θ be a parameter in B^*_Θ, and let $\mathfrak{m}_0, \mathfrak{m}'_0 \in \mathbb{T}_{\mathcal{M}_0}$. Let $(\mathfrak{m}_t)$ and $(\mathfrak{m}'_t)$ be the trajectories associated with θ starting at $\mathfrak{m}_0$ and $\mathfrak{m}'_0$, respectively. Then for all $t \geq 0$,*

$$d\left(\boldsymbol{V}_t\left(\theta,\, \mathfrak{m}_t\right),\, \boldsymbol{V}_t\left(\theta,\, \mathfrak{m}'_t\right)\right) \leq \kappa_1\kappa_5\, m_{\mathrm{H}}(t)\, (1-\alpha)^t\, d\left(\mathfrak{m}_0,\, \mathfrak{m}'_0\right).$$

Proof. This is a direct consequence of Assumption 4.14 and Assumption 4.16. □

Lemma 5.15 (Lipschitz continuity of trajectories)**.** *For any $\bar{\theta} \in B^*_\Theta$ and any sequence of parameters (θ_t) included in B^*_Θ, for any initialization $\mathfrak{m}_0 \in \mathbb{T}_{\mathcal{M}_0}$, the trajectories $(\bar{\mathfrak{m}}_t)$ and $(\mathfrak{m}_t)$, starting at $\mathfrak{m}_0$ with parameters $\bar{\theta}$ and (θ_t) respectively, satisfy*

$$d\left(\bar{\mathfrak{m}}_t,\, \mathfrak{m}_t\right) \leq \frac{\kappa_1\kappa_{\mathrm{lip}\theta}}{\alpha} \sup_{s \leq t-1} d\left(\bar{\theta},\, \theta_s\right),$$

for all $t \geq 0$.

Proof. Let us define a family of trajectories that interpolate between $(\bar{\mathfrak{m}}_t)$ and $(\mathfrak{m}_t)$, by using parameters (θ_t) for the first t_c steps, then parameter $\bar{\theta}$. More precisely, given $t_c \geq 0$, define

$$\mathfrak{m}_t^{t_c} = \begin{cases} \mathcal{A}_t(\theta_{t-1}, \mathfrak{m}_{t-1}^{t_c}) & \text{if } t \leq t_c, \\ \mathcal{A}_t(\bar{\theta}, \mathfrak{m}_{t-1}^{t_c}) & \text{otherwise} \end{cases}$$

so that $\bar{\mathfrak{m}}_t = \mathfrak{m}_t^0$ and $\mathfrak{m}_t = \mathfrak{m}_t^t$. These trajectories lie in $\mathbb{T}_{\mathcal{M}_t}$.

Now

$$d\left(\bar{\mathfrak{m}}_t,\, \mathfrak{m}_t\right) = d\left(\mathfrak{m}_t^0,\, \mathfrak{m}_t^t\right) \leq \sum_{s=0}^{t-1} d\left(\mathfrak{m}_t^s,\, \mathfrak{m}_t^{s+1}\right).$$

Up to time $t = s$, both $(\mathfrak{m}_t^s)$ and $(\mathfrak{m}_t^{s+1})$ use parameter θ_t, therefore $\mathfrak{m}_s^s = \mathfrak{m}_s^{s+1} = \mathfrak{m}_s$. But at time $t = s+1$ they separate:

$$\mathfrak{m}_{s+1}^s = \mathcal{A}_{s+1}(\bar{\theta}, \mathfrak{m}_s^s) = \mathcal{A}_{s+1}(\bar{\theta}, \mathfrak{m}_s)$$

while

$$\mathfrak{m}_{s+1}^{s+1} = \mathcal{A}_{s+1}(\theta_s, \mathfrak{m}_s^{s+1}) = \mathcal{A}_{s+1}(\theta_s, \mathfrak{m}_s).$$

Consequently, by Assumption 4.13,

$$d\left(\mathfrak{m}_{s+1}^s,\, \mathfrak{m}_{s+1}^{s+1}\right) \leq \kappa_{\mathrm{lip}\theta}\, d\left(\theta_s,\, \bar{\theta}\right).$$

Now, from time $s+2$ onwards, both $(\mathfrak{m}_t^s)$ and $(\mathfrak{m}_t^{s+1})$ use parameter $\bar{\theta}$. Therefore for $t \geq s+2$,

$$d\left(\mathfrak{m}_t^s,\, \mathfrak{m}_t^{s+1}\right) \leq \kappa_1 (1-\alpha)^{t-(s+1)} d\left(\mathfrak{m}_{s+1}^s,\, \mathfrak{m}_{s+1}^{s+1}\right)$$

thanks to Assumption 4.14.

Summing, we find

$$d\left(\bar{\mathfrak{m}}_t,\, \mathfrak{m}_t\right) \leq \kappa_1 \kappa_{\mathrm{lip}\theta} \sum_{s=0}^{t-1} (1-\alpha)^{t-(s+1)} d\left(\theta_s,\, \bar{\theta}\right) \leq \frac{\kappa_1 \kappa_{\mathrm{lip}\theta}}{\alpha} \sup_{0 \leq s \leq t-1} d\left(\theta_s,\, \bar{\theta}\right).$$

□

Corollary 5.16 (Continuity of instantaneous gradients)**.** *Let (θ_t) be a sequence of parameters included in B_Θ^*, and let $\bar{\theta} \in B_\Theta^*$. Let $\mathfrak{m}_0 \in \mathbb{T}_{\mathcal{M}_0}$. Let $(\mathfrak{m}_t)$ and $(\bar{\mathfrak{m}}_t)$ be the trajectories starting at $\mathfrak{m}_0$ with parameters (θ_t) and $\bar{\theta}$, respectively. Then for any $t \geq 1$,*

$$d\left(\boldsymbol{V}_t\left(\theta_{t-1},\, \mathfrak{m}_t\right),\, \boldsymbol{V}_t\left(\bar{\theta},\, \bar{\mathfrak{m}}_t\right)\right) = O\left(m_{\mathrm{H}}(t) \sup_{0 \leq s < t} d\left(\theta_s,\, \bar{\theta}\right)\right).$$

Proof. This is a consequence of Assumption 4.16 and Lemma 5.15. □

We now see the finite-time divergence between two trajectories initiated at the same parameter is controlled by the step-size sequence.

Lemma 5.17 (Trajectories for a fixed parameter and different initializations)**.** *Let $\mathfrak{m}_0, \mathfrak{m}_0' \in \mathbb{T}_{\mathcal{M}_0}$ and let $\theta_0 \in B_\Theta^*$ with $d(\theta_0,\, \theta^*) \leq \frac{r_\Theta^*}{3}$. Let $\boldsymbol{\eta} = (\eta_s)_{s \geq 1}$ be a stepsize sequence with $\overline{\eta} \leq \overline{\eta}_{\mathcal{V}}$.*

Then for $t < T_0^{r_\Theta^*}$,

$$d\left(\Phi_{0:t}(\theta_0, \mathfrak{m}_0, (\eta_s)),\, \Phi_{0:t}(\theta_0, \mathfrak{m}_0', (\eta_s))\right) = O\left(m_{\mathrm{H}}(t) \sup_{1 \leq s \leq t} \eta_s\right).$$

Proof. By the definition of the open-loop updates $\Phi_{0:t}$, the distance above is $d(\theta_t,\, \theta_t')$ where we define by induction

$$\begin{cases} \mathfrak{m}_t = \mathcal{A}_t\left(\theta_0,\, \mathfrak{m}_{t-1}\right) \\ v_t = \boldsymbol{V}_t\left(\theta_0,\, \mathfrak{m}_t\right) \\ \theta_t = \Phi_t\left(\theta_{t-1},\, \eta_t\, v_t\right), \end{cases}$$

and likewise with initialization $\mathfrak{m}'_0$

$$\begin{cases} \mathfrak{m}'_t = \mathcal{A}_t\left(\theta_0,\, \mathfrak{m}'_{t-1}\right) \\ v'_t = \boldsymbol{V}_t\left(\theta_0,\, \mathfrak{m}'_t\right) \\ \theta'_t = \Phi_t\left(\theta'_{t-1},\, \eta_t\, v'_t\right), \qquad \theta'_0 = \theta_0. \end{cases}$$

By Assumption 4.15 and Corollary 5.6, for any $t \geq 0$, $\mathfrak{m}_t$ and $\mathfrak{m}'_t$ belong to $\mathbb{T}_{\mathcal{M}_t}$ and v_t and v'_t to $B_{\mathcal{V}_t}$.

Thus, from Lemma 5.13 with $\theta_0 = \theta'_0$, for $t < T_0^{r^*_\Theta}$ we have

$$d\left(\theta_t,\, \theta'_t\right) = O\left(\sum_{1 \leq s \leq t} \eta_s\, d\left(v_s,\, v'_s\right)\right).$$

But thanks to Corollary 5.14, for any $s \leq t$,

$$d\left(v_s,\, v'_s\right) = O\left(m_{\mathrm{H}}(s)\,(1-\alpha)^s\; d\left(\mathfrak{m}_0,\, \mathfrak{m}'_0\right)\right)$$

and therefore, for $t < T_0^{r^*_\Theta}$,

$$\begin{aligned} \sum_{1 \leq s \leq t} \eta_s\, d\left(v_s,\, v'_s\right) &= O\left(\sum_{1 \leq s \leq t} \eta_s\, m_{\mathrm{H}}(s)\,(1-\alpha)^s\, d\left(\mathfrak{m}_0,\, \mathfrak{m}'_0\right)\right) \\ &= O\left(\frac{d\left(\mathfrak{m}_0,\, \mathfrak{m}'_0\right)}{\alpha}\, m_{\mathrm{H}}(t) \sup_{1 \leq s \leq t} \eta_s\right). \end{aligned}$$

Since $\mathbb{T}_{\mathcal{M}_0}$ has a finite diameter and α is fixed, the conclusion follows. □

5.2 Timescales and step sizes

Here we gather some properties that follow from Assumption 4.19 on step sizes and the various scale functions involved.

5.2.1 Sums over intervals $(T; T + L(T)]$

Corollary 5.18 (Sums of stepsizes on an interval are negligible)**.** *Let $T \geq 0$ and let I be the integer interval $I = (T; T + L(T)]$. Under Assumption 4.19:*

1. $\sum_I \eta_t \sim \eta_T\, L(T)$ *when* $T \to \infty$*, and this tends to* 0.
2. $\sum_I \eta_t\, m(t) \sim \eta_T\, L(T)\, m(T)$ *when* $T \to \infty$*, and this tends to* 0.
3. $\sum_I \eta_t\, m_{\mathrm{H}}(t) \sim \eta_T\, L(T)\, m_{\mathrm{H}}(T)$ *when* $T \to \infty$*, and this tends to* 0.
4. $\left(\sum_I \eta_t\, m(t)\right)\left(\sum_I \eta_t\, m_{\mathrm{H}}(t)\right) = o\left(\eta_T\, L(T)\right)$.
5. $\left(\sum_I \eta_t\, m(t)\right)\left(\sum_I \eta_t\, m_{\mathrm{H}}(t)\right) \sim \eta_T^2\, L(T)^2\, m(T)\, m_{\mathrm{H}}(T)$ *when* $T \to \infty$.
6. $\sum_I \eta_t^2\, m(t)^2 \sim \eta_T^2\, m(T)^2\, L(T)$*, when* $T \to \infty$.
7. $\left(\sup_I m_{\mathrm{H}}(t)\right)\left(\sup_I \eta_t\right) = o\left(\eta_T\, L(T)\right)$.
8. *When* $T \to \infty$*,*

$$\frac{\sup_{T < s \leq T + L(T)} \eta_s}{\inf_{T < s \leq T + L(T)} \eta_s} = 1 + o(1/m(T)) = 1 + o(1/m(T + L(T))).$$

9. $T + L(T) \sim T$.

Proof. First, by assumption $L(T) \ll T$, so that $T + L(T) \sim T$.

By the sup/inf assumption in Assumption 4.19, we have $\eta_t \sim \eta_T$ for $t \in I$, so that $\sum_I \eta_t \sim \eta_T\, L(T)$.

Likewise, since $T + L(T) \sim T$ and since $m()$ is a scale function, we have $m(t) \sim m(T)$ for $t \in I$, so that $\sum_I \eta_t\, m(t) \sim \eta_T\, L(T)\, m(T)$. The argument is the same with $m_{\mathrm{H}}(t)$, and with $\sum_I \eta_t^2\, m(t)^2$.

These quantities all tend to 0 by Assumption 4.19.

We have $(\sum_I \eta_t\, m(t))\,(\sum_I \eta_t\, m_{\mathrm{H}}(t)) \sim \eta_T^2\, L(T)^2\, m(T)\, m_{\mathrm{H}}(T)$ by the above. Since $\eta_t L(t) m(t) m_{\mathrm{H}}(t)$ tends to 0 by Assumption 4.19, this is $o\,(\eta_T L(T))$.

Since $m_{\mathrm{H}}(t)$ is a scale function, we have $\sup_I m_{\mathrm{H}}(t) \sim m_{\mathrm{H}}(T)$. By the sup/inf assumption in Assumption 4.19, we have $m_{\mathrm{H}}(T) \sup_I \eta_t \sim m_{\mathrm{H}}(T)\, \eta_T$, which is $o\,(\eta_T\, L(T))$ by Assumption 4.19.

The sup / inf property follows directly from Assumption 4.19 and from $m(t) \sim m(T)$ for $t \in I$. □

Remember that the sequence T_k is defined by $T_{k+1} = T_k + L(T_k)$ (Definition 4.21).

Remark 5.19. *Since the η_t's are nonnegative, and their series diverges according to Assumption 4.19, the first point of Corollary 5.18 implies that the series $\eta_{T_k}\, L(T_k)$ diverges as well.*

Corollary 5.20 (Smallest safe interval k)**.** *There exists an integer $k_0 \geq 1$ such that, for any $\overline{\eta} \leq \overline{\eta}_{\mathcal{V}}$, for any $k \geq k_0$, one has $\sum_{(T_k;T_{k+1}]} \eta_t\, m(t) \leq \frac{r^*_\Theta}{3\mathrm{M}}$ and $\lambda_{\min} \sum_{(T_k;T_{k+1}]} \eta_t \leq 1$. (M is defined in Cor. 5.7.)*

Proof. Using Corollary 5.18, take k_0 such that this holds for $\overline{\eta} = \overline{\eta}_{\mathcal{V}}$. Then the same will hold for smaller $\overline{\eta}$. □

The next lemma justifies the construction of the timescale T_k.

Lemma 5.21. *For any $\overline{\eta} \leq \overline{\eta}_{\mathcal{V}}$, for any $k \geq k_0$, the control time $T_{T_k}^{r^*_\Theta}(\boldsymbol{\eta})$ is (strictly) larger than T_{k+1}.*

Proof. This follows from the Definition 5.8 of $T_t^{r^*_\Theta}$, and from Corollary 5.20. □

We now prove a slight technical strengthening of the sup/inf property on η_t, involving intervals $[T;T+L(T)]$ instead of $(T;T+L(T)]$.

Lemma 5.22. *When $T \to \infty$,*

$$\frac{\sup_{T\leq t\leq T+L(T)} \eta_t}{\inf_{T\leq t\leq T+L(T)} \eta_t} = 1 + o(1/m(T)) = 1 + o(1/m(T + L(T)))$$

and moreover for $T < t \leq T + L(T)$ we have

$$\frac{\eta_T}{\eta_t} = 1 + o(1/m(T)).$$

Proof. The last statement follows from the first by specializing to η_T in the supremum.

For the first statement, write

$$\sup_{T\leq s\leq T+L(T)} \eta_s = \sup\left(\eta_T, \sup_{T<s\leq T+L(T)} \eta_s\right)$$

and likewise for the infimum. By Assumption 4.19 applied to time $t = T-1$, one has $\eta_T \leq \left(1 + o\left(\frac{1}{m(T-1)}\right)\right)\eta_{T+1}$ so that

$$\sup_{T\leq s\leq T+L(T)} \eta_s \leq \left(1 + o\left(\frac{1}{m(T-1)}\right)\right) \sup_{T<s\leq T+L(T)} \eta_s$$

and likewise for the infimum. Thus,

$$\frac{\sup_{T\leq s\leq T+L(T)}\ \eta_s}{\inf_{T\leq s\leq T+L(T)}\ \eta_s} \leq \left(1 + o\left(\frac{1}{m(T-1)}\right)\right)^2 \frac{\sup_{T<s\leq T+L(T)}\ \eta_s}{\inf_{T<s\leq T+L(T)}\ \eta_s}$$

and we can now apply Assumption 4.19 to the rightmost term, yielding

$$\frac{\sup_{T\leq s\leq T+L(T)}\ \eta_s}{\inf_{T\leq s\leq T+L(T)}\ \eta_s} \leq \left(1 + o\left(\frac{1}{m(T-1)}\right)\right)^2 \left(1 + o\left(\frac{1}{m(T)}\right)\right).$$

Now, since $m()$ is a scale function (or 1), we have $m(T-1) \sim m(T)$ when $T \to \infty$, so the above is $(1 + o(1/m(T)))^3$ which is just $1 + o(1/m(T))$.

Finally, as seen above, $T + L(T) \sim T$ so that $m(T + L(T)) \sim m(T)$ as $m()$ is a scale function. □

5.2.2 Constant Stepsizes vs a Sequence of Stepsizes

We now bound the difference between updates $\Phi_{T_k:T_{k+1}}\left(\theta, (\eta_t\, v_t)_{T_k<t\leq T_{k+1}}\right)$ using a variable learning rate η_t, and using the constant learning rate η_{T_k}. This is a consequence of the homogeneity of learning rates on intervals $(T; T + L(T)]$.

Lemma 5.23 (Variable vs constant stepsizes)**.** *Let $\boldsymbol{\eta}$ be a sequence of stepsizes with $\overline{\eta} \leq \overline{\eta}_{\mathcal{V}}/2$. Let L be a scale function such that for T large enough, $T + L(T) < T_T^{r^*_\Theta}(\boldsymbol{\eta})$.*

Let (v_t) be a sequence of gradients with $v_t \in B_{\mathcal{V}_t}$ for all t. Let (θ_t) be a sequence of parameters with $d(\theta_t, \theta^*) \leq \frac{r^*_\Theta}{3}$. Then

$$d\left(\Phi_{T:T+L(T)}(\theta_T, (\eta_t\, v_t)),\ \Phi_{T:T+L(T)}(\theta_T, \eta_T\,(v_t))\right) = o\left(\sum_{T<t\leq T+L(T)} \eta_t\right)$$

when $T \to \infty$.

In particular, letting (v_t) be the sequence of gradients computed along a trajectory $(\mathfrak{m}_T)$ with $\mathfrak{m}_T \in \mathbb{T}_{\mathcal{M}_T}$, we find

$$d\left(\Phi_{T:T+L(T)}(\theta_T,\, \mathfrak{m}_T, (\eta_t)),\ \Phi_{T:T+L(T)}(\theta_T,\, \mathfrak{m}_T,\, \eta_T)\right) = o\left(\sum_{T<t\leq T+L(T)} \eta_t\right).$$

Proof. Let $T < t \leq T + L(T)$. Define v'_t such that

$$\eta_t\, v'_t = \eta_T\, v_t,$$

so that

$$\Phi_{T:T+L(T)}\left(\theta_T, (\eta_T\, v_t)_{T<t\leq T+L(T)}\right) = \Phi_{T:T+L(T)}\left(\theta_T, (\eta_t\, v'_t)_{T<t\leq T+L(T)}\right).$$

By Lemma 5.22, we have $\eta_T/\eta_t = 1 + o(1)$. Therefore, for t large enough, we have $\|v'_t\| \le 2\|v_t\|$ so that if $\overline{\eta} < \overline{\eta}_{\mathcal{V}}/2$, then $\eta_t v'_t$ lies in the control ball $B_{\mathcal{V}_t}$ thanks to Corollary 5.5.

By Lemma 5.13 the distance we want to bound is at most

$$\kappa_6 \sum_{T<t\le T+L(T)} \eta_s \, d(v_s, v'_s)$$

but then $d(v_s, v'_s) = d\left(v_s, \frac{\eta_T}{\eta_s} v_s\right) = \|v_s\| \left|\frac{\eta_T}{\eta_s} - 1\right| = o(1)$ since $v_s = O(m(s))$ and $\frac{\eta_T}{\eta_s} = 1 + o(1/m(T + L(T)))$ with $m(T + L(T)) \ge m(s)$. □

5.3 Finite-Time Divergence Between Trajectories

In this section we consider increasingly easier-to-analyze trajectories. We start with some parameters θ_t computed along a "noisy" trajectory where the states $(\mathfrak{m}_t)$ are not necessarily given by applying the algorithm $\mathcal{A}_t$. We then consider the "regularized" trajectory $\bar{\mathfrak{m}}_t = \mathcal{A}_t(\theta_{t-1}, \bar{\mathfrak{m}}_{t-1})$ defined by $\mathcal{A}_t$, but still using the parameters from the noisy trajectory, and the parameter updates $\bar{\theta}_t$ computed from $\bar{\mathfrak{m}}_t$. These differ by the deviation $D_{0:t}(\theta_0, (\mathfrak{m}_t), \boldsymbol{\eta})$ from Definition 4.24.

Next we consider the "open-loop" trajectory $\mathfrak{m}'_t = \mathcal{A}_t(\theta_0, \mathfrak{m}'_{t-1})$ and the resulting parameter updates θ'_t computed from $\mathfrak{m}'_t$. This open-loop trajectory can be compared to the trajectory with optimal parameter θ^*.

5.3.1 Divergence Between Open-Loop and Closed-Loop Trajectories

Lemma 5.24 (Noisy closed-loop vs open-loop divergence)**.** *Let $\theta_0 \in B^*_\Theta$ with $d(\theta_0, \theta^*) \le r^*_\Theta/3$. Let $(\mathfrak{m}_t)$ be any sequence of states such that $\mathfrak{m}_t \in \mathbb{T}_{\mathcal{M}_t}$. Let $\boldsymbol{\eta} = (\eta_t)_{t\ge1}$ be a stepsize sequence with $\overline{\eta} \le \overline{\eta}_{\mathcal{V}}$.*

Define the "closed-loop" trajectory by induction for $t \ge 1$

$$\begin{cases} v_t = \boldsymbol{V}_t(\theta_{t-1}, \mathfrak{m}_t) \\ \theta_t = \Phi_t(\theta_{t-1}, \eta_t v_t), \end{cases}$$

and let $\theta'_t := \Phi_{0:t}(\theta_0, \mathfrak{m}_0, (\eta_t))$ be the corresponding open-loop value with parameter θ_0, namely, $\mathfrak{m}'_0 = \mathfrak{m}_0$, $\theta'_0 = \theta_0$, and for $t \ge 1$,

$$\begin{cases} \mathfrak{m}'_t = \mathcal{A}_t(\theta_0, \mathfrak{m}'_{t-1}) \\ v'_t = \boldsymbol{V}_t(\theta_0, \mathfrak{m}'_t) \\ \theta'_t = \Phi_t(\theta'_{t-1}, \eta_t v'_t). \end{cases}$$

Then for all $0 \le t < T_0^{r^*_\Theta}(\boldsymbol{\eta})$,

$$d(\theta_t, \theta'_t) = O\left(\left(\sum_{1\le s\le t} \eta_s \, m_{\mathrm{H}}(s)\right)\left(\sum_{1\le s\le t} \eta_s \, m(s)\right)\right) + D_{0:t}(\theta_0, (\mathfrak{m}_s), \boldsymbol{\eta}).$$

In particular, if $(\mathfrak{m}_t)$ itself follows the trajectory $\mathfrak{m}_t = \mathcal{A}_t(\theta_{t-1}, \mathfrak{m}_{t-1})$, we find

$$d(\theta_t, \theta'_t) = O\left(\left(\sum_{1\le s\le t} \eta_s \, m(s)\right)\left(\sum_{1\le s\le t} \eta_s \, m_{\mathrm{H}}(s)\right)\right),$$

as $D_{0:t}(\theta_0, (\mathfrak{m}_s), \boldsymbol{\eta})$ is 0 by definition.

Proof. We first consider the "regularized" trajectory lying between the other two. Define the following trajectory by induction initialized with $\bar{\mathfrak{m}}_0 = \mathfrak{m}_0$, $\bar{\theta}_0 = \theta_0$, and

$$\begin{cases} \bar{\mathfrak{m}}_t = \mathcal{A}_t\left(\theta_{t-1},\, \bar{\mathfrak{m}}_{t-1}\right) \\ \bar{v}_t = \boldsymbol{V}_t\left(\theta_{t-1},\, \bar{\mathfrak{m}}_t\right) \\ \bar{\theta}_t = \Phi_t\left(\bar{\theta}_{t-1},\, \eta_t\, \bar{v}_t\right). \end{cases}$$

By Definition 4.24, for any $t \geq 0$,

$$d\left(\theta_t,\, \bar{\theta}_t\right) = D_{0:t}\left(\theta_0,\, (\mathfrak{m}_s),\, \boldsymbol{\eta}\right).$$

Note that for all three trajectories, for $t < T_0^{r_\Theta^*}$, by Lemma 5.9, all objects at time t belong respectively to B_Θ^*, $B_{\mathcal{V}_t}$, and $\mathbb{T}_{\mathcal{M}_t}$.

We now study the divergence $d\left(\bar{\theta}_t,\, \theta'_t\right)$ between the regularized trajectory and the open-loop trajectory.

Since $\bar{\theta}_0 = \theta'_0$, from Lemma 5.13, we have for $t < T_0^{r_\Theta^*}$

$$d\left(\bar{\theta}_t,\, \theta'_t\right) \leq O\left(\sum_{1 \leq s \leq t} \eta_s\, d\left(\bar{v}_s,\, v'_s\right)\right).$$

Now $\bar{v}_s$ is computed from the trajectory with parameters (θ_s) and v'_s with constant parameter θ_0, so by Corollary 5.16, we have

$$d\left(\bar{v}_s,\, v'_s\right) = O\left(m_{\mathrm{H}}(s) \sup_{p<s} d\left(\theta_p,\, \theta_0\right)\right).$$

But for $0 \leq p < T_0^{r_\Theta^*}$, by Lemma 5.13 we have

$$d\left(\theta_p,\, \theta_0\right) = O\left(\sum_{p' \leq p} \eta_{p'}\, m(p')\right),$$

and therefore, for $s \geq 2$, we have

$$d\left(\bar{v}_s,\, v'_s\right) = O\left(m_{\mathrm{H}}(s) \sum_{p \leq s-1} \eta_p\, m(p)\right) = O\left(m_{\mathrm{H}}(s) \sum_{1 \leq p \leq t} \eta_p\, m(p)\right).$$

The bound still holds for $s = 1$, since $d\left(\bar{v}_1,\, v'_1\right) = 0$, as they are both computed from θ_0 and $\mathfrak{m}_0$. Therefore, for $t < T_0^{r_\Theta^*}$,

$$\begin{aligned} d\left(\bar{\theta}_t,\, \theta'_t\right) &= O\left(\sum_{1 \leq s \leq t} \eta_s\, m_{\mathrm{H}}(s) \left(\sum_{1 \leq p \leq t} \eta_p\, m(p)\right)\right) \\ &= O\left(\left(\sum_{1 \leq s \leq t} \eta_s\, m_{\mathrm{H}}(s)\right) \left(\sum_{1 \leq s \leq t} \eta_s\, m(s)\right)\right), \end{aligned}$$

from which the conclusion follows. ∎

5.3.2 Deviation from the Optimal Parameter in Finite Time

Lemma 5.25 (Deviation from the optimal parameter in finite time)**.** *Let $\theta_0 \in B^*_\Theta$ with $d(\theta_0, \theta^*) \le r^*_\Theta/3$, and let $\mathfrak{m}_0 \in \mathbb{T}_{\mathcal{M}_0}$. Let $\boldsymbol{\eta} = (\eta_t)_{t\ge1}$ be a stepsize sequence with $\overline{\eta} \le \overline{\eta}_{\mathcal{V}}$.*

Consider a trajectory such that for $t \ge 1$

$$\begin{cases} \mathfrak{m}_t = \mathcal{A}_t(\theta_{t-1}, \mathfrak{m}_{t-1}) & \text{or } \mathfrak{m}_t \in \mathbb{T}_{\mathcal{M}_t}, \\ v_t = \boldsymbol{V}_t(\theta_{t-1}, \mathfrak{m}_t), \\ \theta_t = \Phi_t(\theta_{t-1}, \eta_t v_t). \end{cases}$$

Then for any $0 \le t < T_0^{r^*_\Theta}$,

$$\begin{aligned} d(\theta_t, \theta^*) \le{}& d\left(\Phi_{0:t}(\theta_0, \mathfrak{m}_0^*, (\eta_s)),\ \Phi_{0:t}(\theta^*, \mathfrak{m}_0^*, (\eta_s))\right) + d\left(\Phi_{0:t}(\theta^*, \mathfrak{m}_0^*, (\eta_s)), \theta^*\right) \\ &+ D_{0:t}(\theta_0, (\mathfrak{m}_s), \boldsymbol{\eta}) + O\left(m_{\mathrm{H}}(t) \sup_{1\le s\le t} \eta_s\right) \\ &+ O\left(\left(\sum_{1\le s\le t} \eta_s\, m(s)\right)\left(\sum_{1\le s\le t} \eta_s\, m_{\mathrm{H}}(s)\right)\right). \end{aligned}$$

Proof. Let us consider the open-loop trajectory initialized with θ_0 and $\mathfrak{m}_0$, namely

$$\theta'_t := \Phi_{0:t}(\theta_0, \mathfrak{m}_0, (\eta_s)).$$

By Lemma 5.17, for $0 \le t < T_0^{r^*_\Theta}$, we have

$$d\left(\theta'_t, \Phi_{0:t}(\theta_0, \mathfrak{m}_0^*, (\eta_s))\right) = O\left(m_{\mathrm{H}}(t) \sup_{1\le s\le t} \eta_s\right).$$

On the other hand, by Lemma 5.24, for any $t < T_0^{r^*_\Theta}$, we have

$$d(\theta_t, \theta'_t) = O\left(\left(\sum_{1\le s\le t} \eta_s\, m(s)\right)\left(\sum_{1\le s\le t} \eta_s\, m_{\mathrm{H}}(s)\right)\right) + D_{0:t}(\theta_0, (\mathfrak{m}_s), \boldsymbol{\eta}),$$

and the conclusion follows by the triangle inequality. □

5.4 Convergence of Learning

5.4.1 Behavior Around the Local Minimum θ^*

Lemma 5.26 (At first order, θ^* is not updated in I_k)**.** *Assume that $\overline{\eta} \le \min(\overline{\eta}_{\mathrm{op}}, \overline{\eta}_{\mathcal{V}}/2)$. Then when $k \to \infty$,*

$$d\left(\Phi_{T_k:T_{k+1}}\left(\theta^*, \mathfrak{m}^*_{T_k}, (\eta_t)\right), \theta^*\right) = o\left(\eta_{T_k} L(T_k)\right).$$

Proof. By Assumption 4.18 and by the Definition 4.21 of T_k, this holds when using a constant learning rate η_{T_k} instead of η_t between T_k and T_{k+1}; namely, we have

$$d\left(\Phi_{T_k:T_{k+1}}\left(\theta^*, \mathfrak{m}^*_{T_k}, \eta_{T_k}\right), \theta^*\right) = o\left(\eta_{T_k} L(T_k)\right).$$

Lemma 5.23 can transfer this to non-constant step sizes η_s instead of η_{T_k}. Let us check that all the assumptions of Lemma 5.23 are satisfied. Remember that

$T_{k+1} = T_k + L(T_k)$. The condition $T_{T_k}^{r_\Theta^*} > T_{k+1}$ is satisfied for $k \geq k_0$ by Lemma 5.21. The condition on step sizes is satisfied by the last point of Corollary 5.18. Therefore, for $k \geq k_0$ we can apply Lemma 5.23 to $T = T_k$. This provides the conclusion, after observing that $\sum_{I_k} \eta_t \sim \eta_{T_k} L(T_k)$ by Corollary 5.18. What happens for $k < k_0$ is absorbed in the $o()$ notation. □

Lemma 5.27 (Contractivity of open-loop updates on each interval)**.** *Assume that* $\overline{\eta} \leq \min(\overline{\eta}_{\mathrm{op}}, \overline{\eta}_{\mathcal{V}}/2)$. *Then for* $k \geq k_0$, *for any* $\theta \in B_\Theta^*$ *with* $d(\theta, \theta^*) \leq \frac{r_\Theta^*}{3}$,

$$d\left(\Phi_{T_k:T_{k+1}}\left(\theta, \mathfrak{m}_{T_k}^*, (\eta_t)\right), \Phi_{T_k:T_{k+1}}\left(\theta^*, \mathfrak{m}_{T_k}^*, (\eta_t)\right)\right)$$

is at most

$$(1 - \lambda_{\min} \eta_{T_k} L(T_k))\ d(\theta, \theta^*) + o(\eta_{T_k} L(T_k)).$$

Proof. Assumption 4.18 applied to the intervals $T_k < t \leq T_k + L(T_k) = T_{k+1}$ provides the same statement but using constant step size η_{T_k} instead of variable step size (η_t).

As in Lemma 5.26, we can use Lemma 5.23 to bound the distance between constant and variable step sizes. This yields

$$d\left(\Phi_{T_k:T_{k+1}}\left(\theta, \mathfrak{m}_{T_k}^*, (\eta_t)\right), \Phi_{T_k:T_{k+1}}\left(\theta, \mathfrak{m}_{T_k}^*, \eta_{T_k}\right)\right) = o\left(\sum_{I_k} \eta_t\right) = o(\eta_{T_k} L(T_k))$$

and likewise for θ^*. The conclusion follows by the triangle inequality. □

5.4.2 Contraction of Errors from T_k to T_{k+1}

Lemma 5.28 (Contraction of errors from T_k to T_{k+1})**.** *Let* $\overline{\eta} \leq \min(\overline{\eta}_{\mathcal{V}}/2, \overline{\eta}_{\mathrm{op}})$. *Let* $k \geq k_0$ *where* k_0 *is defined in Corollary 5.20.*

Let θ_{T_k} *be such that* $d(\theta_{T_k}, \theta^*) \leq \frac{r_\Theta^*}{3}$, *and let* $\mathfrak{m}_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$. *Consider the learning trajectory from initial parameter* θ_{T_k} *and initial state* $\mathfrak{m}_{T_k}$ *and learning rates* (η_t), *namely,*

$$\begin{cases} \mathfrak{m}_t = \mathcal{A}_t(\theta_{t-1}, \mathfrak{m}_{t-1}) & \text{or } \mathfrak{m}_t \in \mathbb{T}_{\mathcal{M}_t} \\ v_t = \boldsymbol{V}_t(\theta_{t-1}, \mathfrak{m}_t) \\ \theta_t = \Phi_t(\theta_{t-1}, \eta_t v_t). \end{cases}$$

Then for all $T_k \leq t \leq T_{k+1}$, *we have* $\theta_t \in B_\Theta^*$ *and* $\mathfrak{m}_t \in \mathbb{T}_{\mathcal{M}_t}$, *and moreover,*

$$\begin{aligned} d\left(\theta_{T_{k+1}}, \theta^*\right) \leq\ & (1 - \lambda_{\min} \eta_{T_k} L(T_k))\ d(\theta_{T_k}, \theta^*) \\ & + D_{T_k:T_{k+1}}(\theta_{T_k}, (\mathfrak{m}_t), \boldsymbol{\eta}) + o(\eta_{T_k} L(T_k)) \end{aligned}$$

where the $o()$ *is uniform over* θ_{T_k}, $\mathfrak{m}_{T_k}$ *and* $\overline{\eta}$ *satisfying the constraints above.*

Proof. From Lemma 5.21 and since $k \geq k_0$, we have $T_{T_k}^{r_\Theta^*} > T_{k+1}$. Therefore we can apply Lemma 5.9 and, for $T_k \leq t \leq T_{k+1}$, we have

$$(\theta_t, \mathfrak{m}_t) \in B_\Theta^* \times \mathbb{T}_{\mathcal{M}_t}.$$

Thus we can apply Lemma 5.25 starting at time T_k, using again that $T_{T_k}^{r_\Theta^*} > T_{k+1}$. This yields, for any $T_k + 1 \le t \le T_{k+1}$,

$$\begin{aligned} d(\theta_t, \theta^*) \le\ & d\left(\Phi_{T_k:t}\left(\theta_{T_k}, \mathfrak{m}^*_{T_k}, (\eta_s)\right), \Phi_{T_k:t}\left(\theta^*_{T_k}, \mathfrak{m}^*_{T_k}, (\eta_s)\right)\right) \\ & + d\left(\Phi_{T_k:t}\left(\theta^*, \mathfrak{m}^*_{T_k}, (\eta_s)\right), \theta^*\right) \\ & + D_{T_k:t}(\theta_{T_k}, (\mathfrak{m}_t), \boldsymbol{\eta}) \\ & + O\left(\left(\sum_{T_k<s\le t} \eta_s\, m(s)\right)\left(\sum_{T_k<s\le t} \eta_s m_{\mathrm{H}}(s)\right)\right) + O\left(m_{\mathrm{H}}(t) \sup_{T_k<s\le t} \eta_s\right). \end{aligned}$$

Taking $t = T_{k+1}$, by Lemma 5.27, the first term is at most $(1 - \lambda_{\min}\, \eta_{T_k}\, L(T_k))\ d(\theta_{T_k}, \theta^*) + o(\eta_{T_k}\, L(T_k))$.

By Lemma 5.26, the second term is $o(\eta_{T_k}\, L(T_k))$.

By Corollary 5.18, the last two terms are $o(\eta_{T_k}\, L(T_k))$. □

5.4.3 Convergence of the Algorithm

Lemma 5.29. *Let $\boldsymbol{r} = (r_k)$ and $\boldsymbol{b} = (b_k)$ be two non-negative sequences such that*

1. *$r_k \to 0$ and $\sum_k r_k \to \infty$;*
2. *$b_k = o(r_k)$ when $k \to \infty$.*

Let (x_k) be any non-negative sequence such that for $k \ge k_0$,

$$x_{k+1} \le (1 - r_k)\, x_k + b_k.$$

Then $x_k \to 0$.

Proof. Let us prove that $x_k \to 0$. Let $\varepsilon > 0$ and let us prove that ultimately, $x_k \le 2\varepsilon$.

Set $K := \inf\{k \ge k_0 \,|\, \forall k' \ge k,\ b_{k'} \le \varepsilon\, r_{k'}\}$. For $k \ge K$, the interval $[0; \varepsilon]$ is stable by the map $x \mapsto (1 - r_k)x + b_k$. Therefore, if there exists $k \ge K$ such that $x_k \le \varepsilon$, then we have $x_{k'} \le \varepsilon$ for all $k' \ge k$.

If there exists no $k \ge K$ such that $x_k \le \varepsilon$, then we have for all $k \ge K$, $0 \le x_{k+1} - \varepsilon \le (1 - r_k)\ x_k + b_k - \varepsilon \le (1 - r_k)\ x_k + \varepsilon\, r_k - \varepsilon = (1 - r_k)\ (x_k - \varepsilon)$. Therefore,

$$0 \le x_k - \varepsilon \le \left(\prod_{k'=K}^{k-1} (1 - r_{k'})\right)(x_K - \varepsilon).$$

Since $\sum_k r_k$ diverges, the product $\prod(1 - r_k)$ tends to 0. Therefore, $x_k - \varepsilon$ is less than ε for large enough k.

Thus in both cases, x_k is ultimately less that 2ε, for any $\varepsilon > 0$. □

Lemma 5.30 (End of proof of Theorem 4.27)**.** *There exists $\overline{\eta}_{\mathrm{conv}} > 0$ such that, for any $0 \le \overline{\eta} \le \overline{\eta}_{\mathrm{conv}}$, the following convergence holds.*

For any θ_0 with $d(\theta_0, \theta^*) \le \frac{r_\Theta^*}{4}$ and any $\mathfrak{m}_0 \in \mathbb{T}_{\mathcal{M}_0}$, consider a trajectory given by

$$\begin{cases} \mathfrak{m}_t = \mathcal{A}_t(\theta_{t-1}, \mathfrak{m}_{t-1}) \\ v_t = \boldsymbol{V}_t(\theta_{t-1}, \mathfrak{m}_t) \\ \theta_t = \Phi_t(\theta_{t-1}, \eta_t\, v_t), \end{cases}$$

for $t \ge 1$. Then θ_t tends to θ^* as $t \to \infty$.

Proof. Take $\overline{\eta} \leq \min(\overline{\eta}_{\text{op}}, \overline{\eta}_{\mathcal{V}}/2)$. (This is not yet $\overline{\eta}_{\text{conv}}$: there will be an additional constraint on $\overline{\eta}$ below.)

By Lemma 5.28, there exists $k_0 \geq 0$, and a sequence $b_k = o\left(\eta_{T_k} L(T_k)\right)$ such that

$$d\left(\theta_{T_{k+1}}, \theta^*\right) \leq \left(1 - \lambda_{\min} \eta_{T_k} L(T_k)\right) d\left(\theta_{T_k}, \theta^*\right) + b_k$$

holds for those values of $k \geq k_0$ such that $d\left(\theta_{T_k}, \theta^*\right) \leq \frac{r^*_\Theta}{3}$ and $\mathfrak{m}_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$. (Note that $D_{T_k:T_{k+1}}\left(\theta_{T_k}, (\mathfrak{m}_t), \boldsymbol{\eta}\right)$ vanishes by definition because, for all $t \geq 1$, we have $\mathfrak{m}_t = \mathcal{A}_t\left(\theta_{t-1}, \mathfrak{m}_{t-1}\right)$.) By Lemma 5.28, the value of b_k is uniform over $\bar{\eta}$ and the values of θ and $\mathfrak{m}$ satisfying those assumptions.

Since b_k is $o\left(\eta_{T_k} L(T_k)\right)$, there exists $k_1 \geq k_0$ such that b_k is less than $(r^*_\Theta/3)(\lambda_{\min} \eta_{T_k} L(T_k))$ for $k \geq k_1$. (Such a k_1 is uniform in the values of θ, $\mathfrak{m}$ and $\bar{\eta}$ satisfying the assumptions above, because b_k is.)

Define $\overline{\eta}_{\text{conv}} := \min\left(\overline{\eta}_{\text{op}}, \overline{\eta}_{\mathcal{V}}/2, \overline{\eta}^{T_{k_1}}\right)$, where $\overline{\eta}^{T_{k_1}}$ is defined by Lemma 5.11 applied to $T = T_{k_1}$.

The assumptions state that $d\left(\theta_0, \theta^*\right) \leq \frac{r^*_\Theta}{4}$ and $\mathfrak{m}_0 \in \mathbb{T}_{\mathcal{M}_0}$. Therefore, by Lemma 5.11 applied to $T = T_{k_1}$, if $\overline{\eta} \leq \overline{\eta}^{T_{k_1}}$ then $d\left(\theta_{T_{k_1}}, \theta^*\right) \leq \frac{r^*_\Theta}{3}$ and $\mathfrak{m}_{T_{k_1}} \in \mathbb{T}_{\mathcal{M}_{T_{k_1}}}$.

Set $r_k := \lambda_{\min} \eta_{T_k} L(T_k)$. We have $b_k = o\left(r_k\right)$.

By Lemma 5.28, if $d\left(\theta_{T_k}, \theta^*\right) \leq \frac{r^*_\Theta}{3}$ and $\mathfrak{m}_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$, then

$$d\left(\theta_{T_{k+1}}, \theta^*\right) \leq \left(1 - r_k\right) d\left(\theta_{T_k}, \theta^*\right) + b_k$$

and $\mathfrak{m}_{T_{k+1}} \in \mathbb{T}_{\mathcal{M}_{T_{k+1}}}$.

By definition of k_1, if $k \geq k_1$ then $(1 - r_k)\frac{r^*_\Theta}{3} + b_k \leq \frac{r^*_\Theta}{3}$.

Consequently, if $k \geq k_1$ and $d\left(\theta_{T_k}, \theta^*\right) \leq \frac{r^*_\Theta}{3}$ and $\mathfrak{m}_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$, then $d\left(\theta_{T_{k+1}}, \theta^*\right) \leq \frac{r^*_\Theta}{3}$ and $\mathfrak{m}_{T_{k+1}} \in \mathbb{T}_{\mathcal{M}_{T_{k+1}}}$.

Since this holds at time T_{k_1}, by induction this holds for any $k \geq k_1$: if $\overline{\eta} \leq \overline{\eta}^{T_{k_1}}$, then $d\left(\theta_{T_k}, \theta^*\right) \leq \frac{r^*_\Theta}{3}$ and $\mathfrak{m}_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$ for all $k \geq k_1$.

Therefore, for any $k \geq k_1$, we have

$$d\left(\theta_{T_{k+1}}, \theta^*\right) \leq \left(1 - r_k\right) d\left(\theta_{T_k}, \theta^*\right) + b_k.$$

By Remark 5.19, the series $r_k = \lambda_{\min} \eta_{T_k} L(T_k)$ diverges. Since $b_k = o\left(r_k\right)$, by Lemma 5.29 this implies that θ_{T_k} tends to θ^* when $k \to \infty$.

For the intermediate times $T_k < t \leq T_{k+1}$, by Lemma 5.13, we have

$$d\left(\theta_t, \theta_{T_k}\right) \leq \kappa_6 \sum_{T_k < s \leq T_{k+1}} \eta_s \, m(s)$$

(we can apply Lemma 5.13 because we stay in the stable tube for $t \geq T_{k_1}$). By Corollary 5.18, this proves that θ_t tends to θ^* if θ_{T_k} does. □

Lemma 5.31 (End of proof of Theorem 4.28)**.** *Assume that, for any θ_0 with $d\left(\theta_0, \theta^*\right) \leq \frac{r^*_\Theta}{4}$ and any $\mathfrak{m}_0 \in \mathbb{T}_{\mathcal{M}_0}$, we are given a random trajectory $(\mathfrak{m}_t, v_t, \theta_t)_{t \geq 0}$ which respects the stable tube $(\mathbb{T}_{\mathcal{M}_t})_{t \geq 0}$, in the sense of Definition 4.23, that is, satisfies*

$$\begin{cases} \mathfrak{m}_{t-1} \in \mathbb{T}_{\mathcal{M}_{t-1}} \text{ and } \theta_{t-1} \in B^*_\Theta \implies \mathfrak{m}_t \in \mathbb{T}_{\mathcal{M}_t} \\ v_t = \boldsymbol{V}_t\left(\theta_{t-1}, \mathfrak{m}_t\right) \\ \theta_t = \Phi_t\left(\theta_{t-1}, \eta_t \, v_t\right), \end{cases}$$

for $t \geq 1$. Let $\varepsilon > 0$. Assume there exists $K \geq 0$, a non-negative sequence (δ_k) which tends towards 0, and $\overline{\eta}_{\text{noise}} > 0$ such that, for all $\bar{\eta} \leq \overline{\eta}_{\text{noise}}$, with probability greater than $1 - \varepsilon$, this trajectory has negligible noise starting at K, at speed (δ_k) (Definition 4.25).

Then there exists $\overline{\eta}_{\text{conv}} > 0$ such that for $\overline{\eta} \leq \overline{\eta}_{\text{conv}}$, with probability greater than $1 - \varepsilon$, the parameter θ_t tends to θ^* as $t \to \infty$.

Proof. Take $\overline{\eta} \leq \min(\overline{\eta}_{\text{noise}}, \overline{\eta}_{\text{op}}, \overline{\eta}_{\mathcal{V}}/2)$. (This is not yet $\overline{\eta}_{\text{conv}}$: there will be an additional constraint on $\overline{\eta}$ below.)

Define $b_k^1 := \delta_k \, \eta_{T_k} \, L(T_k)$, where (δ_k) is the sequence controlling the negligible noise in the assumptions.

Let $k \geq 0$ such that $\mathfrak{m}_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$, $\theta_{T_k} \in B_\Theta^*$, and $d(\theta_{T_k}, \theta^*) \leq r_\Theta^*/3$. Since $\overline{\eta} \leq \overline{\eta}_{\mathcal{V}}$, by Lemma 5.21, for any $k \geq 0$, we have $T_{k+1} < T_{T_k}^{r_\Theta^*}(\boldsymbol{\eta})$. Therefore, by Lemma 5.10, we stay in the stable tube between T_k and T_{k+1}, namely, we have $\mathfrak{m}_t \in \mathbb{T}_{\mathcal{M}_t}$ and $\theta_t \in B_\Theta^*$ for all $T_k + 1 \leq t \leq T_{k+1}$.

Since the random trajectory is in the stable tube for $T_k \leq t \leq T_{k+1}$, we can apply Lemma 5.28. Thus, there exists $k_0 \geq 0$, and a sequence $b_k^2 = o(\eta_{T_k} \, L(T_k))$ such that

$$d\left(\theta_{T_{k+1}}, \theta^*\right) \leq (1 - \lambda_{\min} \, \eta_{T_k} \, L(T_k)) \, d(\theta_{T_k}, \theta^*) + D_{T_k:T_{k+1}}(\theta_{T_k}, (\mathfrak{m}_t), \boldsymbol{\eta}) + b_k^2$$

holds for those values of $k \geq k_0$ such that $d(\theta_{T_k}, \theta^*) \leq \frac{r_\Theta^*}{3}$ and $\mathfrak{m}_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$. By Lemma 5.28, the value of b_k^2 is uniform over $\overline{\eta}$ and the values of θ and $\mathfrak{m}$ satisfying those assumptions.

Set $b_k := b_k^1 + b_k^2$. Since b_k is $o(\eta_{T_k} \, L(T_k))$, there exists $k_1 \geq \max(K, k_0)$ such that b_k is less than $(r_\Theta^*/3)(\lambda_{\min} \, \eta_{T_k} \, L(T_k))$ for $k \geq k_1$. Such a k_1 is uniform over the values of θ, $\mathfrak{m}$ and $\overline{\eta}$ satisfying the assumptions above, because b_k is.

Define $\overline{\eta}_{\text{conv}} := \min\left(\overline{\eta}_{\text{noise}}, \overline{\eta}_{\text{op}}, \overline{\eta}_{\mathcal{V}}/2, \overline{\eta}^{T_{k_1}}\right)$, where $\overline{\eta}^{T_{k_1}}$ is the value provided by Lemma 5.12 applied to $T = T_{k_1}$.

For $k \geq 0$, let $\mathcal{E}_k = \left\{\theta \in \Theta \,\middle|\, d(\theta, \theta^*) \leq \frac{r_\Theta^*}{3}\right\} \times \mathbb{T}_{\mathcal{M}_{T_k}}$. Consider the event that the random trajectory has negligible noise; more precisely, define the event

$$\mathfrak{S}(\bar{\eta}) := \left\{1_{(\theta_{T_k}, \mathfrak{m}_{T_k}) \in \mathcal{E}_k} \, D_{T_k:T_{k+1}}(\theta_{T_k}, (\mathfrak{m}_t), \boldsymbol{\eta}) \leq \delta_k \, \eta_{T_k} \, L(T_k), \quad \forall k \geq K\right\}.$$

Thus, on this event, we have $D_{T_k:T_{k+1}}(\theta_{T_k}, (\mathfrak{m}_t), \boldsymbol{\eta}) \leq b_k^1$ for any $k \geq K$ such that $\theta_{T_k} \in B_\Theta^*$, $d(\theta_{T_k}, \theta^*) \leq r_\Theta^*/3$ and $\mathfrak{m}_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$. By assumption and by definition of negligible noise, for $\overline{\eta} < \overline{\eta}_{\text{noise}}$ this event has probability at least $1 - \varepsilon$. We now assume the trajectory is such that this event holds.

Set $r_k := \lambda_{\min} \, \eta_{T_k} \, L(T_k)$. We have $b_k = o(r_k)$.

By Lemma 5.28, if $d(\theta_{T_k}, \theta^*) \leq \frac{r_\Theta^*}{3}$ and $\mathfrak{m}_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$, then

$$d\left(\theta_{T_{k+1}}, \theta^*\right) \leq (1 - r_k) \, d(\theta_{T_k}, \theta^*) + b_k$$

and $\mathfrak{m}_{T_{k+1}} \in \mathbb{T}_{\mathcal{M}_{T_{k+1}}}$.

By definition of k_1, if $k \geq k_1$ then $(1 - r_k)\frac{r_\Theta^*}{3} + b_k \leq \frac{r_\Theta^*}{3}$.

Consequently, if $k \geq k_1$ and $d(\theta_{T_k}, \theta^*) \leq \frac{r_\Theta^*}{3}$ and $\mathfrak{m}_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$, then $d\left(\theta_{T_{k+1}}, \theta^*\right) \leq \frac{r_\Theta^*}{3}$ and $\mathfrak{m}_{T_{k+1}} \in \mathbb{T}_{\mathcal{M}_{T_{k+1}}}$.

Therefore, by induction, if $d\left(\theta_{T_{k_1}}, \theta^*\right) \leq \frac{r_\Theta^*}{3}$ and $\mathfrak{m}_{T_{k_1}} \in \mathbb{T}_{\mathcal{M}_{T_{k_1}}}$, then this holds for any $k \geq k_1$.

By assumption, the random trajectory respects the stable tube. Moreover, we have assumed that $d(\theta_0, \theta^*) \leq \frac{r_\Theta^*}{4}$ and $\mathfrak{m}_0 \in \mathbb{T}_{\mathcal{M}_0}$. Therefore, we can apply Lemma 5.12: the value $\overline{\eta}^{T_{k_1}}$ provided by Lemma 5.12 (and used in the definition of $\overline{\eta}_{\text{conv}}$ above) is such that, for any $\overline{\eta} \leq \overline{\eta}^{T_{k_1}}$, we have $d\left(\theta_{T_{k_1}}, \theta^*\right) \leq \frac{r_\Theta^*}{3}$ and $\mathfrak{m}_{T_{k_1}} \in \mathbb{T}_{\mathcal{M}_{T_{k_1}}}$. We have defined $\overline{\eta}_{\text{conv}}$ above to be no greater than $\overline{\eta}^{T_{k_1}}$, so the constraint $\overline{\eta} \leq \overline{\eta}^{T_{k_1}}$ is satisfied for any $\overline{\eta} \leq \overline{\eta}_{\text{conv}}$.

Thus, if $\overline{\eta} \leq \overline{\eta}^{T_{k_1}}$, then $d(\theta_{T_k}, \theta^*) \leq \frac{r_\Theta^*}{3}$ and $\mathfrak{m}_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$ for all $k \geq k_1$.

For any $\bar{\eta} \leq \overline{\eta}_{\text{conv}}$, conditionally on the event $\mathfrak{S}(\bar{\eta})$, all of the above applies. Therefore, for any $k \geq k_1$ we have

$$d\left(\theta_{T_{k+1}}, \theta^*\right) \leq (1 - r_k)\, d(\theta_{T_k}, \theta^*) + b_k.$$

Since $b_k = o(r_k)$, by Lemma 5.29 this implies that θ_{T_k} tends to θ^* when $k \to \infty$.

For the intermediate times $T_k < t \leq T_{k+1}$, by Lemma 5.13, we have

$$d(\theta_t, \theta_{T_k}) \leq \kappa_6 \sum_{T_k < s \leq T_{k+1}} \eta_s\, m(s)$$

(we can apply Lemma 5.13 because we stay in the stable tube for $t \geq T_{k_1}$). By Corollary 5.18, this proves that θ_t tends to θ^* if θ_{T_k} does.

Therefore, for each $\bar{\eta} \leq \overline{\eta}_{\text{conv}}$, convergence occurs for each trajectory such that the event $\mathfrak{S}(\bar{\eta})$ holds, which, by assumption, happens with probability greater than $1 - \varepsilon$. We have thus proven our claim. □

5.4.4 Convergence of the Open-Loop Algorithm

We now prove convergence of the open-loop algorithm as used in Theorem 4.29. All results established so far still hold true, except for Lemma 5.28 and Lemma 5.30 (and Lemma 5.31); we now prove the respective analogues of Lemmas 5.28 and 5.30 for the open-loop algorithm, Lemma 5.33 and Lemma 5.34. The proofs are actually simpler, since the open-loop case on intervals $(T_k; T_{k+1}]$ is actually the basis of the analysis of the previous case.

First, Lemma 5.32 deals with the piecewise constant stepsizes of the open-loop algorithm.

Lemma 5.32 (Using piecewise constant step-sizes)**.** *Define a modified stepsize sequence $(\tilde{\eta}_t)$ by setting*

$$\tilde{\eta}_t := \eta_{T_{k+1}}$$

for each $T_k + 1 \leq t \leq T_{k+1}$. Then this new stepsize sequence still satisfies Assumption 4.19.

Consequently, all previous results also apply with this new stepsize sequence.

Thus, for the rest of this section, we assume that the stepsize sequence (η_t) is constant on each time interval $(T_k; T_{k+1}]$.

Proof. Fof the first point of Assumption 4.19, write

$$\sum_{t \geq 0} \tilde{\eta}_t = \sum_{k \geq 0} \sum_{T_k + 1}^{T_{k+1}} \eta_{T_{k+1}} = \sum_{k \geq 0} \eta_{T_{k+1}}\, L(T_k).$$

Now, $\eta_{T_{k+1}} L(T_k) \sim \eta_{T_k} L(T_k) \sim \sum_{t=T_k+1}^{T_{k+1}} \eta_t$ by Corollary 5.18. So $\sum \tilde{\eta}_t$ diverges if and only if $\sum \eta_t$ does. There is nothing to check for the third point of the assumption. Let us check the second point. For every $t \geq 1$, write k_t the unique integer such that $T_{k_t} < t \leq T_{k_t+1}$. Then

$$\tilde{\eta}_t \, L(t) \, m(t) \, m_{\mathrm{H}}(t) = \eta_{T_{k_t+1}} \, L(t) \, m(t) \, m_{\mathrm{H}}(t)$$

Since $k_t \to \infty$, when $t \to \infty$, and since $T_{k_t} < t \leq T_{k_t+1}$ with $T_k \sim T_{k+1}$ when $k \to \infty$ (by Lemma 4.22), we have $T_{k_t+1} \sim t$, when $t \to \infty$. As a result, since L, m and $m_{\mathrm{H}}(\cdot)$ are scale functions, and consequently preserve asymptotic equivalence at infinity, we have $L(t) \, m(t) \, m_{\mathrm{H}}(t) \sim L(T_{k_t+1}) \, m(T_{k_t+1}) \, m_{\mathrm{H}}(T_{k_t+1})$, as $t \to \infty$. Therefore,

$$\tilde{\eta}_t \, L(t) \, m(t) \, m_{\mathrm{H}}(t) \sim \eta_{T_{k_t+1}} \, L(T_{k_t+1}) \, m(T_{k_t+1}) \, m_{\mathrm{H}}(T_{k_t+1}),$$

as $k \to \infty$. Now, since the sequence (η_t) satisfies Assumption 4.19, the right-hand side converges to 0, as $t \to \infty$, so that the sequence $(\tilde{\eta}_t)$ indeeds satisfies the second point of Assumption 4.19.

For the last point, let $t \geq 1$. We want to bound $(\sup_{t<s\leq t+L(t)} \tilde{\eta}_s)/(\inf_{t<s\leq t+L(t)} \tilde{\eta}_s)$. We have

$$1 \leq \frac{\sup_{t<s\leq t+L(t)} \tilde{\eta}_s}{\inf_{t<s\leq t+L(t)} \tilde{\eta}_s} = \frac{\sup_{t<s\leq t+L(t)} \eta_{T_{k_s+1}}}{\inf_{t<s\leq t+L(t)} \eta_{T_{k_s+1}}}.$$

The maps $s \mapsto k_s$ is non-decreasing. Therefore, when s ranges from t to $t + L(t)$, k_s ranges at most from k_t to $k_{t+L(t)}$, so that T_{k_s+1} ranges at most from T_{k_t+1} to $T_{k_{t+L(t)}+1}$. Therefore,

$$\frac{\sup_{t<s\leq t+L(t)} \eta_{T_{k_s+1}}}{\inf_{t<s\leq t+L(t)} \eta_{T_{k_s+1}}} \leq \frac{\sup_{T_{k_t+1}\leq s\leq T_{k_{t+L(t)}+1}} \eta_s}{\inf_{T_{k_t+1}\leq s\leq T_{k_{t+L(t)}+1}} \eta_s}.$$

Next, by definition of $k_{t+L(t)}$, we have $T_{k_{t+L(t)}} < t+L(t)$. Moreover, $T_{k_t} < t \leq T_{k_t+1}$, and L is non-decreasing, so that we have $T_{k_t+1} = T_{k_t} + L(T_{k_t}) < t + L(t) \leq T_{k_t+1} + L(T_{k_t+1}) = T_{k_t+2}$. As a result, $k_{t+L(t)} = k_t + 1$. Therefore,

$$\frac{\sup_{t<s\leq t+L(t)} \eta_{T_{k_s+1}}}{\inf_{t<s\leq t+L(t)} \eta_{T_{k_s+1}}} \leq \frac{\sup_{T_{k_t+1}\leq s\leq T_{k_t+2}} \eta_s}{\inf_{T_{k_t+1}\leq s\leq T_{k_t+2}} \eta_s}.$$

By Lemma 5.22, this is $1 + o\left(1/m(T_{k_t+1})\right)$.

Finally, remember that $T_{k_t+1} \sim t$. So, since $m()$ is a scale function, we have $m(T_{k_t+1}) \sim m(t)$ and $1 + o(1/m(T_{k_t+1})) = 1 + o(1/m(t))$. Thus, we have proven that

$$1 \leq \frac{\sup_{t<s\leq t+L(t)} \tilde{\eta}_s}{\inf_{t<s\leq t+L(t)} \tilde{\eta}_s} \leq 1 + o(1/m(t))$$

namely, $\tilde{\eta}_s$ satisfies the last point of Assumption 4.19. □

Lemma 5.33 (Contraction of errors from T_k to T_{k+1} for the open-loop algorithm)**.** *Let $\overline{\eta} \leq \min(\overline{\eta}_{\mathcal{V}}/2, \overline{\eta}_{\mathrm{op}})$. Let $k \geq k_0$ where k_0 is defined in Corollary 5.20.*

Let θ_{T_k} be such that $d\left(\theta_{T_k}, \theta^*\right) \leq \frac{r^*_\Theta}{3}$, and let $\mathfrak{m}_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$. Consider the learning trajectory from initial parameter θ_{T_k} and initial state $\mathfrak{m}_{T_k}$ and learning rates (η_t), namely,

$$\begin{cases} \mathfrak{m}_t = \mathcal{A}_t\left(\theta_{T_k}, \, \mathfrak{m}_{t-1}\right) \\ v_t = \boldsymbol{V}_t\left(\theta_{T_k}, \, \mathfrak{m}_t\right) \\ \theta_t = \Phi_t\left(\theta_{t-1}, \, \eta_t \, v_t\right). \end{cases}$$

for $T_k < t \le T_{k+1}$. Then for all $T_k \le t \le T_{k+1}$, we have $\theta_t \in B^*_\Theta$ and $\mathfrak{m}_t \in \mathbb{T}_{\mathcal{M}_t}$, and moreover,

$$d\left(\theta_{T_{k+1}}, \theta^*\right) \le (1 - \lambda_{\min}\, \eta_{T_k}\, L(T_k))\; d\left(\theta_{T_k}, \theta^*\right) + o\left(\eta_{T_k}\, L(T_k)\right)$$

where the $o\,()$ is uniform over θ_{T_k}, $\mathfrak{m}_{T_k}$ and $\overline{\eta}$ satisfying the constraints above.

Proof. The proof is similar to that of Lemma 5.28.

From Lemma 5.21 and since $k \ge k_0$, we have $T^{r^*_\Theta}_{T_k} > T_{k+1}$. Therefore we can apply Lemma 5.9 and, for $T_k \le t \le T_{k+1}$, we have

$$(\theta_t,\, \mathfrak{m}_t) \in B^*_\Theta \times \mathbb{T}_{\mathcal{M}_t}.$$

By construction of the sequence θ_t and by definition of $\Phi_{t_1:t_2}(\theta_{t_1}, \mathfrak{m}_{t_1}, (\eta_s))$ (Def. 4.6), we have

$$\theta_t = \Phi_{T_k:t}(\theta_{T_k}, \mathfrak{m}_{T_k}, (\eta_s)).$$

Then thanks to Lemma 5.17 starting at time T_k, for any $T_k + 1 \le t \le T_{k+1}$ (using again that $T^{r^*_\Theta}_{T_k} > T_{k+1}$), we have

$$d\left(\Phi_{T_k:t}(\theta_{T_k}, \mathfrak{m}_{T_k}, (\eta_s)),\; \Phi_{T_k:t}(\theta_{T_k}, \mathfrak{m}^*_{T_k}, (\eta_s))\right) = O\left(m_{\mathrm{H}}(t) \sup_{T_k+1 \le s \le t} \eta_s\right)$$

and, thanks to the triangle inequality, we obtain, for any $T_k + 1 \le t \le T_{k+1}$,

$$\begin{aligned} d\left(\theta_t,\, \theta^*\right) \le\; & d\left(\Phi_{T_k:t}\left(\theta_{T_k}, \mathfrak{m}^*_{T_k}, (\eta_s)\right),\; \Phi_{T_k:t}\left(\theta^*, \mathfrak{m}^*_{T_k}, (\eta_s)\right)\right) \\ & + d\left(\Phi_{T_k:t}\left(\theta^*, \mathfrak{m}^*_{T_k}, (\eta_s)\right),\; \theta^*\right) + O\left(m_{\mathrm{H}}(t) \sup_{T_k < s \le t} \eta_s\right). \end{aligned}$$

Apply this to $t = T_{k+1}$. By Lemma 5.27, the first term is at most $(1 - \lambda_{\min}\, \eta_{T_k}\, L(T_k))\; d\left(\theta_{T_k}, \theta^*\right) + o\left(\eta_{T_k}\, L(T_k)\right)$.

By Lemma 5.26, the second term is $o\left(\eta_{T_k}\, L(T_k)\right)$.

By Corollary 5.18, the last term is $o\left(\eta_{T_k}\, L(T_k)\right)$. □

Lemma 5.34 (Convergence of the open-loop algorithm: end of proof of Theorem 4.29)**.** *There exists $\overline{\eta}_{\mathrm{conv}} > 0$ such that, for any $0 \le \overline{\eta} \le \overline{\eta}_{\mathrm{conv}}$, the following convergence holds.*

Let $\theta_0 \in \Theta$ with $d\left(\theta_0, \theta^*\right) \le \frac{r^*_\Theta}{4}$, and let $\mathfrak{m}'_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$ be any sequence of "reset states". Consider the trajectory $(\theta_t)_{t \ge 0}$ computed for every $k \ge 0$ and $T_k < t \le T_{k+1}$ by resetting

$$\mathfrak{m}_{T_k} \leftarrow \mathfrak{m}'_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$$

and then

$$\begin{cases} \mathfrak{m}_t = \mathcal{A}_t\left(\theta_{T_k}, \mathfrak{m}_{t-1}\right) \\ v_t = \boldsymbol{V}_t\left(\theta_{T_k}, \mathfrak{m}_t\right) \\ \theta_t = \Phi_t\left(\theta_{t-1}, \eta_t\, v_t\right). \end{cases}$$

Then θ_t tends to θ^* as $t \to \infty$.

Proof. The proof unfolds much like that of Lemma 5.30. It is simpler, in that the relations $\mathfrak{m}_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$, after the substitutions $\mathfrak{m}_{T_k} \leftarrow \mathfrak{m}'_{T_k}$, hold by assumption.

Take $\overline{\eta} \leq \min(\overline{\eta}_{\mathrm{op}}, \overline{\eta}_{\mathcal{V}}/2)$. (This is not yet $\overline{\eta}_{\mathrm{conv}}$: there will be an additional constraint on $\overline{\eta}$ below.)

By Lemma 5.33, there exists $k_0 \geq 0$, and a sequence $b_k = o\left(\eta_{T_k} L(T_k)\right)$ such that

$$d\left(\theta_{T_{k+1}}, \theta^*\right) \leq (1 - \lambda_{\min} \eta_{T_k} L(T_k))\, d\left(\theta_{T_k}, \theta^*\right) + b_k \tag{15}$$

holds for those values of $k \geq k_0$ such that $d\left(\theta_{T_k}, \theta^*\right) \leq \frac{r^*_\Theta}{3}$ and $\mathfrak{m}_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$. By Lemma 5.33, the value of b_k is uniform over $\bar{\eta}$ and the values of θ and $\mathfrak{m}$ satisfying those assumptions.

Since b_k is $o\left(\eta_{T_k} L(T_k)\right)$, there exists $k_1 \geq k_0$ such that b_k is less than $(r^*_\Theta/3)(\lambda_{\min} \eta_{T_k} L(T_k))$ for $k \geq k_1$. (Such a k_1 is uniform in the values of θ, $\mathfrak{m}$ and $\bar{\eta}$ satisfying the assumptions above, because b_k is.)

Define $\overline{\eta}_{\mathrm{conv}} := \min\left(\overline{\eta}_{\mathrm{op}}, \overline{\eta}_{\mathcal{V}}/2, \overline{\eta}^{T_{k_1}}\right)$, where $\overline{\eta}^{T_{k_1}}$ is defined in Lemma 5.11.

Set $r_k := \lambda_{\min} \eta_{T_k} L(T_k)$. We have $b_k = o\left(r_k\right)$.

By Lemma 5.33, if $d\left(\theta_{T_k}, \theta^*\right) \leq \frac{r^*_\Theta}{3}$ and $\mathfrak{m}_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$, then

$$d\left(\theta_{T_{k+1}}, \theta^*\right) \leq (1 - r_k)\, d\left(\theta_{T_k}, \theta^*\right) + b_k.$$

By definition of k_1, if $k \geq k_1$ then $(1 - r_k)\frac{r^*_\Theta}{3} + b_k \leq \frac{r^*_\Theta}{3}$.

Consequently, if $k \geq k_1$ and $d\left(\theta_{T_k}, \theta^*\right) \leq \frac{r^*_\Theta}{3}$ and $\mathfrak{m}_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$, then $d\left(\theta_{T_{k+1}}, \theta^*\right) \leq \frac{r^*_\Theta}{3}$.

By assumption, for every k, after the substitution $\mathfrak{m}_{T_k} \leftarrow \mathfrak{m}'_{T_k}$, we have $\mathfrak{m}_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$. Therefore, by induction, if $d\left(\theta_{T_{k_1}}, \theta^*\right) \leq \frac{r^*_\Theta}{3}$, this holds for any $k \geq k_1$.

Since we assume $d\left(\theta_0, \theta^*\right) \leq \frac{r^*_\Theta}{4}$ and $\mathfrak{m}_0 \in \mathbb{T}_{\mathcal{M}_0}$, by Lemma 5.11 applied to $T = T_{k_1}$, for $\overline{\eta} \leq \overline{\eta}^{T_{k_1}}$, we have $d\left(\theta_{T_{k_1}}, \theta^*\right) \leq \frac{r^*_\Theta}{3}$.

Thus, if $\overline{\eta} \leq \overline{\eta}^{T_{k_1}}$, then $d\left(\theta_{T_k}, \theta^*\right) \leq \frac{r^*_\Theta}{3}$ for all $k \geq k_1$.

Therefore, for any $k \geq k_1$, we have

$$d\left(\theta_{T_{k+1}}, \theta^*\right) \leq (1 - r_k)\, d\left(\theta_{T_k}, \theta^*\right) + b_k.$$

Since $b_k = o\left(r_k\right)$, by Lemma 5.29 this implies that θ_{T_k} tends to θ^* when $k \to \infty$.

For the intermediate times $T_k < t \leq T_{k+1}$, by Lemma 5.13, we have

$$d\left(\theta_t, \theta_{T_k}\right) \leq \kappa_6 \sum_{T_k < s \leq T_{k+1}} \eta_s\, m(s)$$

(we can apply Lemma 5.13 because we stay in the stable tube for $t \geq T_{k_1}$). By Corollary 5.18 this proves that θ_t tends to θ^* if θ_{T_k} does. □

6 Controlling RTRL and Imperfect RTRL Algorithms around the Target Trajectory

We now turn back to the setting of Section 2. We proceed by making the connection with the more abstract setting of Section 4, with a suitable abstract state $\mathfrak{m}_t = (s_t, J_t)$ where s_t is the state of the original dynamical system, and J_t is the quantity

maintained by RTRL. Notably, we relate the assumptions of Section 2 to those of Section 4. Convergence will then result from Theorems 4.27, 4.28, or 4.29, depending on the case.

Thus, we now work under the assumptions of Section 2. As before, throughout the proof, the constants implied in the $O()$ notation only depend on the constants and $O()$ directly appearing in the assumptions (Remark 5.2).

6.1 Applying the Abstract Convergence Theorem to RTRL

To prove convergence of the RTRL algorithm, we will apply Theorem 4.27 or Theorem 4.28 to the state of the algorithm. The latter is composed not only of the state of the system $s_t \in \mathcal{S}_t$, but also of the Jacobian J_t, which is maintained by the algorithm. The Jacobian J_t is an element of the space of linear maps $\mathrm{L}(\Theta, \mathcal{S}_t)$. Thus, the state of the algorithm will be the pair $\mathfrak{m}_t = (s_t, J_t)$. The definition of RTRL provides the transition function on $\mathfrak{m}_t$, together with the way to compute gradients.

This is gathered in the following definition. The purpose here is to bring the RTRL algorithm as defined in Definition 2.9 into the framework of Section 4.1.

Definition 6.1 (RTRL as an abstract gradient descent algorithm)**.** *Given a parameterized dynamical system (Defs. 2.6–2.7) and an extended RTRL algorithm (Def. 2.9), the transition operators $(\mathcal{A}_t)$ (Def. 4.2) associated with this RTRL algorithm are defined as follows. The parameter space is Θ and the state space is*

$$\mathcal{M}_t := \mathcal{S}_t \times \mathrm{L}(\Theta, \mathcal{S}_t)$$

equipped with the norm $\|(s, J)\| = \max(\|s\|, \|J\|)$. The transition operators $\mathcal{A}_t \colon \Theta \times \mathcal{M}_{t-1} \to \mathcal{M}_t$ are defined by

$$\mathcal{A}_t(\theta, (s, J)) := \left(\mathbf{T}_t(s, \theta), \frac{\partial \mathbf{T}_t(s, \theta)}{\partial s} J + \frac{\partial \mathbf{T}_t(s, \theta)}{\partial \theta} \right)$$

and the gradient computation operators $\boldsymbol{V}_t \colon \Theta \times \mathcal{M}_t \to \Theta$ (Def. 4.3) are set to

$$\boldsymbol{V}_t(\theta, (s, J)) := \mathcal{U}_t \left(\frac{\partial \mathcal{L}_t(s)}{\partial s} J,\ s,\ \theta \right).$$

Finally, the update operators Φ_t of Def. 4.5 are those of the RTRL algorithm (Def. 2.9).

The rest of the text is devoted to proving that this abstract algorithm satisfies all the assumptions of Section 4.

We have to prove that these assumptions hold for θ in some ball B_Θ^* (Section 4). We start by setting B_Θ^* to the ball $B_\Theta(\theta^*, r_\Theta)$ where the assumptions of Section 2 hold. This ball B_Θ^* will be reduced several times in the course of the proof so that elements $\theta \in B_\Theta^*$ satisfy further properties.

Definition 6.2 (Notation for iterates)**.** *Let $0 \leq t_1 \leq t_2$. Given $s_{t_1} \in \mathcal{S}_{t_1}$ and a sequence of parameters $(\theta_t)_{t \geq 0}$ in Θ, we denote*

$$\mathbf{T}_{t_1:t_2}(s_{t_1}, (\theta_t)) := s_{t_2}$$

where the sequence (s_t) is defined inductively via $s_t = \mathbf{T}_t(s_{t-1}, \theta_{t-1})$ for $t > t_1$. If $(\theta_t) \equiv \theta$ is constant we just write $\mathbf{T}_{t_1:t_2}(s_{t_1}, \theta)$.

Next, we define the norm on Θ that will be used in the proof. Indeed, convergence in Section 4 is based on a contractivity property in a certain distance (Assumption 4.18.2). But the dynamics of learning is not contractive for any distance on Θ, only for distances built from a suitable Lyapunov function.

For non-extended RTRL (no $\mathcal{U}_t$), the suitable norm on Θ is directly given by the Hessian of the average loss at θ^*. For extended RTRL algorithms, remember the notation from Assumption 2.11.b: the Jacobian of the update direction, over time, averages to a matrix Λ whose eigenvalues have positive real part, which plays the role of an extended Hessian of the average loss. This matrix controls the asymptotic dynamics of learning around θ^*, which is equivalent to $(\theta - \theta^*)' = -\Lambda(\theta - \theta^*)$ in the continuous-time limit when θ is close to θ^* (see Section 7.3).

We select a norm on Θ based on Λ, such that this dynamics is contractive. This is based on a classical linear algebra result.

Lemma 6.3 (Existence of a suitable Lyapunov function)**.** *There exists a positive definite matrix B such that $B\,\Lambda + \Lambda^T\,B$ is positive definite.*

Proof. This is a consequence of the fact the eigenvalues of Λ have positive real part. See Appendix A. □

From now on we endow Θ with the norm given by B, namely, we set

$$\|\theta\|^2 := \theta^\top B\,\theta$$

where B is such that $B\Lambda + \Lambda^\top B$ is positive definite and Λ is given by Assumption 2.11.b. This norm will be used as an approximate Lyapunov function for the algorithm.

Note that the assumptions in Section 2 have been expressed with respect to an unspecified norm on Θ. Since Θ is finite-dimensional, all norms are equivalent; in particular, we can find a ball for the new norm that is included in the original ball $B_\Theta(\theta^*, r_\Theta)$ on which the assumptions hold. Assumptions 2.11.a, 2.11.b, 2.17, 2.21, 2.23, 2.25.a, and 2.25.b also involve a norm on Θ via norms on objects such as derivatives with respect to θ; a change to an equivalent norm only introduces a constant factor which is absorbed in the $O\,()$ in these assumptions. Therefore, up to restriction to the smaller ball, all the assumptions of Section 2 hold with respect to the norm we just defined.

6.2 RTRL Computes the Correct Derivatives

We first prove that RTRL indeed computes the correct derivatives of the loss (this is actually how RTRL is built in the first place) when the parameter is kept fixed. This implies that over a time interval, the open-loop (fixed-parameter) algorithm computes a parameter update equal to the derivative of the loss, summed over this interval. (When the parameter is actually updated at every step, this will be true only up to some higher-order terms, controlled in Section 7.)

Lemma 6.4 (RTRL computes the correct derivatives for the open-loop trajectory)**.** *Call* open-loop RTRL *the algorithm of Definition 2.9 with $\eta_t = 0$ for all t (i.e., θ is kept fixed).*

Then the quantities J_t and v_t computed by open-loop RTRL starting at s_0 and $\theta_0 = \theta$, are respectively equal to the Jacobian of the state with respect to the parameter,

$$J_t = \frac{\partial \mathbf{s}_t(s_0, \theta)}{\partial \theta}$$

and to the derivative of the loss with respect to the parameter fed to the extended update rule,

$$v_t = \mathcal{U}_t\left(\frac{\partial \mathcal{L}_{\rightsquigarrow t}(s_0,\theta)}{\partial \theta},\, \mathbf{s}_t\left(s_0,\,\theta\right),\,\theta\right)$$

for all $t \geq 1$.

In other words, the RTRL algorithm $\mathcal{A}_t$ *(Definition 6.1) satisfies, for any* $\theta \in \Theta$ *and* $s_0 \in \mathcal{S}_0$,

$$\mathcal{A}_{0:t}(\theta,(s_0,0)) = \left(\mathbf{s}_t(s_0,\theta),\,\frac{\partial \mathbf{s}_t(s_0,\theta)}{\partial \theta}\right)$$

and

$$\boldsymbol{V}_t\left(\theta,\,\mathcal{A}_{0:t}(\theta,(s_0,0))\right) = \mathcal{U}_t\left(\frac{\partial \mathcal{L}_{\rightsquigarrow t}(s_0,\theta)}{\partial \theta},\,\mathbf{s}_t\left(s_0,\,\theta\right),\,\theta\right).$$

Proof. RTRL is actually built to obtain this property, as explained before Definition 2.9. Indeed, when $\eta_t = 0$, the parameter θ is constant along the trajectory. Then by definition the state s_t in Definition 2.9 is $s_t = \mathbf{s}_t(s_0,\theta) = \mathbf{T}_t(\mathbf{s}_{t-1}(s_0,\theta),\theta)$. By differentiation, we find that $\partial \mathbf{s}_t/\partial\theta$ satisfies the linear evolution equation

$$\frac{\partial \mathbf{s}_t(s_0,\theta)}{\partial \theta} = \frac{\partial \mathbf{T}_t}{\partial s}\frac{\partial \mathbf{s}_{t-1}(s_0,\theta)}{\partial \theta} + \frac{\partial \mathbf{T}_t}{\partial \theta}$$

where the derivatives of $\mathbf{T}_t$ are evaluated at (s_{t-1},θ). This is the evolution equation for J in Definition 2.9, so J is equal to this quantity. (The initialization $J = 0$ corresponds to $\partial \mathbf{s}_0(s_0,\theta)/\partial\theta = \partial s_0/\partial\theta = 0$.)

Then the expressions for v_t and $\boldsymbol{V}_t$ follow from their definitions in Defs. 2.9 and 6.1, and the chain rule applied to the definition of $\mathcal{L}_{\rightsquigarrow t}$ (Def. 2.7). □

Recall that by Definition 4.5 and Definition 4.6, the RTRL algorithm defines an iterated update operator $\Phi_{t_1:t_2}(\theta,(v_t))$ and an open-loop update operator $\Phi_{t_1:t_2}(\theta,\mathfrak{m}_{t_1},(\eta_t))$.

Corollary 6.5. *Set* $\mathfrak{m}_0^* := (s_0^*, 0)$ *(state of the RTRL algorithm initialized at* s_0^* *with* $J_0 = 0$*).*

Then for any $\theta \in \Theta$*, for any* $1 \leq t_1 \leq t_2$*, for any sequence of learning rates* $(\eta_{t;\,t_1,t_2})$ *(not necessarily satisfying Assumption 2.26), the open-loop operator of RTRL updates* θ *by the recurrent derivatives of the loss, fed to the extended update rule:*

$$\Phi_{t_1:t_2}(\theta,\mathfrak{m}_{t_1},(\eta_{t;\,t_1,t_2})_t) = \Phi_{t_1:t_2}\left(\theta,\left(\eta_{t;\,t_1,t_2}\,\mathcal{U}_t\left(\frac{\partial \mathcal{L}_{\rightsquigarrow t}(s_0^*,\theta)}{\partial \theta},\,\mathbf{s}_t\left(s_0^*,\,\theta\right),\,\theta\right)\right)_t\right)$$

where $\mathfrak{m}_{t_1} = \mathcal{A}_{0:t_1}(\theta,\mathfrak{m}_0^*)$ *is the RTRL state obtained at time* t_1 *from parameter* θ*.*

Proof. By Definition 4.6, for any family of numbers (η_t), the quantity $\Phi_{t_1:t_2}(\theta,\mathfrak{m}_{t_1},(\eta_t))$ is equal to $\Phi_{t_1:t_2}(\theta,(\eta_t\,v_t))$ where $v_t = \boldsymbol{V}_t(\theta,\mathcal{A}_{t_1:t}(\theta,\mathfrak{m}_{t_1}))$. With $\mathfrak{m}_{t_1} = \mathcal{A}_{0:t_1}(\theta,\mathfrak{m}_0^*)$ we have $\mathcal{A}_{t_1:t}(\theta,\mathfrak{m}_{t_1}) = \mathcal{A}_{0:t}(\theta,\mathfrak{m}_0^*)$. The result follows by the expression for $\boldsymbol{V}_t$ in Lemma 6.4. □

6.3 On the Sequence of Step Sizes

Next, let us deal with the assumptions on the learning rate sequence. We start by building the scale function L used in Section 4 (notably Assumption 4.18, and the timescale of Definition 4.21) from the assumptions in Section 2.

Lemma 6.6 (Intervals for averaging)**.** *Let A and a' be numbers such that $\max(a, \gamma) < a' < A < b - 2\gamma$ for RTRL and extended RTRL algorithms, and such that $\max(a, \gamma + 1/2) < a' < A < b - 2\gamma$ for imperfect RTRL algorithms. In both cases, the set of such pairs (a', A) is non-empty under Assumption 2.26.*

Define the scale functions

$$L(t) := t^A, \qquad e_0(t) := t^{a'}, \qquad m(t) := t^\gamma.$$

Then $e_0(t)$ and $m(t)$ are negligible in front of L, L is negligible in front of the identity function, and $\eta_t\, L(t)\, m(t)^2 \to 0$ as t tends to infinity.

Moreover, Assumptions 2.11.a and 2.11.b are still satisfied with a' instead of a.

Remark 6.7. *In Section 2, we have presented the assumptions and results using rates t^a and t^γ. Actually all our results are valid as long as these expressions are scale functions (Definition 4.8). This is why we use the more abstract notation with scale functions L, e_0, and m in the following.*

Proof. First, note that the range of values for A is non-empty: indeed, by Assumption 2.26 we have $\max(a, \gamma) + 2\gamma < b$. For imperfect RTRL algorithms, Assumption 2.26 further states that $\max(a, \gamma + 1/2) + 2\gamma < b$. Therefore, the requirements that $A < b - 2\gamma$ and $A > \max(a, \gamma + 1/2)$ are mutually compatible and the range for A is non-empty.

We know that e_0 and $m(\cdot)$ are negligible in front of L since $\gamma < a' < A$. We have $\eta_t\, L(t)\, m(t)^2 \to 0$ when t tends to infinity, since $-b + A + 2\gamma < 0$.

Finally, since $a' > a$, Assumptions 2.11.a and 2.11.b are a fortiori satisfied with a' instead of a. □

Lemma 6.8 (Comparison relations for scale functions)**.** *Assume the overall learning rate $\overline{\eta}$ is small enough so that $\eta_t \leq 1$ for all t. Then under Assumption 2.26, the sequence $1/\eta_t$ is a scale function. Moreover, $\eta_t\, e_0(t)\, m(t)^2 \to 0$ and $m(t) = o(e_0(t))$ as $t \to \infty$.*

Proof. By the choice of $\overline{\eta}$, we have $1/\eta_t \geq 1$ for all t. Moreover, $1/\eta_t$ is non-decreasing by assumption on η_t. Now, by Assumption 2.26, $1/\eta_t$ is equivalent to t^b which is a scale function, and therefore, $1/\eta_t$ preserves asymptotic equivalence at ∞.

The last statements are rewritings of the conditions $a' + 2\gamma < b$ and $\gamma < a'$ from Lemma 6.6. □

Lemma 6.9 (Timescale for extended RTRL algorithms)**.** *With this choice of L, the timescale (T_k) of Definition 4.21 amounts to $T_0 = 0$, $T_1 = 1$ and, for $k \geq 1$,*

$$T_{k+1} = T_k + {T_k}^A.$$

Moreover, it satisfies $T_k \sim c\, k^{1/(1-A)}$ for some $c > 0$ as $k \to \infty$.

Proof. The first statement is by direct substitution of $L(t) = t^A$ in Definition 4.21. For the second statement, let $\beta \geq 0$. We have

$$\begin{aligned} {T_{k+1}}^\beta &= \left(T_k + {T_k}^A\right)^\beta = {T_k}^\beta \left(1 + \frac{1}{{T_k}^{1-A}}\right)^\beta = {T_k}^\beta \left(1 + \frac{\beta}{{T_k}^{1-A}} + o\left(\frac{1}{{T_k}^{1-A}}\right)\right) \\ &= {T_k}^\beta + \frac{\beta}{{T_k}^{1-A-\beta}} + o\left(\frac{1}{{T_k}^{1-A-\beta}}\right), \end{aligned}$$

as $k \to \infty$. Taking $\beta = 1 - A > 0$, we obtain that ${T_{k+1}}^{1-A} - {T_k}^{1-A} \sim 1 - A$, as $k \to \infty$, so that ${T_k}^{1-A} \sim (1 - A)k$ as $k \to \infty$. □

Lemma 6.10 (Homogeneity satisfied)**.** *For $t \geq 0$, let I_t be the segment $I_t = [t+1, t+L(t)]$. Then*

$$\frac{\sup_{s\in I_t} \eta_s}{\inf_{s\in I_t} \eta_s} = 1 + o\left(\frac{1}{m(t)}\right)$$

as t tends to infinity.

Proof. For every $t \geq 1$, by the definition of η_t in Assumption 2.26, we have

$$\frac{\sup_{I_t} \eta_s}{\inf_{I_t} \eta_s} = \frac{\eta_{t+1}}{\eta_{t+L(t)}} = \left(\frac{t+L(t)}{t+1}\right)^b \frac{\left(1+o\left(\frac{1}{m(t+1)}\right)\right)}{\left(1+o\left(\frac{1}{m(t+L(t))}\right)\right)}.$$

For $t \to \infty$, we have

$$\begin{aligned}\left(\frac{t+L(t)}{t+1}\right)^b &= \left(1+\frac{L(t)}{t}\right)^b \left(1+o\left(\frac{1}{t}\right)\right)^b = \left(1+\frac{1}{t^{1-A}}\right)^b \left(1+o\left(\frac{1}{t}\right)\right) \\ &= \left(1+O\left(\frac{1}{t^{1-A}}\right)\right)\left(1+o\left(\frac{1}{t}\right)\right) = 1+O\left(\frac{1}{t^{1-A}}\right) = 1+o\left(\frac{1}{t^\gamma}\right),\end{aligned}$$

since, according to Lemma 6.6, we have $0 \leq A < b - 2\,\gamma \leq 1-\gamma$, so that $1-A \leq 1$ and $1-A > \gamma$. Moreover, still when $t \to \infty$, we have

$$\begin{aligned}\frac{\left(1+o\left(\frac{1}{m(t+1)}\right)\right)}{\left(1+o\left(\frac{1}{m(t+L(t))}\right)\right)} &= \frac{1+o\left(\frac{1}{(t+1)^\gamma}\right)}{1+o\left(\left(\frac{1}{t+t^A}\right)^\gamma\right)} = \frac{1+o\left(\frac{1}{t^\gamma}\right)}{1+o\left(\frac{1}{t^\gamma}\right)} \\ &= 1+o\left(\frac{1}{t^\gamma}\right).\end{aligned}$$

As a result, when $t \to \infty$, we have

$$\frac{\sup_{I_t} \eta_s}{\inf_{I_t} \eta_s} = 1+o\left(\frac{1}{t^\gamma}\right),$$

which ends the proof, since for every $t \geq 1$, $m(t) = t^\gamma$. □

Corollary 6.11 (Suitable stepsizes)**.** *The stepsize sequence $\boldsymbol{\eta} = (\eta_t)$, together with the scale function L, satisfy Assumption 4.19, taking $m_{\mathrm{H}}(t) = m(t)$.*

Proof. This is a direct consequence of Assumption 2.26 and Lemma 6.6, and of the fact we use $m_{\mathrm{H}}(t) = m(t)$. The homogeneity assumption is a consequence of Lemma 6.10. □

6.4 Local Boundedness of Derivatives, Short-Time Control

Lemma 6.12 (Controlling the derivatives of the transition operators around θ^*)**.** *Let B be the bound on second derivatives appearing in Assumption 2.23. Then for $\theta \in B_\Theta(\theta^*, r_\Theta)$ and $s \in B_{\mathcal{S}_{t-1}}(s^*_{t-1}, r_{\mathcal{S}})$, one has*

$$\left\|\frac{\partial \mathbf{T}_t}{\partial(s,\theta)}(s,\theta) - \frac{\partial \mathbf{T}_t}{\partial(s,\theta)}(s^*,\theta^*)\right\|_{\mathrm{op}} \leq B \max(\|s-s^*\|, \|\theta-\theta^*\|)$$

and therefore

$$\sup_{t\geq 1} \sup_{\substack{\theta\in B_\Theta(\theta^*,r_\Theta) \\ s\in B_{\mathcal{S}_{t-1}}(s^*_{t-1},r_{\mathcal{S}})}} \left\|\frac{\partial \mathbf{T}_t}{\partial(s,\theta)}(s,\theta)\right\|_{\mathrm{op}} < \infty.$$

Proof. This is a direct consequence of Assumption 2.23. Indeed, let $\theta \in B_\Theta(\theta^*, r_\Theta)$ and $s \in B_{\mathcal{S}_{t-1}}(s^*_{t-1}, r_\mathcal{S})$. For $0 \le u \le 1$, set

$$s_u = (1-u)s^* + us, \qquad \theta_u = (1-u)\theta^* + u\theta$$

so that

$$\frac{\partial \mathbf{T}_t}{\partial(s,\theta)}(s,\theta) = \frac{\partial \mathbf{T}_t}{\partial(s,\theta)}(s^*,\theta^*) + \int_{u=0}^1 \left(\frac{\partial^2 \mathbf{T}_t}{\partial(s,\theta)^2}(s_u,\theta_u)\right) \cdot \frac{\mathrm{d}(s_u,\theta_u)}{\mathrm{d}u}\,\mathrm{d}u$$

and now the operator norm of $\frac{\partial^2 \mathbf{T}_t}{\partial(s,\theta)^2}$ is bounded by Assumption 2.23, and $\frac{\mathrm{d}(s_u,\theta_u)}{\mathrm{d}u} = (s-s^*, \theta-\theta^*)$. This proves the first claim.

The second claim follows since $\frac{\partial \mathbf{T}_t}{\partial(s,\theta)}(s^*,\theta^*)$ is bounded by assumption, and $s-s^*$ and $\theta-\theta^*$ are bounded by definition in the balls considered. □

Lemma 6.13. *The operators $\mathbf{T}_t$ are uniformly Lipschitz on $B_\Theta(\theta^*, r_\Theta)$ and $B_{\mathcal{S}_{t-1}}(s^*_{t-1}, r_\mathcal{S})$. Namely, there exists a constant $\kappa_7 \ge 1$ such that for any $t \ge 1$, for any $\theta, \theta' \in B_\Theta(\theta^*, r_\Theta)$, for any $s, s' \in B_{\mathcal{S}_{t-1}}(s^*_{t-1}, r_\mathcal{S})$, one has*

$$\left\|\mathbf{T}_t(s,\theta) - \mathbf{T}_t(s',\theta')\right\| \le \kappa_7 \max(\left\|s-s'\right\|, \left\|\theta-\theta'\right\|).$$

Proof. This is a consequence of Lemma 6.12. Indeed, for $0 \le u \le 1$, set as above

$$s_u = (1-u)s + us', \qquad \theta_u = (1-u)\theta + u\theta'$$

so that

$$\mathbf{T}_t(s',\theta') = \mathbf{T}_t(s,\theta) + \int_{u=0}^1 \left(\frac{\partial \mathbf{T}_t}{\partial(s,\theta)}(s_u,\theta_u)\right) \cdot \frac{\mathrm{d}(s_u,\theta_u)}{\mathrm{d}u}\,\mathrm{d}u$$

and now the operator norm of $\frac{\partial \mathbf{T}_t}{\partial(s,\theta)}$ is bounded by Lemma 6.12, and $\frac{\mathrm{d}(s_u,\theta_u)}{\mathrm{d}u} = (s-s^*, \theta-\theta^*)$. This proves the claim. □

Corollary 6.14. *Let $0 \le t_1 \le t_2$. Let (θ_t) and (θ'_t) be two sequences of parameters with $\sup_t \|\theta_t - \theta^*\| \le \min(r_\Theta, r_\mathcal{S}/\kappa_7^{t_2-t_1})$ and likewise for θ'_t. Let $s, s' \in \mathcal{S}_{t_1}$ with $\|s - s^*_{t_1}\| \le r_\mathcal{S}/\kappa_7^{t_2-t_1}$ and likewise for s'.*

Then for every $t_1 \le t \le t_2$,

$$\left\|\mathbf{T}_{t_1:t}(s,(\theta_t)) - \mathbf{T}_{t_1:t}(s',(\theta'_t))\right\| \le \kappa_7^{t-t_1} \max(\left\|s-s'\right\|, \sup_{t'} \left\|\theta_{t'} - \theta'_{t'}\right\|)$$

and both $\mathbf{T}_{t_1:t}(s,(\theta_t))$ and $\mathbf{T}_{t_1:t}(s',(\theta'_t))$ lie in $B_\mathcal{S}(s^*_t, r_\mathcal{S})$.

Proof. By induction from Lemma 6.13. First consider the case $s' = s^*_{t_1}$ and $\theta'_t = \theta^*$: by induction from Lemma 6.13, we obtain that

$$\left\|\mathbf{T}_{t_1:t}(s,(\theta_t)) - \mathbf{T}_{t_1:t}(s^*_{t_1},\theta^*)\right\| \le \kappa_7^{t-t_1} \max(\left\|s-s^*_{t_1}\right\|, \sup_{t'} \left\|\theta_{t'} - \theta^*\right\|) \le r_\mathcal{S}$$

and therefore, since $\mathbf{T}_{t_1:t}(s^*_{t_1},\theta^*) = s^*_t$ by definition, we obtain that $\mathbf{T}_{t_1:t}(s,(\theta_t)) \in B_s(s^*_t, r_\mathcal{S})$. Thus Lemma 6.13 can be applied at the next step of the induction.

Next, consider the case of general s'. By the first step above, both $\mathbf{T}_{t_1:t}(s',(\theta_t)) \in B_s(s^*_t, r_\mathcal{S})$ and $\mathbf{T}_{t_1:t}(s',(\theta'_t)) \in B_s(s^*_t, r_\mathcal{S})$ lie in the ball $B_s(s^*_t, r_\mathcal{S})$. So Lemma 6.13 can be applied at all times $t \le t_2$, which gives the result by induction. □

6.5 Spectral Radius Close to θ^*

Proposition 6.15 (Continuity of spectral radius for sequences)**.** *Let $(A_t)_{t\geq 0}$ be a sequence of linear operators over a normed vector space, with bounded operator norm. Assume that (A_t) has spectral radius $\leq 1-\alpha$ at horizon h. Then there exists $\varepsilon > 0$ such that if (A'_t) is a sequence of linear operators with $\|A_t - A'_t\|_{\mathrm{op}} \leq \varepsilon$ for all t, then the sequence (A'_t) has spectral radius $\leq 1-\alpha/2$ at horizon h.*

Proof. Writing $A'_t =: A_t + r_t$ and expanding the product $A'_{t+h-1}\dots...A'_{t+1}A'_t$, one finds 2^h terms, one of which is $A_{t+h-1}\dots...A_{t+1}A_t$ and all the others involve at least one r_s factor. Therefore, if $\|r_s\|_{\mathrm{op}} \leq \frac{\alpha}{2^{h+1}}\min(1, (1/\sup\|A_t\|_{\mathrm{op}})^h)$, each of those terms has operator norm $\leq \frac{\alpha}{2^{h+1}}$. So the sum of all the terms with at least one r_s factor has operator norm $\leq \alpha/2$ and the conclusion follows. □

Corollary 6.16 (Balls with spectral radius bounded away from 1)**.** *Let h be the horizon for the spectral radius in Assumption 2.13.*

There exist $r'_\Theta > 0$, $r'_{\mathcal{S}} > 0$, and $M > 0$ such that, for any sequence of parameters $(\theta_t)_{t\geq 0}$ with $\theta_t \in B_\Theta(\theta^*, r'_\Theta)$ and any sequence of states $(s_t)_{t\geq 0}$ with $s_t \in B_{\mathcal{S}_t}(s^*_t, r'_{\mathcal{S}})$, the sequence of operators

$$\frac{\partial \mathbf{T}_t}{\partial s}(s_{t-1}, \theta_{t-1})$$

has spectral radius at most $1-\alpha/2$ at horizon h. Moreover, any product of such consecutive operators has operator norm bounded by

$$\left\| \prod_{t_1 < t \leq t_2} \frac{\partial \mathbf{T}_t}{\partial s}(s_{t-1}, \theta_{t-1}) \right\|_{\mathrm{op}} \leq M(1-\alpha/2)^{(t_2-t_1)/h}.$$

Proof. One has

$$\left\| \frac{\partial \mathbf{T}_t}{\partial(s,\theta)}(s,\theta) - \frac{\partial \mathbf{T}_t}{\partial(s,\theta)}(s^*,\theta^*) \right\|_{\mathrm{op}} \geq \left\| \frac{\partial \mathbf{T}_t}{\partial s}(s,\theta) - \frac{\partial \mathbf{T}_t}{\partial s}(s^*,\theta^*) \right\|_{\mathrm{op}}$$

since any change of s can be seen as a change of (s,θ) with no change on θ.

Therefore by Lemma 6.12, if (s_t, θ_t) is close enough to (s^*_t, θ^*), then $\frac{\partial \mathbf{T}_{t+1}}{\partial s}(s_t, \theta_t)$ is arbitrarily close to $\frac{\partial \mathbf{T}_{t+1}}{\partial s}(s^*_t, \theta^*)$ in operator norm. The spectral radius property follows by Assumption 2.13 and Proposition 6.15.

For the last inequality, divide the time interval $(t_1; t_2]$ into blocks of length h, plus a remainder of length $< h$. On each consecutive block of length h, by definition of the spectral radius of a sequence (Def. 2.12), the operator norm of the product is at most $(1-\alpha/2)$. For the remaining interval of length $< h$, define

$$M_0 = \max\left(1, \sup_{t\geq 1} \sup_{\substack{\theta\in B_\Theta(\theta^*, r_\Theta) \\ s\in B_{\mathcal{S}_{t-1}}(s^*_{t-1}, r_{\mathcal{S}})}} \left\| \frac{\partial \mathbf{T}_t}{\partial s}(s,\theta) \right\|_{\mathrm{op}}\right),$$

which is finite thanks to Lemma 6.12. Thus, a product of $< h$ consecutive operators has operator norm at most M_0^h. Defining $M = M_0^h/(1-\alpha/2)$ proves the claim (the $1/(1-\alpha/2)$ compensates for $t_2 - t_1$ not being an exact multiple of h). □

Lemma 6.17 (Balls with contractivity at the horizon)**.** *Let h be the horizon for the spectral radius in Assumption 2.13, and $1-\alpha$ the corresponding operator norm.*

Define $r''_\Theta = \min(r_\Theta, r'_\Theta, r_\mathcal{S}/\kappa_7^h, r'_\mathcal{S}/\kappa_7^h)$ and $r''_\mathcal{S} = \min(r_\mathcal{S}, r'_\mathcal{S})/\kappa_7^h$ with κ_7 as in Corollary 6.14.

Let $\theta \in B_\Theta(\theta^*, r''_\Theta)$ and let $s, s' \in B_{\mathcal{S}_t}(s_t^*, r''_\mathcal{S})$.

Then for all $t \leq t' \leq t+h$, $\mathbf{T}_{t:t'}(s,\theta)$ and $\mathbf{T}_{t:t'}(s',\theta)$ belong to the ball $B_{\mathcal{S}_{t'}}(s_{t'}^*, r'_\mathcal{S})$, and moreover

$$\left\|\mathbf{T}_{t:t+h}(s,\theta) - \mathbf{T}_{t:t+h}(s',\theta)\right\| \leq (1-\alpha/2)\left\|s-s'\right\|.$$

In particular, taking $s' = s_t^*$, we see that s_{t+h} belongs to $B_{\mathcal{S}_{t+h}}(s_{t+h}^*, r''_\mathcal{S})$.

Proof. For $0 \leq u \leq 1$ set $s_u = (1-u)s' + us$, which belongs to $B_{\mathcal{S}_t}(s_t^*, r''_\mathcal{S})$.

Then

$$\mathbf{T}_{t:t'}(s,\theta) = \mathbf{T}_{t:t'}(s',\theta) + \int_{u=0}^{1} \left(\frac{\partial \mathbf{T}_{t:t'}}{\partial s}(s_u,\theta)\right) \cdot \frac{\mathrm{d}s_u}{\mathrm{d}u}\,\mathrm{d}u.$$

Denote $s_{u,t'} = \mathbf{T}_{t:t'}(s_u,\theta)$ for $t' \geq t$ the trajectory starting at s_u with parameter θ. Since $\mathbf{T}_{t:t'+1} = \mathbf{T}_{t'+1}(\mathbf{T}_{t:t'})$, by induction the derivative of $\mathbf{T}_{t:t'}$ is the product of derivatives along the trajectory:

$$\frac{\partial \mathbf{T}_{t:t'}}{\partial s}(s_u,\theta) = \frac{\partial \mathbf{T}_{t'}}{\partial s}(s_{u,t'-1},\theta)\frac{\partial \mathbf{T}_{t'-1}}{\partial s}(s_{u,t'-2},\theta)\cdots\frac{\partial \mathbf{T}_{t+1}}{\partial s}(s_u,\theta).$$

Since $s_u \in B_{\mathcal{S}_t}(s_t^*, r''_\mathcal{S})$, by Corollary 6.14, for any $t \leq t' \leq t+h$ we have

$$\left\|\mathbf{T}_{t:t'}(s_u,\theta)) - \mathbf{T}_{t:t'}(s_t^*,\theta^*)\right\| \leq \kappa_7^h \max(\left\|s_u - s_t^*\right\|, \left\|\theta-\theta^*\right\|) \leq \kappa_7^h \max(r''_\mathcal{S}, r''_\Theta) \leq r'_\mathcal{S} \tag{16}$$

by our definition of r''_Θ and $r''_\mathcal{S}$.

Since $\mathbf{T}_{t:t'}(s_t^*,\theta^*) = s_{t'}^*$ and $\mathbf{T}_{t:t'}(s_u,\theta) = s_{u,t'}$ by definition, this means that $s_{u,t'}$ belongs to $B_{\mathcal{S}_{t'}}(s_{t'}^*, r'_\mathcal{S})$.

Therefore we can apply Corollary 6.16. We obtain that the sequence $\frac{\partial \mathbf{T}_{t'}}{\partial s}(s_{u,t'-1},\theta)$ for $t \leq t' \leq t+h$, has spectral radius at most $1-\alpha/2$ at horizon h. Therefore, taking $t' = t+h$ we have

$$\left\|\frac{\partial \mathbf{T}_{t:t+h}}{\partial s}(s_u,\theta)\right\|_{\mathrm{op}} \leq 1-\alpha/2.$$

Since $\left\|\frac{\mathrm{d}s_u}{\mathrm{d}u}\right\| = \|s-s'\|$, the conclusion follows. □

6.6 Stable Tubes for RTRL and Imperfect RTRL

6.6.1 Existence of a Stable Tube for the States s_t

We are now ready to construct stable tubes for $\mathbf{T}$. We cannot construct stable balls in a straightforward way, as contractivity needs h iterations to operate. As a result, we construct two sets of balls in the state spaces $\mathcal{S}_t$ around the target trajectory (s_t^*) (of course, all the balls are included in the balls where smoothness, and Lipschitz assumptions, are satisfied). The successive radii are smaller as the number of "primes" increases.

1. The balls of radius $r_\mathcal{S}$ are those where the regularity Assumption 2.23 is satisfied.

2. The balls of radius $r'_\mathcal{S}$ are those where the several differentials of the model are bounded, and were the $\frac{\partial \mathbf{T}_t}{\partial s}$'s have spectral radius less than $1-\alpha$.

3. The balls of radius $r''_{\mathcal{S}}$ are those the states in which cannot escape from the balls of radius $r'_{\mathcal{S}}$ in h iterations.

4. The balls of radius $r''''_{\mathcal{S}}$ are stable by h successive iterations of the $\mathbf{T}_t$'s, provided parameters in $B_\Theta(\theta^*, r'''_\Theta)$ are used.

5. In between times t and $t+h$, the states of trajectories issuing from a ball $B_{\mathcal{S}_t}(s^*_t, r''''_{\mathcal{S}})$, and using parameters in $B_\Theta(\theta^*, r'''_\Theta)$, may get out of balls of radius $r''''_{\mathcal{S}}$, but remain in balls of radius $r'''_{\mathcal{S}}$.

As a result, every trajectory issuing from a ball $B_{\mathcal{S}_t}(s^*_t, r''''_{\mathcal{S}})$ at some time t, and using parameters in $B_\Theta(\theta^*, r'''_\Theta)$, will behave as follows.

1. At every time $t+n\,h$, where $n \geq 0$ is an integer, $s_{t+n\,h}$ is in $B_{\mathcal{S}_{t+n\,h}}(s^*_{t+n\,h}, r''''_{\mathcal{S}})$.

2. At times $t+r+n\,h$, with $r < h$, $s_{t+r+n\,h}$ is in a ball $B_{\mathcal{S}_{t+r+n\,h}}(s^*_{t+r+n\,h}, r'''_{\mathcal{S}})$.

Finally, at any time t, any state $s_t \in \mathbb{T}_t$ is guaranteed to stay in the balls where the $\mathbf{T}_t$'s are smooth, and "have spectral radius less than $1-\alpha$" in h iterations: for every $t \leq t' \leq t+k$, for every sequence (θ_p) of parameters in $B_\Theta(\theta^*, r'''_\Theta)$, we have $\mathbf{T}_{t:t'}(s_t, (\theta_p)) \in B_{\mathcal{S}_{t'}}(s^*_{t'}, r_{\mathcal{S}}) \cap B_{\mathcal{S}_{t'}}(s^*_{t'}, r'_{\mathcal{S}})$ (this is a consequence of Corollary 6.14).

Lemma 6.18 (Existence of a stable tube for s)**.** *There exist a ball $B^*_\Theta := B_\Theta(\theta^*, r'''_\Theta)$ with positive radius, and sets $\mathbb{T}_t \subset \mathcal{S}_t$ with the following properties:*

1. *Stability: for any $\theta \in B^*_\Theta$ and any $s_t \in \mathbb{T}_t$, then $\mathbf{T}_{t+1}(s_t, \theta) \in \mathbb{T}_{t+1}$;*

2. *The sets $\mathbb{T}_t$ contain a neighborhood of s^*_t and have bounded diameter; more precisely, there exist $r'''_{\mathcal{S}} > 0$ and $r''''_{\mathcal{S}} > 0$ such that $B_{\mathcal{S}_t}(s^*_t, r''''_{\mathcal{S}}) \subset \mathbb{T}_t \subset B_{\mathcal{S}_t}(s^*_t, r'''_{\mathcal{S}})$ for all $t \geq 0$.*

3. *$r'''_\Theta \leq \min(r_\Theta, r'_\Theta, r''_\Theta)$ and likewise for $r'''_{\mathcal{S}}$, so that inside B^*_Θ and $\mathbb{T}_t$, all assumptions of Section 2.4 as well as all results 6.12–6.17 apply (with $t_2 \leq t_1 + h$ for Corollary 6.14).*

Proof. Let h be the horizon for the spectral radius in Assumption 2.13.

Let ε_θ and $\varepsilon_{\mathcal{S}}$ be small enough, to be determined later. Let (θ_t) be a sequence of parameters with $\|\theta_t - \theta^*\| \leq \varepsilon_\theta$ and let $s \in \mathcal{S}_t$ with $\|s - s^*_t\| \leq \varepsilon_{\mathcal{S}}$.

By Corollary 6.14, for all $t \leq t' \leq t+h$ one has

$$\|\mathbf{T}_{t:t'}(s, \theta^*) - \mathbf{T}_{t:t'}(s^*_t, \theta^*)\| \leq \kappa_7^h\, \varepsilon_{\mathcal{S}}$$

provided we take $\varepsilon_\theta \leq \varepsilon_{\mathcal{S}}$ small enough so that the assumption of Corollary 6.14 is met.

Take $\varepsilon_{\mathcal{S}}$ smaller than $r''_{\mathcal{S}}$ from Lemma 6.17. Then we can apply Lemma 6.17 to obtain

$$\|\mathbf{T}_{t:t+h}(s, \theta^*) - \mathbf{T}_{t:t+h}(s^*_t, \theta^*)\| \leq (1-\alpha/2)\, \|s - s^*_t\|\,.$$

Now we have

$$\begin{aligned}\|\mathbf{T}_{t:t+h}(s, (\theta_t)) - s^*_{t+h}\| &= \|\mathbf{T}_{t:t+h}(s, (\theta_t)) - \mathbf{T}_{t:t+h}(s^*_t, \theta^*)\| \\ &\leq \|\mathbf{T}_{t:t+h}(s, (\theta_t)) - \mathbf{T}_{t:t+h}(s, \theta^*)\| + \|\mathbf{T}_{t:t+h}(s, \theta^*) - \mathbf{T}_{t:t+h}(s^*_t, \theta^*)\| \\ &\leq \kappa_7^h\, \varepsilon_\theta + (1-\alpha/2)\, \|s - s^*_t\|\,,\end{aligned}$$

where the last inequality follows by applying Corollary 6.14 to $(s,(\theta_t))$ and (s,θ^*).

Therefore, if $\|s-s_t^*\| \le 2\kappa_7^h\,\varepsilon_\theta/\alpha$, then

$$\|\mathbf{T}_{t:t+h}(s,(\theta_t)) - s_{t+h}^*\| \le \kappa_7^h\,\varepsilon_\theta + (1-\alpha/2)2\kappa_7^h\,\varepsilon_\theta/\alpha = 2\kappa_7^h\,\varepsilon_\theta/\alpha$$

again. This means that the balls of radius $2\kappa_7^h\,\varepsilon_\theta/\alpha$ around s_t^* are stable by the application of k consecutive steps of the transition operator $\mathbf{T}$, using any sequence of parameters (θ_t) such that $\|\theta_t-\theta^*\| \le \varepsilon_\theta$.

So if we define $\varepsilon_{\mathcal{S}} = 2\kappa_7^h\,\varepsilon_\theta/\alpha$ (still subject to the constraints on ε_θ and $\varepsilon_{\mathcal{S}}$ above), by induction we obtain that if (θ_t) is any sequence of parameters with $\|\theta_t-\theta^*\| \le \varepsilon_\theta$, and $s \in \mathcal{S}_t$ with $\|s-s_t^*\| \le \varepsilon_{\mathcal{S}}$, then

$$\|\mathbf{T}_{t:t+nh}(s,(\theta_t)) - s_{t+nh}^*\| \le \varepsilon_{\mathcal{S}}$$

for all $n \ge 0$.

This establishes that iterates of an element of a ball of radius $\varepsilon_{\mathcal{S}}$ around s_t^*, stay in such a ball at times that are multiples of h.

For times in between multiples of h, write $nh \le t < nh+h$ and assume that $\|s_{nh}-s_{nh}^*\| \le \varepsilon_{\mathcal{S}}$. Then by Corollary 6.14, one has

$$\|\mathbf{T}_{nh:t}(s_{nh},(\theta_t)) - \mathbf{T}_{nh:t}(s_{nh}^*,\theta^*)\| \le \kappa_7^h\,\max(\varepsilon_{\mathcal{S}},\varepsilon_\theta)$$

which is bounded.

We can thus set $r_\Theta''' := \varepsilon_\theta$, $r_{\mathcal{S}}''' = \kappa_7^h\,\max(\varepsilon_{\mathcal{S}},\varepsilon_\theta)$ and $r_{\mathcal{S}}'''' = \varepsilon_{\mathcal{S}}$. We then set $B_\Theta^* := B_\Theta(\theta^*, r_\Theta''')$ and define, inductively for $t \ge 1$,

$$\mathbb{T}_t := \mathbf{T}_t(\mathbb{T}_{t-1}, B_\Theta^*) \cup B_{\mathcal{S}_t}(s_t^*,\varepsilon_{\mathcal{S}}), \qquad \mathbb{T}_0 := B_{\mathcal{S}_0}(s_0^*,\varepsilon_{\mathcal{S}})$$

so that the sets $\mathbb{T}_t$ are stable under $\mathbf{T}_t$ and contain a neighborhood of s_t^*.

Then every element of $\mathbb{T}_t$ is an iterate of an element of $B_{\mathcal{S}_{t'}}(s_{t'}^*,\varepsilon_{\mathcal{S}})$ for some $t' \le t$. Therefore, by the above, $\mathbb{T}_t$ is contained in a ball of radius $r_{\mathcal{S}}'''$ around s_t^*. □

Corollary 6.19 (Forgetting of states with a fixed parameter)**.** *Let h be the horizon for the spectral radius in Assumption 2.13, and $1-\alpha$ the corresponding operator norm.*

Let $\theta \in B_\Theta^*$ and let $s, s' \in \mathbb{T}_t$.

Then there exists a constant $\kappa_8 \ge 0$ such that, for any $t' \ge t$,

$$\|\mathbf{T}_{t:t'}(s,\theta) - \mathbf{T}_{t:t'}(s',\theta)\| \le \kappa_8\,(1-\alpha/2)^{(t'-t)/h}\,\|s-s'\|\,.$$

Proof. Write $t'-t = r+nh$ with $r<h$. By Lemma 6.18, Corollary 6.14 can be applied inside $\mathbb{T}_t$ provided $t_2 \le t+h$. With $t_2 = t+r$, this yields

$$\|\mathbf{T}_{t:t+r}(s,\theta) - \mathbf{T}_{t:t+r}(s',\theta)\| \le \kappa_7^h\,\|s-s'\|\,.$$

Then by induction from Lemma 6.17 (whose assumptions are satisfied in $\mathbb{T}_{t+r}$, since $r_{\mathcal{S}}''' \le r_{\mathcal{S}}''$, according to Lemma 6.18), we obtain

$$\|\mathbf{T}_{t:t+r+nh}(s,\theta) - \mathbf{T}_{t:t+r+nh}(s',\theta)\| \le \kappa_7^h(1-\alpha/2)^n\,\|s-s'\|\,,$$

from which the conclusion follows by setting $\kappa_8 = \kappa_7^h/(1-\alpha/2)$ where the factor $1/(1-\alpha/2)$ accounts for the rounding in the division $(t'-t)/h$. □

6.6.2 Existence of a Stable Tube for the Jacobians J_t and $\tilde{J}_t$

Remark 6.20. *Let $A\colon \mathcal{S}_t \to \mathcal{S}_{t+1}$ be a linear operator. Equip $\mathrm{L}(\Theta, \mathcal{S}_t)$ with the operator norm. Then the operator norm of A acting on $\mathrm{L}(\Theta, \mathcal{S}_t)$ via $J \in \mathrm{L}(\Theta, \mathcal{S}_t) \mapsto AJ \in \mathrm{L}(\Theta, \mathcal{S}_{t+1})$ is the same as the operator norm of A acting on $\mathcal{S}_t$.*

Lemma 6.21. *Let $(A_t)_{t\geq 1}$ be a sequence of linear operators on normed vector spaces, with spectral radius at most $1-\alpha$ at horizon h. Assume the A_t's have operator norm at most ρ.*

Let $(J_t)_{t\geq 0}$ and $(J'_t)_{t\geq 0}$ be two sequences of elements of the spaces on which the A_t's act, and suppose that

$$J_t = A_t J_{t-1} + B_t, \qquad J'_t = A_t J'_{t-1} + B'_t$$

for some B_t and B'_t. Then for any $0 \leq t_1 \leq t_2$,

$$\|J_{t_2} - J'_{t_2}\| \leq \frac{\max(1,\rho^h)}{1-\alpha} \max\left((1-\alpha)^{\frac{t_2-t_1}{h}} \|J_{t_1} - J'_{t_1}\|, \sup_{t_1\leq t\leq t_2} (1-\alpha)^{\frac{t_2-t}{h}} \|B_t - B'_t\| \right).$$

Proof. By induction we have

$$J_{t_2} = A_{t_2} A_{t_2-1} \cdots A_{t_1+1} J_{t_1} + \sum_{t=t_1+1}^{t_2} A_{t_2} A_{t_2-1} \cdots A_{t+1} B_t$$

and likewise for J' and thus also for $J - J'$.

Now, for any $t_1 \leq t_2$ the product of the operators $A_{t_2} A_{t_2-1} \cdots A_{t_1+1}$ has operator norm at most

$$\|A_{t_2} A_{t_2-1} \cdots A_{t_1+1}\|_{\mathrm{op}} \leq \max(1,\rho^h)(1-\alpha)^{\frac{t_2-t_1}{h}-1}.$$

Indeed, we can decompose $t_2 - t_1 = r + nh$ with $r < h$, and the product of the first r factors has operator norm at most ρ^r (and $\rho^r \leq \max(1,\rho)^h$ because $r \leq h$ and the max accounts for whether ρ is larger than 1 or not). Finally, the product of the remaining nh factors has operator norm at most $(1-\alpha)^n$, and $n \geq (t_2-t_1)/h - 1$. □

Proposition 6.22. *Let $0 \leq \alpha \leq 1$, $\rho > 0$, and $h \geq 0$. Then there exists $\varepsilon > 0$ with the following property.*

Let $(A_t)_{t\geq 1}$ be any sequence of linear operators on normed vector spaces, with operator norm bounded by ρ, and with spectral radius at most $1-\alpha$ at horizon h.

Let $c \geq 0$. Let $(J_t)_{t\geq 0}$ be any sequence of elements of the spaces on which the A_t's act, such that

$$J_t = A_t J_{t-1} + E_t$$

with $\|E_t\| \leq c + \varepsilon \|J_{t-1}\|$.

Then $\|J_t\|$ is bounded when $t \to \infty$. More precisely, there exist constants a and b such that for any $t \geq 0$ and any $t' \geq t$, $\|J_{t'}\| \leq a\|J_t\| + b$.

Moreover, the coefficient a depends on ρ, α and h, while b depends on ρ, α, h and c.

Proof. Up to increasing ρ, we can assume $\rho \geq 1$.

By induction, for $t' \geq t$ one finds

$$\|J_{t'}\| \leq (\rho+\varepsilon)^{t'-t} \|J_t\| + (t'-t)\rho^{t'-t} c. \tag{17}$$

Moreover, by induction, for $t' \geq t$,

$$J_{t'} = A_{t'} A_{t'-1} \cdots A_{t+1} J_t + \sum_{s=1}^{t'-t} A_{t'} A_{t'-1} \cdots A_{t+s+1} E_{t+s}.$$

Taking $t' = t + h$, using the spectral radius property, then substituting (17), one finds

$$\begin{aligned}
\|J_{t+h}\| &\leq (1-\alpha)\|J_t\| + \sum_{s=1}^{h} \rho^{h-s} \|E_{t+s}\| \\
&\leq (1-\alpha)\|J_t\| + \sum_{s=1}^{h} \rho^{h-s}(c + \varepsilon \|J_{t+s-1}\|) \\
&\leq (1-\alpha)\|J_t\| + \sum_{s=1}^{h} \rho^{h-s}\left(c + \varepsilon(\rho+\varepsilon)^{s-1}\|J_t\| + \varepsilon(s-1)\rho^{s-1}c\right) \\
&= \left(1 - \alpha + \varepsilon \sum_{s=1}^{h} \rho^{h-s}(\rho+\varepsilon)^{s-1}\right)\|J_t\| + \sum_{s=1}^{h} \rho^{h-s}\left(c + \varepsilon(s-1)\rho^{s-1}c\right).
\end{aligned}$$

Since h and ρ are fixed, by taking ε small enough one can ensure that $1 - \alpha + \varepsilon \sum_{s=1}^{h} \rho^{h-s}(\rho+\varepsilon)^{s-1}$ is less than 1. Moreover, the term $\sum_{s=1}^{h} \rho^{h-s}\left(c + \varepsilon(s-1)\rho^{s-1}c\right)$ does not depend on t, so is bounded when $t \to \infty$.

It results that if $t' = t + nh$ for some $n \geq 0$, then $\|J_{t'}\|$ is bounded by $a\|J_t\| + b$ for some constants a and b.

For $t' - t$ not a multiple of h, write $t' = t + nh + r$ with $r < h$. Then by (17), we have

$$\|J_{t'}\| \leq (\rho+\varepsilon)^r \|J_{t+nh}\| + r\rho^r c$$

which is bounded as well, hence the conclusion. □

Corollary 6.23 ($\tilde{J}$ is bounded for imperfect RTRL algorithms.)**.** *Let (θ_t) and (s_t) be sequences of parameters and states with $\theta_t \in B_\Theta^*$ and $s_t \in \mathbb{T}_t$ for all $t \geq 0$.*

Consider a sequence $(\tilde{J}_t)_{t \geq t_0}$ computed as in an imperfect RTRL algorithm (Definition 2.10) starting at time t_0, namely

$$\tilde{J}_t = \frac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial s} \tilde{J}_{t-1} + \frac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial \theta} + E_t, \qquad \tilde{J}_{t_0} \in \mathrm{L}(\Theta, \mathcal{S}_{t_0})$$

where $E_t \in \mathrm{L}(\Theta, \mathcal{S}_t)$ satisfies

$$\|E_t\|_{\mathrm{op}} \leq \phi\left(\left\|\tilde{J}_{t-1}\right\|_{\mathrm{op}}, \left\|\frac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial(s,\theta)}\right\|_{\mathrm{op}}\right)$$

for some gauge error ϕ.

Then the sequence $(\tilde{J}_t)$ is bounded. More precisely, there exist constants a and b such that for any such sequence $(\tilde{J}_t)$, for any $t \geq t_0$ and any $t' \geq t$, $\left\|\tilde{J}_{t'}\right\| \leq a\left\|\tilde{J}_t\right\| + b$.

Proof. By Lemma 6.18, the stable tubes B_Θ^* and $\mathbb{T}_t$ are included in balls on which all the results up to Lemma 6.18 hold.

So by Lemma 6.12, the operator norm $\left\|\frac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial(s,\theta)}\right\|_{\mathrm{op}}$ is bounded by some $\rho \geq 0$. Therefore, for all $t \geq 1$, we have

$$\|E_t\|_{\mathrm{op}} \leq \sup_{|y| \leq \rho} \phi\left(\left\|\tilde{J}_{t-1}\right\|_{\mathrm{op}}, y\right),$$

and likewise

$$\left\| \frac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial \theta} + E_t \right\|_{\mathrm{op}} \leq \rho + \sup_{|y| \leq \rho} \phi\left(\left\| \tilde{J}_{t-1} \right\|_{\mathrm{op}}, y \right)$$

since $\left\| \frac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial \theta} \right\|_{\mathrm{op}} \leq \left\| \frac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial (s, \theta)} \right\|_{\mathrm{op}}$.

By Corollary 6.16, the sequence of operators $\left(\frac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial s} \right)_{t \geq 0}$ has spectral radius at most $1 - \alpha/2$ at horizon h. Consider the value $\varepsilon > 0$ provided by Proposition 6.22 for this sequence of operators.

Since ϕ is an error gauge, thanks to the two properties of Definition 2.19, we can find a constant $c = c(\rho, \phi) \geq 0$ such that $\rho + \sup_{|y| \leq \rho} \phi\left(\left\| \tilde{J}_{t-1} \right\|_{\mathrm{op}}, y \right) \leq c + \varepsilon \left\| \tilde{J}_{t-1} \right\|_{\mathrm{op}}$. Therefore we can apply Proposition 6.22 which yields the conclusion. □

Corollary 6.24 (Stable tubes for RTRL and imperfect RTRL algorithms)**.** *The RTRL algorithm and all imperfect RTRL algorithms with errors controlled by the same error gauge ϕ (Assumption 2.21) admit a common stable tube on $(s_t, \tilde{J}_t)$.*

More precisely, with $(\mathbb{T}_t)_{t \geq 0}$ the stable tube for (s_t) alone from Lemma 6.18, we can find sets $(\mathbb{T}_t^J)_{t \geq 0}$ (depending on the error gauge ϕ) such that $(\mathbb{T}_t \times \mathbb{T}_t^J)_{t \geq 0}$ is a stable tube for the pair (s_t, J_t) of RTRL, as well as for the pair $(s_t, \tilde{J}_t)$ of any imperfect RTRL algorithm with error gauge ϕ. Namely, if $\theta_t \in B_\Theta^*$ and $(s_t, J_t) \in \mathbb{T}_t \times \mathbb{T}_t^J$, then (s_{t+1}, J_{t+1}) in $\mathbb{T}_{t+1} \times \mathbb{T}_{t+1}^J$ and likewise for $\tilde{J}$.

In particular, trajectories of the imperfect RTRL algorithms respect this stable tube in the sense of Definition 4.23.

Moreover, the sets $(\mathbb{T}_t^J)_{t \geq 0}$ are bounded, and for every t, $\mathbb{T}_t^J$ contains a ball around 0, whose radius does not depend on t.

Therefore, Assumption 4.11 is satisfied, with the same stable tube, for the RTRL algorithm and any imperfect RTRL algorithm which admits ϕ as an error gauge.

Proof. First, the RTRL algorithm is also an imperfect RTRL algorithm where errors $E_t = 0$ vanish, and therefore it satisfies Assumption 2.21 (for any error gauge); thus, the previous results of this section for imperfect RTRL algorithms also apply to RTRL.

We have already established the existence of stable tubes for the state s_t, defined by B_Θ^* and $\mathbb{T}_t$ (for which J and $\tilde{J}$ play no role).

Let $\tilde{\mathcal{A}}_t$ be the transition function of the imperfect RTRL algorithm, which computes $(s_t, \tilde{J}_t)$ from θ and $(s_{t-1}, \tilde{J}_{t-1})$. Let $\tilde{\mathcal{A}}_t^{\tilde{J}}$ be the part of this function that just returns $\tilde{J}_t$.

Let J_t^* be the value of J_t computed by RTRL along the target trajectory s_t^* defined by θ^*, namely, $J_t^* = \frac{\partial \mathbf{T}_t(s_{t-1}^*, \theta^*)}{\partial s} J_{t-1}^* + \frac{\partial \mathbf{T}_t(s_{t-1}^*, \theta^*)}{\partial \theta}$ initialized with $J_0^* = 0$. By Corollary 6.23, $\|J_t^*\|_{\mathrm{op}}$ is bounded by some value r_J^*.

Let $r_J > 0$ be any positive value. Define the stable tubes over J by taking, at each step, the image of the previous values under any imperfect RTRL algorithm and adjoining a ball of fixed radius around J_t^*, namely, define the set

$$\mathbb{T}_0^J := \{ \tilde{J}_0 \colon \left\| \tilde{J}_0 - J_0^* \right\|_{\mathrm{op}} \leq r_J \}$$

and then, by induction over $t \geq 1$, define $\mathbb{T}_t^J$ to be the union of a ball around J_t^* and of all possible values $\tilde{J}_t$ obtained from a value $\tilde{J}_{t-1}$ in $\mathbb{T}_{t-1}^J$ by a transition that

respects the assumptions for imperfect RTRL algorithms; namely,

$$\mathbb{T}_t^J := \{\tilde{J}_t \colon \left\|\tilde{J}_t - J_t^*\right\|_{\text{op}} \leq r_J\}$$
$$\cup \left\{ \frac{\partial \mathbf{T}_t(s,\theta)}{\partial s}\, \tilde{J}_{t-1} + \frac{\partial \mathbf{T}_t(s,\theta)}{\partial \theta} + E_t, \qquad s \in \mathbb{T}_{t-1},\ \theta \in B_\Theta^*,\ \tilde{J}_{t-1} \in \mathbb{T}_{t-1}^J, \right.$$
$$\left. \|E_t\|_{\text{op}} \leq \phi\left(\left\|\tilde{J}_{t-1}\right\|_{\text{op}}, \left\| \frac{\partial \mathbf{T}_t(s,\theta)}{\partial(s,\theta)} \right\|_{\text{op}} \right) \right\}.$$

By construction of $\mathbb{T}_t^J$, any element of $\mathbb{T}_t^J$ is the iterate by some imperfect RTRL algorithm, of an element of the ball $\{\tilde{J}_{t_0} \colon \left\|\tilde{J}_{t_0} - J_{t_0}^*\right\|_{\text{op}} \leq r_J\}$ for some $t_0 \leq t$, where the algorithm is used at a sequence of states s_t and parameters θ_t in the stable tube.

Since $\|J_t^*\|_{\text{op}}$ is bounded by r_J^*, the elements of the ball $\{\tilde{J}_{t_0} \colon \left\|\tilde{J}_{t_0} - J_{t_0}^*\right\|_{\text{op}} \leq r_J\}$ are bounded by $r_J^* + r_J$. By construction of $\mathbb{T}_t^J$, any sequence $\tilde{J}_t \in \mathbb{T}_t^J$ respects the assumptions of Corollary 6.23. Therefore, by Corollary 6.23, any such sequence $\tilde{J}_t$ is bounded by $a(r_J^* + r_J) + b$. So the sets $\mathbb{T}_t^J$ are bounded.

Besides, $\mathbb{T}_t^J$ contains a ball around J_t^* by definition. Therefore, the sets $\mathbb{T}_t \times \mathbb{T}_t^J$ constitute a stable tube for the pair $(s_t, \tilde{J}_t)$.

Finally, since $\mathbb{T}_t^J$ contains a ball of radius r_J around J_t^*, and since $\|J_t^*\| \leq r_J^*$, if we choose $r_J > r_J^*$, we ensure that $\mathbb{T}_t^J$ contains a ball around 0. □

Remark 6.25 (Taylor expansions in the stable tube)**.** *When we say "in the stable tube" in the following, it means that both the state s and the Jacobian J considered belong to their respective stable tubes, and that the parameter θ considered belongs to B_Θ^*. Therefore, all the assumptions of Section 2 hold "on the stable tube"; more precisely, on balls containing the stable tube.*

In the following, we shall repeatedly use the following argument: "Since s and s' are in the stable tube, and since $\partial_s f$ is bounded on the stable tube by one of the assumptions of Section 2, then $\|f(s) - f(s')\| = O(\|s - s'\|)$." However, the stable tube is not convex, so integrating $\partial_s f$ on the segment from s to s' may exit the stable tube. Still, every assumption of Section 2 holds on a ball containing the stable tube; these balls are convex so the argument is valid. We will implicitly use this argument and just say "the derivative of f is bounded on the stable tube."

6.7 Lipschitz-Type Properties of the Transition Operator of RTRL

Remember that $\mathbb{T}_t$ and $\mathbb{T}_t^J$ are the stable tubes for s_t and J_t introduced in Lemma 6.18 and Corollary 6.24.

Lemma 6.26 (Exponential forgetting of (s, J) for RTRL with fixed θ)**.** *Let $\mathcal{A}$ be the RTRL algorithm on (s, J) from Definition 6.1.*

Then there exists a constant κ_9 such that the following holds. For any $t_0 \geq 0$, for any $\theta \in B_\Theta^*$, for any $s, s' \in \mathbb{T}_{t_0}$ and $J, J' \in \mathbb{T}_{t_0}^J$, for any $t \geq t_0$,

$$\left\|\mathcal{A}_{t_0:t}(\theta,(s,J)) - \mathcal{A}_{t_0:t}(\theta,(s',J'))\right\| \leq \kappa_9 \left(1-\alpha/2\right)^{\frac{t-t_0}{h}} \max(\left\|s-s'\right\|, \left\|J-J'\right\|)$$

and in particular, Assumption 4.14 is satisfied for the RTRL algorithm.

Proof. Define the trajectory $(s_t, J_t) = \mathcal{A}_{t_0:t}(\theta, (s, J))$ and likewise for (s', J'). Thus, we have to bound $\|(s_t, J_t) - (s_t', J_t')\|$.

The norm on pairs (s, J) is the max-norm (Definition 6.1), so it is enough to prove the statement separately for $\|s_t - s'_t\|$ and $\|J_t - J'_t\|$.

By Definition 6.1 of the RTRL algorithm, we have $s_t = \mathbf{T}_{t_0:t}(s, \theta)$, so the conclusion for s is exactly Corollary 6.19.

By Definition 6.1, the sequence (J_t) satisfies

$$J_t = \frac{\partial \mathbf{T}_t(s_{t-1}, \theta)}{\partial s} J_{t-1} + \frac{\partial \mathbf{T}_t(s_{t-1}, \theta)}{\partial \theta}$$

for $t \geq t_0 + 1$, and likewise for J'. Denote

$$A_t := \frac{\partial \mathbf{T}_t(s_{t-1}, \theta)}{\partial s}, \qquad B_t := \frac{\partial \mathbf{T}_t(s_{t-1}, \theta)}{\partial \theta}$$

and likewise for J'. Therefore, we have $J_t = A_t J_{t-1} + B_t$ and

$$J'_t = A'_t J'_{t-1} + B' = A_t J'_{t-1} + B'_t + (A'_t - A_t) J'_{t-1}.$$

Since the trajectories lie in the stable tube, by Corollary 6.16, the sequence (A_t) has spectral radius at most $1 - \alpha/2$ at horizon h.

In order to apply Lemma 6.21 to J_t and J'_t, we have to bound

$$B_t - B'_t - (A'_t - A_t) J'_{t-1}$$

for each t.

By definition $B_t = \frac{\partial \mathbf{T}_t(s_{t-1}, \theta)}{\partial \theta}$. By Assumption 2.23, the second derivative $\frac{\partial^2 \mathbf{T}_t(s_{t-1}, \theta)}{\partial (s, \theta)^2}$ is bounded on the stable tube. This implies that $\frac{\partial \mathbf{T}_t(s_{t-1}, \theta)}{\partial \theta}$ is a Lipschitz function of s_{t-1} on the stable tube, and in particular

$$B_t - B'_t = O(\|s_t - s'_t\|).$$

The same reasoning applies to $A_t = \frac{\partial \mathbf{T}_t(s_{t-1}, \theta)}{\partial s}$), so that

$$A_t - A'_t = O(\|s_t - s'_t\|)$$

and the sequence (J'_t) belongs to the stable tube, so by Corollary 6.24, (J'_t) is bounded. Consequently,

$$(A'_t - A_t) J'_t = O(\|s_t - s'_t\|)$$

as well, and therefore

$$B_t - B'_t - (A'_t - A_t) J'_{t-1} = O(\|s_t - s'_t\|).$$

So Lemma 6.21 applied to the sequences J_t and J'_t provides

$$\|J_t - J'_t\| \leq O\left(\max\left((1 - \alpha/2)^{\frac{t-t_0}{h}} \|J_{t_0} - J'_{t_0}\|, \sup_{t_0 \leq t' \leq t} (1 - \alpha/2)^{\frac{t-t'}{h}} \|s_{t'} - s'_{t'}\|\right)\right).$$

But by Corollary 6.19,

$$\|s_t - s'_t\| \leq \kappa_8 (1 - \alpha/2)^{\frac{t-t_0}{h}} \|s_{t_0} - s'_{t_0}\|$$

so that for $t_0 \leq t' \leq t$, one has

$$\begin{aligned}(1-\alpha/2)^{\frac{t-t'}{h}} \left\|s_{t'} - s'_{t'}\right\| &\leq \kappa_8 (1-\alpha/2)^{\frac{t-t'}{h}} (1-\alpha/2)^{\frac{t'-t_0}{h}} \left\|s_{t_0} - s'_{t_0}\right\| \\ &= \kappa_8 (1-\alpha/2)^{\frac{t-t_0}{h}} \left\|s_{t_0} - s'_{t_0}\right\|\end{aligned}$$

and therefore

$$\left\|J_t - J'_t\right\| \leq O\left((1-\alpha/2)^{\frac{t-t_0}{h}} \max\left(\left\|J_{t_0} - J'_{t_0}\right\|, \left\|s_{t_0} - s'_{t_0}\right\|\right)\right)$$

as needed.

□

Corollary 6.27 (RTRL is Lipschitz wrt θ)**.** *Let $\mathcal{A}$ be the RTRL algorithm on (s, J) from Definition 6.1. Then $\mathcal{A}$ satisfies Assumption 4.13. Namely, for any $(s_t, J_t) \in \mathbb{T}_t \times \mathbb{T}_t^J$ and any $\theta, \theta' \in B_\Theta^*$, one has*

$$\left\|\mathcal{A}_{t+1}(\theta, (s_t, J_t)) - \mathcal{A}_{t+1}(\theta', (s_t, J_t))\right\| \leq \kappa_{\text{lip}\theta} \left\|\theta - \theta'\right\|$$

for some $\kappa_{\text{lip}\theta} \geq 0$.

Proof. By definition of $\mathcal{A}$, we have $\mathcal{A}_{t+1}(\theta, (s_t, J_t)) = (s_{t+1}, J_{t+1})$ where $s_{t+1} = \mathbf{T}_{t+1}(s_t, \theta)$ and

$$J_{t+1} = \frac{\partial \mathbf{T}_{t+1}(s_t, \theta)}{\partial s} J_t + \frac{\partial \mathbf{T}_{t+1}(s_t, \theta)}{\partial \theta}.$$

Now, by Lemma 6.12, $\frac{\partial \mathbf{T}_{t+1}(s_t, \theta)}{\partial \theta}$ is bounded on the stable tube, so that $\mathbf{T}_{t+1}$ is Lipschitz with respect to θ on the stable tube. This shows that s_{t+1} depends on θ in a Lipschitz way. Since the bound in Lemma 6.12 is uniform in t, the Lipschitz constant is uniform in t.

For J_{t+1}, by Assumption 2.23, the second derivative $\left\|\frac{\partial^2 \mathbf{T}_{t+1}}{\partial (s,\theta)^2}(s, \theta)\right\|_{\text{op}}$ is bounded on the stable tube. This implies that both $\frac{\partial \mathbf{T}_{t+1}(s_t, \theta)}{\partial \theta}$ and $\frac{\partial \mathbf{T}_{t+1}(s_t, \theta)}{\partial s}$ are Lipschitz with respect to θ in the stable tube. Since J_t is bounded in the stable tube, this implies that $J_{t+1} = \frac{\partial \mathbf{T}_{t+1}(s_t, \theta)}{\partial s} J_t + \frac{\partial \mathbf{T}_{t+1}(s_t, \theta)}{\partial \theta}$ is Lipschitz with respect to θ. Since the bound in Assumption 2.23 is uniform in t, the Lipschitz constant is uniform in t. □

6.8 Boundedness of Gradients for RTRL

Lemma 6.28 (Boundedness of gradients for RTRL)**.** *The gradient computation operators $\boldsymbol{V}_t$ from Definition 6.1 satisfy Assumption 4.15 and Assumption 4.16; the scale functions $m(t)$ and $m_{\text{H}}(t)$ appearing in these assumptions are both equal to the scale function $m(t)$ appearing in Assumption 2.24.*

Consistently with Assumption 4.15, we denote $B_{\mathcal{V}_t}$ the ball of $\mathcal{V}_t$ with radius $\sup_{\theta \in B_\Theta^*} \sup_{\mathfrak{m} \in \mathbb{T}_{\mathcal{M}_t}} \left\|\boldsymbol{V}_t(\theta, \mathfrak{m})\right\|$.

Proof. From Definition 6.1,

$$\boldsymbol{V}_t(\theta, (s, J)) = \mathcal{U}_t\left(\frac{\partial \mathcal{L}_t(s)}{\partial s} \cdot J,\ s,\ \theta\right)$$

and we have to prove that this quantity is bounded by $O(m(t))$ on the stable tube (Assumption 4.15) and is $O(m(t))$-Lipschitz with respect to $(\theta, (s, J))$ on the stable tube (Assumption 4.16).

Let us first study the auxiliary quantity

$$\boldsymbol{V}_t^1(s,\, J) = \frac{\partial \mathcal{L}_t(s)}{\partial s} \cdot J.$$

By Assumption 2.24, the second derivative of $\mathcal{L}_t$ with respect to s is $O\left(m(t)\right)$ on the stable tube. This proves that $\frac{\partial \mathcal{L}_t(s)}{\partial s}$ is $O\left(m(t)\right)$–Lipschitz with respect to s on the stable tube. Since J is bounded on the stable tube, it follows that $\boldsymbol{V}_t^1(s, J)$ is $O\left(m(t)\right)$–Lipschitz with respect to s on the stable tube.

Since $\frac{\partial \mathcal{L}_t(s)}{\partial s}$ is Lipschitz with respect to s on the stable tube, and its values at (s_t^*) are bounded by $O(m(t))$ by Assumption 2.24, and since the stable tube is bounded around (s_t^*), it follows that $\frac{\partial \mathcal{L}_t(s)}{\partial s}$ is bounded by $O(m(t))$ on the stable tube.

This proves, first, that $\boldsymbol{V}_t^1(s, J))$ is $O(m(t))$-Lipschitz with respect to J on the stable tube. Second, since J is bounded on the stable tube, this proves that $\boldsymbol{V}_t^1(s, J)$ is bounded by $O(m(t))$ on the stable tube.

Thus, $\boldsymbol{V}_t^1(s, J)$ is $O(m(t))$-Lipschitz with respect to $(s,\, J)$ on the stable tube.

Let us now consider $(\theta,\, (s,\, J))$ and $(\theta',\, (s',\, J'))$ on the stable tube. Let $t \geq 0$. Then, the difference $\boldsymbol{V}_t\left(\theta,\, (s,\, J)\right) - \boldsymbol{V}_t\left(\theta',\, (s',\, J')\right)$ equals

$$\begin{aligned}
&\mathcal{U}_t\left(\frac{\partial \mathcal{L}_t(s)}{\partial s} \cdot J,\, s,\, \theta\right) - \mathcal{U}_t\left(\frac{\partial \mathcal{L}_t}{\partial s}\left(s'\right) \cdot J',\, s',\, \theta'\right) \\
&\qquad = \mathcal{U}_t\left(\frac{\partial \mathcal{L}_t(s)}{\partial s} \cdot J,\, s,\, \theta\right) - \mathcal{U}_t\left(\frac{\partial \mathcal{L}_t}{\partial s}\left(s\right) \cdot J,\, s',\, \theta'\right) \\
&\qquad\qquad + \mathcal{U}_t\left(\frac{\partial \mathcal{L}_t}{\partial s}\left(s\right) \cdot J,\, s',\, \theta'\right) - \mathcal{U}_t\left(\frac{\partial \mathcal{L}_t}{\partial s}\left(s'\right) \cdot J',\, s',\, \theta'\right).
\end{aligned}$$

As a result, the difference $\boldsymbol{V}_t\left(\theta,\, (s,\, J)\right) - \boldsymbol{V}_t\left(\theta',\, (s',\, J')\right)$ is bounded by (writing B_t^J a ball containing $\mathbb{T}_t^J$, which radius we can chose independant of t since the stable tube is bounded)

$$\begin{aligned}
&\sup_{B_\Theta^* \times B_{\mathcal{S}_t}(s_t^*, r_{\mathcal{S}}) \times B_t^J} \left(\left\| \frac{\partial \mathcal{U}_t}{\partial\left(s,\, \theta\right)}\left(\boldsymbol{V}_t^1(s'',\, J''),\, s'',\, \theta''\right)\right\|_{\text{op}}\right) \left\|(s,\, \theta) - (s',\, \theta')\right\| \\
&+ \sup_{B_\Theta^* \times B_{\mathcal{S}_t}(s_t^*, r_{\mathcal{S}}) \times B_t^J} \left(\left\| \frac{\partial \mathcal{U}_t}{\partial v}\left(\boldsymbol{V}_t^1(s'',\, J''),\, s'',\, \theta''\right)\right\|_{\text{op}}\right) \left\|\boldsymbol{V}_t^1(s,\, J) - \boldsymbol{V}_t^1(s',\, J')\right\|.
\end{aligned}$$

Thanks to Assumption 2.14, we therefore know that the difference is bounded by

$$\begin{aligned}
&\sup_{B_\Theta^* \times B_{\mathcal{S}_t}(s_t^*, r_{\mathcal{S}}) \times B_t^J} O\left(1 + \left\|\boldsymbol{V}_t^1(s'',\, J'')\right\|\right) \left\|(s,\, \theta) - (s',\, \theta')\right\| \\
&+ \sup_{B_\Theta^* \times B_{\mathcal{S}_t}(s_t^*, r_{\mathcal{S}}) \times B_t^J} \left(\left\| \frac{\partial \mathcal{U}_t}{\partial v}\left(\boldsymbol{V}_t^1(s'',\, J''),\, s'',\, \theta''\right)\right\|_{\text{op}}\right) \left\|\boldsymbol{V}_t^1(s,\, J) - \boldsymbol{V}_t^1(s',\, J')\right\| \\
&= O\left(m(t)\right) \left\|(s,\, \theta) - (s',\, \theta')\right\| + \sup_{B_\Theta^* \times B_{\mathcal{S}_t}(s_t^*, r_{\mathcal{S}}) \times B_t^J} \left(\left\| \frac{\partial \mathcal{U}_t}{\partial v}\left(\boldsymbol{V}_t^1(s'',\, J''),\, s'',\, \theta''\right)\right\|_{\text{op}}\right) \times \\
&\qquad\qquad O\left(m(t)\right) \left\|(s,\, J) - (s',\, J')\right\|.
\end{aligned}$$

Thanks to Assumption 2.14, the derivative $\frac{\partial \mathcal{U}_t}{\partial v}$ is bounded on a ball which contains the stable tube. This shows that $\boldsymbol{V}_t$ is $O\left(m(t)\right)$-Lipschitz with respect to θ, s and J on the stable tube.

Now, for every $t \geq 0$, for every $(\theta, (s, J))$ in the stable tube, we have

$$\boldsymbol{V}_t(\theta,\,(s,\,J)) = \boldsymbol{V}_t(\theta,\,(s,\,J)) - \boldsymbol{V}_t(\theta^*,\,(s_t^*,\,0)) + \boldsymbol{V}_t(\theta^*,\,(s_t^*,\,0))$$

and the last term is $\boldsymbol{V}_t(\theta^*,\,(s_t^*,\,0)) = \mathcal{U}_t(0,\,s_t^*,\,\theta^*)$ by definition.

Thanks to Assumption 2.14, this last term is $O(m(t))$. Finally, since $\|\theta - \theta^*\|$, $\|s - s_t^*\|$ and $\|J\|$ are bounded on the stable tube, and since $\boldsymbol{V}_t$ is $O(m(t))$-Lipschitz with respect to θ, s and J on the stable tube, the difference $\boldsymbol{V}_t(\theta,\,(s,\,J)) - \boldsymbol{V}_t(\theta^*,\,(s_t^*,\,0))$ is $O(m(t))$ on the stable tube. As a result, $\boldsymbol{V}_t(\theta,\,(s,\,J))$ is $O(m(t))$ on the stable tube. □

7 Proving Convergence of the RTRL Algorithm and of Imperfect RTRL Algorithms

Let us go on with applying the results of Section 4 under the assumptions of Section 2. So far, we have shown that imperfect RTRL algorithms admit a stable tube, thus satisfying Assumption 4.11, that the exponential forgetting and Lipschitz Assumptions 4.14 and 4.13 are satisfied, that the gradient computation operators are bounded on the stable tube, and Lipschitz, so that Assumptions 4.15 and 4.16 are satisfied, and that the sequence of learning rates satisfies Assumption 4.19. This leaves Assumption 4.17 and Assumption 4.18.

We now check that the local optimality Assumption 4.18 is valid for the RTRL algorithm. This requires to check that the update operator leaves θ^* almost unchanged over time intervals $[t; t + L(t)]$, and brings other parameters closer to θ^* over these intervals. This follows from Assumption 2.11.b which states that θ^* is a critical point of the average loss function, asymptotically; for this we have to transfer this asymptotic property to finite but long enough intervals $[t; t + L(t)]$. This is the object of Sections 7.2 and 7.3.

We also show that the parameter update operator satisfies Assumption 4.17 in Section 7.1, and that imperfect RTRL algorithms have negligible noise, in the sense of Definition 4.25, in Section 7.4.

7.1 Parameter Updates at First Order in η

Here we study the parameter update operator and its iterates at first order in the step size.

Proposition 7.1 (Checking conformity of the parameter update operator)**.** *Assumption 2.17 implies Assumption 4.17, over balls of the same radius $r_{\mathcal{V}} = \tilde{r}_{\mathcal{V}}$.*

We then define $\overline{\eta}_{\mathcal{V}}$ as in Definition 5.3.

Proof. Let us check the conclusions of Assumption 4.17. The first point is trivial under Assumption 2.17. For the second point, under Assumption 2.17 we have

$$\begin{aligned}\Phi_t(\theta, v) - \Phi_t(\theta', v') &= \theta - \theta' - v + v' + \|v\|^2 \Phi_t^{(2)}(\theta, v) - \|v'\|^2 \Phi_t^{(2)}(\theta', v') \\ &= \theta - \theta' - (v - v') + \|v\|^2 (\Phi_t^{(2)}(\theta, v) - \Phi_t^{(2)}(\theta', v')) \\ &+ \left(\|v\|^2 - \|v'\|^2\right) \Phi_t^{(2)}(\theta', v')\end{aligned}$$

and let us control each term separately. Since $\|v\|^2$ is bounded for $v \in B_{\mathcal{V}_t}(0, \tilde{r}_{\mathcal{V}})$, and since $\Phi_t^{(2)}$ is Lipschitz, the norm of $\|v\|^2 (\Phi_t^{(2)}(\theta, v) - \Phi_t^{(2)}(\theta', v'))$ is controlled

by $\|v - v'\| + \|\theta - \theta'\|$. Finally, writing $\|v\|^2 - \|v'\|^2 = (\|v\| - \|v'\|)(\|v\| + \|v'\|)$, we see that the norm of $\left(\|v\|^2 - \|v'\|^2\right) \Phi_t^{(2)}(\theta', v')$ is controlled by $\|v\| - \|v'\|$ times the supremum of $(\|v\| + \|v'\|)\Phi_t^{(2)}(\theta', v')$ over the ball; since both $\|v\| + \|v'\|$ and $\Phi_t^{(2)}$ are uniformly bounded in the ball, the conclusion follows. □

Lemma 7.2 (Intervals of lengths $L(t)$)**.** *For any $\overline{\eta} \leq \overline{\eta}_{\mathcal{V}}$, for t large enough, we have $L(t) < T_t^{r_\Theta^*}$ (Def. 5.8).*

As a consequence, we can apply the abstract results to intervals of length $L(t)$.

Proof. This is the same argument as in Lemma 5.21, which actually applies to any interval $(t, t + L(t)]$, not only to intervals $(T_k, T_k + L(T_k)]$. □

Lemma 7.3 (Parameter updates at first order)**.** *Let $t_0 \geq 0$. Let $\theta_{t_0} \in B_\Theta^*$ with $d(\theta_{t_0}, \theta^*)$ at most $r_\Theta^*/3$, and let (v_t) be a gradient sequence with $v_t \in B_{\mathcal{V}_t}$ for all $t \geq t_0$ (where $B_{\mathcal{V}_t}$ is defined just after Lemma 6.28). Let (η_t) be a stepsize sequence with $\overline{\eta} \leq \overline{\eta}_{\mathcal{V}}$. Define by induction for $t > t_0$,*

$$\theta_t = \Phi_t(\theta_{t-1}, \eta_t v_t) = \Phi_{t_0:t}(\theta_{t_0}, (\eta_t v_t)_t)$$

Then for any $t_0 \leq t < t_0 + T_{t_0}^{r_\Theta^*}(\boldsymbol{\eta})$,

$$\theta_t = \theta_{t_0} - \sum_{s=t_0+1}^{t} \eta_s v_s + O\left(\sum_{s=t_0+1}^{t} \eta_s^2 \, m(s)^2\right).$$

Proof. By Lemma 5.9, the trajectory stays in the control ball for $t < T_{t_0}^{r_\Theta^*}(\boldsymbol{\eta})$.

By induction from Assumption 2.17 we find

$$\theta_t = \theta_0 - \sum_{s=t_0+1}^{t} \eta_s \, v_s + \sum_{s=t_0+1}^{t} \eta_s^2 \, \|v_s\|^2 \, \Phi_s^{(2)}(\theta_{s-1}, \eta_s v_s).$$

Let us bound the last term. We can apply Corollary 5.5, because all assumptions used for proving this corollary have been checked. Corollary 5.5 yields $\|\eta_s v_s\| \leq r_{\mathcal{V}}$, and in Proposition 7.1 we ensured $r_{\mathcal{V}} \leq \tilde{r}_{\mathcal{V}}$. Since θ_{s-1} is in the control ball, the term $\Phi_s^{(2)}(\theta_{s-1}, \eta_s v_s)$ is bounded by Assumption 2.17. By Corollary 5.5, $\|v_s\|^2$ is $O(m(s)^2)$. Therefore, the last term is $O(\sum \eta_s^2 \, m(s)^2)$, so that, for any $t_0 \leq t < t_0 + T_0^{r_\Theta^*}(\boldsymbol{\eta})$, we have

$$\theta_t = \theta_{t_0} - \sum_{s=t_0+1}^{t} \eta_s v_s + O\left(\sum_{s=t_0+1}^{t} \eta_s^2 \, m(s)^2\right).$$

□

We now prove a similar property using constant step sizes.

Lemma 7.4 (Parameter updates at first order with fixed step size)**.** *Let $T \geq 0$. Let $\theta_T \in B_\Theta^*$ with $d(\theta_T, \theta^*)$ at most $r_\Theta^*/3$, and let (v_t) be a gradient sequence with $v_t \in B_{\mathcal{V}_t}$ for all $t \geq T$ ($B_{\mathcal{V}_t}$ is defined just after Lemma 6.28). Let (η_t) be a stepsize sequence with $\overline{\eta} \leq \overline{\eta}_{\mathcal{V}}/2$. Define by induction for $t > T$,*

$$\theta_t = \Phi_t(\theta_{t-1}, \eta_T v_t) = \Phi_{T:t}(\theta_T, (\eta_T v_t)_t).$$

Then for any $T \leq t \leq T + L(T)$,

$$\theta_t = \theta_T - \eta_T \sum_{s=T+1}^{t} v_s + O\left(L(T) \, \eta_T^2 \, m(T)^2\right).$$

Proof. For $T \le t \le T + L(T)$, define $v'_t := (\eta_T/\eta_t)v_t$, so that

$$\Phi_{T:t}(\theta_T, (\eta_T\, v_t)) = \Phi_{T:t}(\theta_T, (\eta_t\, v'_t)).$$

Since we have assumed $\overline{\eta} \le \overline{\eta}_{\mathcal{V}}/2$, we can apply Corollary 5.5 with stepsize sequence $2\eta_t$: this proves that $\|2\eta_t v_t\| \le r_{\mathcal{V}}$. By Lemma 5.22, η_T/η_t is $1+o(1)$. Therefore, for T large enough, we have $\eta_T/\eta_t < 2$. Therefore, $\|\eta_t v'_t\| = \|(\eta_T/\eta_t)\,\eta_t v_t\| \le r_{\mathcal{V}}$ for T large enough. (What happens for small T is absorbed in the $O\,()$ notation.)

Since $\|\eta_t v'_t\| \le r_{\mathcal{V}}$, we can reason exactly as in Lemma 7.3 using v'_t instead of v_t. This yields

$$\theta_t = \theta_T - \sum_{s=T+1}^{t} \eta_s v'_s + O\left(\sum_{s=T+1}^{t} \eta_s^2\, m(s)^2\right)$$

valid at least up to time $t = T + L(T)$ thanks to Lemma 7.2. This yields the conclusion after substituting $v'_s = (\eta_T/\eta_s)v_s$, and after observing that

$$\sum_{s=T+1}^{T+L(T)} \eta_s^2\, m(s)^2 = O\left(L(T)\,\eta_T^2\, m(T)^2\right)$$

because for $T < s \le T + L(T)$ we have $\eta_s/\eta_T = 1 + o(1)$ (by Lemma 5.22) and because $m(s) \sim m(T)$ as $m()$ is a scale function. □

7.2 Stability of θ^* on Intervals $(T; T + L(T)]$

Lemma 7.5 (Averages over time intervals)**.** *Let (u_t) be a sequence with values in some normed vector space. Assume that the average of u_t tends to 0 fast enough, namely,*

$$\frac{1}{T}\sum_{t=1}^{T} u_t = O(e_0(T)/T)$$

for some scale function $e_0(T) \ll T$ when $T \to \infty$.

Let $L(T)$ be any scale function with $e_0(T) \ll L(T) \ll T$ when $T \to \infty$. Then the averages of u over intervals $[T; T + L(T)]$ tend to 0, and more precisely

$$\frac{1}{L(T)}\sum_{t=T+1}^{T+L(T)} u_t = O(e_0(T)/L(T))$$

when $T \to \infty$.

Proof. For any $T \ge 1$,

$$\sum_{t=T+1}^{T+L(T)} u_t = \sum_{t=1}^{T+L(T)} u_t - \sum_{t=1}^{T} u_t = O(e_0(T + L(T))) + O(e_0(T))$$

by assumption.

When $T \to \infty$ we have $T + L(T) \sim T$ because $L(T) \ll T$. Since e_0 is a scale function, we thus have $e_0(T + L(T)) \sim e_0(T)$.

Therefore

$$\sum_{t=T+1}^{T+L(T)} u_t = O(e_0(T))$$

hence the conclusion. □

Remember the scale function $e_0(t) = t^{a'}$ was defined in Lemma 6.6, together with the scale function L. The constraint $e_0 \ll L$ is satisfied thanks to the same Lemma.

Lemma 7.6 (Average of $\frac{\partial \mathcal{L}_{\rightsquigarrow t}(s_0^*, \theta^*)}{\partial \theta}$ on intervals)**.** *When $T \to \infty$, we have*

$$\frac{1}{L(T)} \sum_{t=T+1}^{T+L(T)} \mathcal{U}_t \left(\frac{\partial \mathcal{L}_{\rightsquigarrow t}(s_0^*, \theta^*)}{\partial \theta}, \mathbf{s}_t\left(s_0^*, \theta^*\right), \theta^* \right) = O(e_0(T)/L(T)).$$

Proof. This is a direct consequence of Lemma 7.5 and the definition of a' in Lemma 6.6. Indeed, by the latter we know that Assumption 2.11.b is satisfied with a' instead of a, so that we have

$$\frac{1}{T} \sum_{t=1}^{T} \mathcal{U}_t \left(\frac{\partial}{\partial \theta} \mathcal{L}_{\rightsquigarrow t}(s_0^*, \theta^*), \mathbf{s}_t\left(s_0^*, \theta^*\right), \theta^* \right) = O(e_0(T)/T).$$

Therefore we can apply Lemma 7.5 to the average of $\mathcal{U}_t \left(\frac{\partial \mathcal{L}_{\rightsquigarrow t}(s_0^*, \theta^*)}{\partial \theta}, \mathbf{s}_t\left(s_0^*, \theta^*\right), \theta^* \right)$. □

Lemma 7.7 (Deviation from the optimal parameter for updates computed along the optimal trajectory)**.** *Set $\mathfrak{m}_0^* := (s_0^*, 0)$ (RTRL initialized at s_0^* with $J_0 = 0$) and $\mathfrak{m}_t^* = \mathcal{A}_{0:t}(\theta^*, \mathfrak{m}_0^*)$, the RTRL state at time t using the optimal parameter. Assume that $\overline{\eta} \leq \overline{\eta}_{\mathcal{V}}/2$. Then*

$$\Phi_{T:T+L(T)}(\theta^*, \mathfrak{m}_T^*, \eta_T) = \theta^* + o\left(\eta_T\, L(T)\right).$$

Consequently, the first part of Assumption 4.18 (first-order stability of θ^*) is satisfied for extended RTRL algorithms.

Proof. For any $T \geq 0$, by Corollary 6.5 applied to θ^* with the stepsize sequence $\eta_{t;\, t_1, t_2} := \eta_{t_1}$, we have

$$\Phi_{T:T+L(T)}(\theta^*, \mathfrak{m}_T^*, \eta_T) = \Phi_{T:T+L(T)}\left(\theta^*, (\eta_T\, v_t)_t\right)$$

where

$$v_t = \mathcal{U}_t \left(\frac{\partial \mathcal{L}_{\rightsquigarrow t}(s_0^*, \theta^*)}{\partial \theta}, \mathbf{s}_t\left(s_0^*, \theta^*\right), \theta^* \right).$$

Since the optimal trajectory stays in the stable tube, we have $v_t \in B_{\mathcal{V}_t}$ by definition of $B_{\mathcal{V}_t}$. Then by Lemma 7.4,

$$\Phi_{T:T+L(T)}\left(\theta^*, (\eta_T\, v_t)_t\right) = \theta^* - \eta_T \sum_{t=T+1}^{T+L(T)} \mathcal{U}_t \left(\frac{\partial \mathcal{L}_{\rightsquigarrow t}(s_0^*, \theta^*)}{\partial \theta}, \mathbf{s}_t\left(s_0^*, \theta^*\right), \theta^* \right) + O\left(\eta_T^2\, L(T)\, m(T)^2\right)$$

and by Lemma 7.6,

$$\sum_{t=T+1}^{T+L(T)} \mathcal{U}_t \left(\frac{\partial \mathcal{L}_{\rightsquigarrow t}(s_0^*, \theta^*)}{\partial \theta}, \mathbf{s}_t\left(s_0^*, \theta^*\right), \theta^* \right) = O(e_0(T)).$$

Finally, by Lemma 6.6, both $\eta_T\, e_0(T)$ and $\eta_T^2\, L(T)\, m(T)^2$ are $o\left(\eta_T\, L(T)\right)$. □

7.3 Contractivity Around θ^*

We now turn to the second part of Assumption 4.18, contractivity around θ^*: we have to prove that

$$d\left(\Phi_{t:t+L(t)}\left(\theta, \mathfrak{m}_t^*, \eta_t\right), \Phi_{t:t+L(t)}\left(\theta^*, \mathfrak{m}_t^*, \eta_t\right)\right) \\ \leq \left(1 - \lambda_{\min} \eta_t L(t)\right) d\left(\theta, \theta^*\right) + o\left(\eta_t L(t)\right).$$

We will use a suitable Lyapunov function to define a suitable Euclidean distance $d(\theta, \theta^*)$ for which this holds.

Remember the notation from Assumption 2.11.b, and in particular, the matrices $\mathcal{H}_t(\theta)$ (Jacobian of the parameter update) and Λ (time average of $\mathcal{H}_t(\theta^*)$). Notably, remember that we have endowed Θ with the norm

$$\|\theta\|^2 := \theta^\top B \theta$$

where B, defined in Lemma 6.3, is such that $B\Lambda + \Lambda^\top B$ is positive definite and Λ is given by Assumption 2.11.b. This norm is used as an approximate Lyapunov function for the algorithm.

Lemma 7.8 (Controlling different initialisations for open-loop trajectories)**.** *As before, let* $\mathfrak{m}_0^* := (s_0^*, 0)$. *Let* $\theta \in B_\Theta^*$ *with* $d(\theta, \theta^*) \leq r_\Theta^*/3$. *For* $T \geq 0$, *let* $\mathfrak{m}_T = \mathcal{A}_{0:T}(\theta, \mathfrak{m}_0^*)$. *Assume* $\overline{\eta} \leq \overline{\eta}_{\mathcal{V}}/2$. *Then*

$$\Phi_{T:T+L(T)}(\theta, \mathfrak{m}_T, \eta_T) - \Phi_{T:T+L(T)}(\theta, \mathfrak{m}_T^*, \eta_T) = o\left(\eta_T L(T)\right).$$

Proof. For $t \geq T$, let us write

$$v_t := \boldsymbol{V}_t(\theta, \mathcal{A}_{T:t}(\theta, \mathfrak{m}_T)).$$

($\boldsymbol{V}_t$ and $\mathcal{A}$ for RTRL are given by Def. 6.1). Since θ belongs to B_Θ^*, $\mathfrak{m}_T$ and $\mathcal{A}_{T:t}(\theta, \mathfrak{m}_T)$ belong to the stable tube, so that $v_t \in B_{\mathcal{V}_t}$ by definition of $B_{\mathcal{V}_t}$.

By definition of the open-loop updates Φ (Def. 4.6), we have

$$\Phi_{T:T+L(T)}(\theta, \mathfrak{m}_T, \eta_T) = \Phi_{T:T+L(T)}(\theta, (\eta_T v_t)_t).$$

Likewise with $\mathfrak{m}_T^*$ instead of $\mathfrak{m}_T$, set $v_t' := \boldsymbol{V}_t(\theta, \mathcal{A}_{T:t}(\theta, \mathfrak{m}_T^*))$ so that $\Phi_{T:T+L(T)}(\theta, \mathfrak{m}_T^*, \eta_T) = \Phi_{T:T+L(T)}(\theta, (\eta_T v_t')_t)$.

By Lemma 7.4, we have

$$\Phi_{T:T+L(T)}\left(\theta, (\eta_T v_t)_t\right) - \Phi_{T:T+L(T)}\left(\theta, (\eta_T v_t')_t\right) = \\ \theta - \theta - \eta_T \sum_{t=T+1}^{T+L(T)} (v_t - v_t') + O\left(\eta_T^2 L(T)\, m(T)^2\right).$$

By Assumption 4.16 (which has been checked for RTRL in the previous section) we have

$$\|v_t - v_t'\| = O\left(m(t) \|\mathcal{A}_{T:t}(\theta, \mathfrak{m}_T) - \mathcal{A}_{T:t}(\theta, \mathfrak{m}_T^*)\|\right).$$

By Assumption 4.14 (which has been checked for RTRL in the previous section, with constant $(1 - \alpha/2)$),

$$\|\mathcal{A}_{T:t}(\theta, \mathfrak{m}_T) - \mathcal{A}_{T:t}(\theta, \mathfrak{m}_T^*)\| = O\left((1 - \alpha/2)^{\frac{t-T}{k}} \|\mathfrak{m}_T - \mathfrak{m}_T^*\|\right)$$

and $\|\mathfrak{m}_T - \mathfrak{m}_T^*\|$ is bounded because both belong to the stable tube. Therefore,

$$\|v_t - v_t'\| = O\left(m(t)\,(1-\alpha/2)^{\frac{t-T}{k}}\right).$$

As a result,

$$\begin{aligned}\Phi_{T:T+L(T)}\left(\theta, (\eta_T\, v_t)_t\right) - \Phi_{T:T+L(T)}\left(\theta, (\eta_T\, v_t')_t\right) = \\ O\left(\eta_T \sum_{t=T+1}^{T+L(T)} m(t)\ (1-\alpha/2)^{\frac{t-T}{k}}\right) + O\left(\eta_T^2\, L(T)\, m(T)^2\right) = \\ O\left(\eta_T\, m(T)\right) + O\left(\eta_T^2\, L(T)\, m(T)^2\right) = O\left(\eta_T\, m(T)\right) = o\left(\eta_T\, L(T)\right),\end{aligned}$$

thanks to Corollary 5.18 and the fact $m(T) = o\left(L(T)\right)$. □

Lemma 7.9 (Difference between open-loop trajectories)**.** *Let* $\mathfrak{m}_0^* := (s_0^*, 0)$. *Let* $\theta \in B_\Theta^*$ *with* $d\left(\theta,\, \theta^*\right) \leq r_\Theta^*/3$. *For* $T \geq 0$, *let* $\mathfrak{m}_T = \mathcal{A}_{0:T}(\theta, \mathfrak{m}_0^*)$ *be the RTRL state obtained at time* T *from parameter* θ, *and* $\mathfrak{m}_T^* = \mathcal{A}_{0:T}(\theta^*, \mathfrak{m}_0^*)$. *Assume* $\overline{\eta} \leq \overline{\eta}_{\mathcal{V}}/2$. *For* $u \in [0,1]$, *denote* $\theta^u := \theta + u\ (\theta^* - \theta)$. *Then,*

$$\begin{aligned}\Phi_{T:T+L(T)}(\theta, \mathfrak{m}_T, \eta_T) - \Phi_{T:T+L(T)}(\theta^*, \mathfrak{m}_T^*, \eta_T) = \\ \theta - \theta^* - \eta_T \int_0^1 \sum_{t=T+1}^{T+L(T)} \mathcal{H}_t\left(\theta^u\right) \cdot (\theta - \theta^*)\ du + O\left(\eta_T^2\, L(T)\, m(T)^2\right).\end{aligned}$$

Proof. By Corollary 6.5 with stepsize sequence $\eta_{t;\,t_1,t_2} = \eta_{t_1}$, we have

$$\Phi_{T:T+L(T)}(\theta, \mathfrak{m}_T, \eta_T) = \Phi_{T:T+L(T)}\left(\theta, \eta_T \left(\mathcal{U}_t\left(\frac{\partial \mathcal{L}_{\rightsquigarrow t}(s_0^*, \theta)}{\partial \theta},\, \mathbf{s}_t\left(s_0^*,\, \theta\right),\, \theta\right)\right)_t\right).$$

Let us abbreviate

$$v_t(\theta) := \mathcal{U}_t\left(\frac{\partial \mathcal{L}_{\rightsquigarrow t}(s_0^*, \theta)}{\partial \theta},\, \mathbf{s}_t\left(s_0^*,\, \theta\right),\, \theta\right).$$

Since $\theta \in B_\Theta^*$ and $s_0^* \in \mathbb{T}_0$, $\mathfrak{m}_T$ belongs to the stable tube at time T. Consequently, $v_t(\theta) \in B_{\mathcal{V}_t}$ by definition of $B_{\mathcal{V}_t}$.

By Lemma 7.4,

$$\Phi_{T:T+L(T)}\left(\theta, (\eta_T\, v_t(\theta))_t\right) = \theta - \eta_T \sum_{t=T+1}^{T+L(T)} v_t(\theta) + O\left(\eta_T^2\, L(T)\, m(T)^2\right).$$

Therefore, by writing the same result at $\theta = \theta^*$ and taking differences, we find

$$\begin{aligned}\Phi_{T:T+L(T)}\left(\theta, (\eta_T\, v_t(\theta))_t\right) - \Phi_{T:T+L(T)}(\theta^*, \mathfrak{m}_T^*, \eta_T) = \\ \theta - \theta^* - \eta_T \sum_{t=T+1}^{T+L(T)} \left(v_t(\theta) - v_t(\theta^*)\right) + O\left(\eta_T^2\, L(T)\, m(T)^2\right).\end{aligned}$$

Now, by the definitions of $\mathcal{H}_t\left(\theta\right)$ (Assumption 2.11.b) and of $v_t(\theta)$ above, we have

$$v_t(\theta) = v_t(\theta^*) + \int_0^1 \mathcal{H}_t\left(\theta^u\right) \cdot (\theta - \theta^*)\ du$$

hence the conclusion. □

Lemma 7.10 (Average of Hessians over time intervals)**.** *For $0 \leq u \leq 1$ and $\theta \in B^*_\Theta$, one has*

$$\sum_{t=T+1}^{T+L(T)} \mathcal{H}_t\left(\theta^u\right) = L(T)\left(\Lambda + O\left(\rho\left(\|\theta - \theta^*\|\right)\right) + O(e_0(T)/L(T))\right).$$

Moreover, the $O\left(\rho\left(\|\theta - \theta^*\|\right)\right)$ term is uniform over $0 \leq \overline{\eta} \leq \overline{\eta}_{\mathcal{V}}$ and over $\theta \in B^*_\Theta$.

Proof. Thanks to Assumption 2.25.b, we know that, for all $0 \leq u \leq 1$, we have

$$\begin{aligned}\sum_{t=T+1}^{T+L(T)} \mathcal{H}_t\left(\theta^u\right) &= \sum_{t=T+1}^{T+L(T)} \left(\mathcal{H}_t\left(\theta^*\right) + O\left(\rho\left(\|\theta^u - \theta^*\|\right)\right)\right) \\ &= \sum_{t=T+1}^{T+L(T)} \left(\mathcal{H}_t\left(\theta^*\right) + O\left(\rho\left(\|\theta - \theta^*\|\right)\right)\right).\end{aligned}$$

Now, thanks to the definition of a' in Lemma 6.6, we know that Assumption 2.11.b is satisfied with $e_0(t) = t^{a'}$, so that we have

$$\begin{aligned}\sum_{t=T+1}^{T+L(T)} \mathcal{H}_t\left(\theta^*\right) &= \sum_{t=1}^{T+L(T)} \mathcal{H}_t\left(\theta^*\right) - \sum_{t=1}^{T} \mathcal{H}_t\left(\theta^*\right) \\ &= (T + L(T))\Lambda + O(e_0(T)) - T\Lambda + O(e_0(T)) \\ &= L(T)\left(\Lambda + O(e_0(T)/L(T))\right).\end{aligned}$$

Combining these results yields the statement. □

We now prove the second part of Assumption 4.18 (contractivity around θ^*). Unfortunately, this does not necessarily hold in the ball B^*_Θ that we have used so far, but in a smaller ball B'_Θ. This smaller ball depends on the various quantities involved in the assumptions (such as the constants in the $O\,()$ notation appearing in the various assumptions, or the function $\rho\,()$ in Assumption 2.25.b). Thus, we will have proved all assumptions of Section 4, but over this smaller ball B'_Θ instead of B^*_Θ. We will thus get the convergence of Theorems 4.27, 4.28 and 4.29 for θ in this smaller ball.

Lemma 7.11 (Contractivity around θ^*)**.** *There exists a ball $B'_\Theta \subset B^*_\Theta$ centered at θ^* with positive radius, and $\lambda > 0$, such that the following holds.*

Let $\mathfrak{m}^*_0 := (s^*_0, 0)$ (RTRL state initialized at s^*_0 with $J_0 = 0$) and $\mathfrak{m}^*_t = \mathcal{A}_{0:t}(\theta^*, \mathfrak{m}^*_0)$ (RTRL state at time t using the optimal parameter). Assume $\overline{\eta} \leq \overline{\eta}_{\mathcal{V}}/2$.

For every $\theta \in B'_\Theta$,

$$\left\|\Phi_{T:T+L(T)}(\theta, \mathfrak{m}^*_T, \eta_T) - \Phi_{T:T+L(T)}(\theta^*, \mathfrak{m}^*_T, \eta_T)\right\|$$

is at most

$$\left(1 - \lambda\,\eta_T\,L(T)\right)\|\theta - \theta^*\| + o\left(\eta_T\,L(T)\right)$$

and the $o\,()$ term is uniform over $0 \leq \overline{\eta} \leq \overline{\eta}_{\mathcal{V}}/2$ and over $\theta \in B'_\Theta$.

Therefore, the second part of Assumption 4.18 is satisfied by an extended RTRL algorithm for θ in the ball B'_Θ, for the distance $d\left(\theta, \theta'\right)^2 := (\theta - \theta')^\top B(\theta - \theta')$.

Proof. Let $\theta \in B^*_\Theta$ with $d(\theta, \theta^*) \leq r^*_\Theta/3$. (We will have further constraints below to define the smaller ball B'_Θ.)

By combining the last three lemmas, we obtain

$$\begin{aligned}
&\Phi_{T:T+L(T)}(\theta, \mathfrak{m}^*_T, \eta_T) - \Phi_{T:T+L(T)}(\theta^*, \mathfrak{m}^*_T, \eta_T) = \\
&\Phi_{T:T+L(T)}(\theta, \mathfrak{m}_T, \eta_T) - \Phi_{T:T+L(T)}(\theta^*, \mathfrak{m}^*_T, \eta_T) + \Phi_{T:T+L(T)}(\theta, \mathfrak{m}^*_T, \eta_T) - \Phi_{T:T+L(T)}(\theta, \mathfrak{m}_T, \eta_T) \\
&\quad = (\theta - \theta^*) - \eta_T L(T) \left(\Lambda + O\left(\rho\left(\|\theta - \theta^*\|\right)\right) + O(e_0(T)/L(T))\right) \cdot (\theta - \theta^*) \\
&\qquad\qquad + O\left(\eta_T^2\, L(T)\, m(T)^2\right) + o\left(\eta_T\, L(T)\right)
\end{aligned}$$

which equals

$$\begin{aligned}
&(\mathrm{Id}_\Theta - \eta_T L(T)\Lambda) \cdot (\theta - \theta^*) + O\left(\eta_T L(T)\, \rho\left(\|\theta - \theta^*\|\right) \|\theta - \theta^*\|\right) \\
&\qquad + O\left(\eta_T\, e_0(T)\, \|\theta - \theta^*\|\right)) + O\left(\eta_T^2\, L(T)\, m(T)^2\right) + o\left(\eta_T\, L(T)\right).
\end{aligned}$$

Since $\|\theta - \theta^*\|$ is bounded on B^*_Θ, the term $\eta_T\, e_0(T)\ \|\theta - \theta^*\|$ is $O(\eta_T e_0(T))$. By Lemma 6.6, both $\eta_T\, e_0(T)$ and $\eta_T^2\, L(T)\, m(T)^2$ are $o\left(\eta_T\, L(T)\right)$. So the last two $O\left(\right)$ terms above are absorbed in the $o\left(\eta_T\, L(T)\right)$ term.

Remember that the norm we use on Θ is defined by $\|\theta\|^2 = \theta^\top B\, \theta$. Therefore, we have

$$\begin{aligned}
\|(\mathrm{Id}_\Theta - \eta_T L(T)\Lambda) \cdot (\theta - \theta^*)\|^2 &= (\theta - \theta^*)^\top B\, (\theta - \theta^*) \\
&\quad - \eta_T L(T) (\theta - \theta^*)^\top \left(B\,\Lambda + \Lambda^T\, B\right) (\theta - \theta^*) + O\left(\eta_T^2 L(T)^2\, \|\theta - \theta^*\|^2\right).
\end{aligned}$$

Thanks to Assumption 2.11.b and to Lemma 6.3, we know that $B\,\Lambda + \Lambda^T\, B$ is positive definite, so that for some $\lambda > 0$ we have

$$(\theta - \theta^*)^\top \left(B\,\Lambda + \Lambda^T\, B\right) (\theta - \theta^*) \geq 4\,\lambda\, \|\theta - \theta^*\|^2,$$

and we obtain

$$\|(\mathrm{Id}_\Theta - \eta_T L(T)\Lambda) \cdot (\theta - \theta^*)\|^2 \ \leq\ \|\theta - \theta^*\|^2 \left(1 - 4\,\lambda\, \eta_T L(T) + O\left(\eta_T^2 L(T)^2\right)\right).$$

Therefore, the quantity we want to compute is

$$\begin{aligned}
&\left\|\Phi_{T:T+L(T)}(\theta, \mathfrak{m}^*_T, \eta_T) - \Phi_{T:T+L(T)}(\theta^*, \mathfrak{m}^*_T, \eta_T)\right\| \leq \\
&\qquad \|\theta - \theta^*\| \left(1 - 4\,\lambda\, \eta_T L(T) + O\left(\eta_T^2 L(T)^2\right)\right)^{\frac{1}{2}} \\
&\quad + O\left(\eta_T L(T)\, \rho\left(\|\theta - \theta^*\|\right)\, \|\theta - \theta^*\|\right) + o\left(\eta_T\, L(T)\right) \leq \\
&\qquad \|\theta - \theta^*\| \left(1 - 2\,\lambda\, \eta_T L(T) + O\left(\eta_T^2 L(T)^2\right)\right) \\
&\qquad\qquad + O\left(\eta_T L(T)\, \rho\left(\|\theta - \theta^*\|\right)\, \|\theta - \theta^*\|\right) + o\left(\eta_T\, L(T)\right),
\end{aligned}$$

as $\sqrt{1-x} \leq 1 - x/2$.

Since $\eta_T L(T) \to 0$, the term $O\left(\eta_T^2 L(T)^2\right)$ is ultimately smaller than $(\lambda/2)\eta_T\, L(T)$.

Consider the term $O\left(\eta_T L(T)\, \rho\left(\|\theta - \theta^*\|\right) \|\theta - \theta^*\|\right)$. The constant in the $O\left(\right)$ notation is independent of T or $\theta \in B^*_\Theta$. Let κ be that constant. Since $\rho \to 0$ at 0, there is a ball B'_Θ of some fixed radius around θ^*, in which $\rho\left(\|\theta - \theta^*\|\right)$

is smaller than $\lambda/2\kappa$. Therefore, in that ball, $O\left(\eta_T L(T)\rho\left(\|\theta-\theta^*\|\right)\|\theta-\theta^*\|\right) \leq (\lambda/2)\,\eta_T\,L(T)\,\|\theta-\theta^*\|$. On this smaller ball B'_Θ, one has

$$\left\|\Phi_{T:T+L(T)}(\theta,\mathfrak{m}_T^*,\eta_T)-\Phi_{T:T+L(T)}(\theta^*,\mathfrak{m}_T^*,\eta_T)\right\| \leq \\ \|\theta-\theta^*\|\left(1-2\lambda\eta_T\,L(T)+(\lambda/2)\eta_T\,L(T)+(\lambda/2)\eta_T\,L(T)\right)+o\left(\eta_T\,L(T)\right),$$

as needed. □

7.4 Noise Control for Imperfect RTRL Algorithms

In this section we bound the divergence between imperfect RTRL algorithms and exact RTRL. We consider an imperfect RTRL algorithm (Def. 2.10) whose errors E_t satisfy the unbiasedness Assumption 2.18 and the error control Assumption 2.21 with some error gauge ϕ (Def. 2.19). (Actually the results presented here will be valid simultaneously for all imperfect RTRL algorithms sharing the same error gauge.) Moreover, we assume that the extended update rules $\mathcal{U}_t$ are linear with respect to their first argument (Assumption 2.22).

We compare such imperfect RTRL algorithms to the corresponding RTRL algorithm with error $E_t = 0$ but the same underlying system.

Lemma 7.12 (Imperfect Jacobians expressed in terms of exact Jacobians plus noise)**.** *Let $(s_0,\,J_0) \in \mathbb{T}_0 \times \mathbb{T}_0^J$, and let $\boldsymbol{\theta} = (\theta_t)$ be a sequence of parameters included in B_Θ^*. Let (s_t) be the trajectory of states starting at s_0 computed from (θ_t), namely, via $s_t = \mathbf{T}_t\left(s_{t-1},\,\theta_{t-1}\right)$ for $t \geq 1$.*

We now compare Jacobians computed with the exact RTRL updates, and with imperfect updates. Precisely, consider

1. *the Jacobians $\left(\tilde{J}_t\right)$ starting at J_0 and following the imperfect RTRL updates*

$$\tilde{J}_t = \frac{\partial \mathbf{T}_t(s_{t-1},\theta_{t-1})}{\partial s}\,\tilde{J}_{t-1} + \frac{\partial \mathbf{T}_t(s_{t-1},\theta_{t-1})}{\partial \theta} + E_t$$

   for $t \geq 1$, where the errors E_t satisfy Assumption 2.21;

2. *the Jacobians (J'_t) also starting at J_0 and following the exact RTRL updates that is, for $t \geq 1$,*

$$J'_t = \frac{\partial \mathbf{T}_t}{\partial s}\left(s_{t-1},\,\theta_{t-1}\right)\cdot J'_{t-1} + \frac{\partial \mathbf{T}_t}{\partial \theta}\left(s_{t-1},\,\theta_{t-1}\right).$$

Then, for every $t \geq 1$,

$$\tilde{J}_t = J'_t + \sum_{s\leq t}\left(\prod_{p=s+1}^{t}\frac{\partial \mathbf{T}_p}{\partial s}\left(\mathbf{T}_{0:p-1}\left(s_0,\,\theta_0\right),\,\theta_0\right)\right)E_s + O\left(\sup_{s\leq t-1} d\left(\theta_s,\,\theta_0\right)\right),$$

where the constant in the $O\left(\right)$ term only depends on the constants appearing in the assumptions and on the error gauge. (By convention, for $s = t$ the empty product $\prod_{p=s+1}^{t}$ is equal to Id.*)*

Proof. For every $t \geq 1$,

$$\tilde{J}_t - J'_t = \frac{\partial \mathbf{T}_t}{\partial s}\left(s_{t-1},\,\theta_{t-1}\right)\left(\tilde{J}_{t-1} - J'_{t-1}\right) + E_t.$$

Now, we have seen above that the non-imperfect RTRL operator on (s, J) (Definition 6.1) satisfies Assumption 4.13 and Assumption 4.14. Setting $(s''_t, J''_t) := \mathcal{A}_{0:t}(\theta_0, (s_0, J_0))$, we may therefore apply Lemma 5.15 to compare the (non-imperfect) RTRL trajectories with fixed parameter θ_0 and with variable parameter (θ_t): this yields, for all $t \geq 1$,

$$d\left((s_t, J'_t),\, (s''_t, J''_t)\right) = O\left(\sup_{s \leq t-1} d\left(\theta_s,\, \theta_0\right)\right).$$

Setting $S_t := \sup_{s \leq t-1} d\left(\theta_s,\, \theta_0\right)$, we have a fortiori

$$d\left(s_t,\, s''_t\right) = O\left(S_t\right).$$

Also note that $s''_t = \mathbf{T}_{0:t}(s_0, \theta_0)$ by definition of the RTRL operator $\mathcal{A}_{0:t}$.

Moreover, for all $t \geq 0$, θ_t and s_t belong to the stable tube for RTRL. Thanks to Assumption 2.23 the second derivatives of the transition operator on the states are bounded. As a result, for all $t \geq 1$,

$$\begin{aligned}\frac{\partial \mathbf{T}_t}{\partial s}\left(s_{t-1},\, \theta_{t-1}\right) &= \frac{\partial \mathbf{T}_t}{\partial s}\left(s''_{t-1},\, \theta_0\right) + O\left(d\left(s_{t-1},\, s''_{t-1}\right) + d\left(\theta_{t-1},\, \theta_0\right)\right) \\ &= \frac{\partial \mathbf{T}_t}{\partial s}\left(s''_{t-1},\, \theta_0\right) + O\left(S_t\right).\end{aligned}$$

As a consequence, for all $t \geq 1$,

$$\begin{aligned}\tilde{J}_t - J'_t = \frac{\partial \mathbf{T}_t}{\partial s}\left(s''_{t-1},\, \theta_0\right)\, \left(\tilde{J}_{t-1} - J'_{t-1}\right) \\ + O\left(S_t \left\|\tilde{J}_{t-1} - J'_{t-1}\right\|\right) + E_t.\end{aligned}$$

Now, since the sequences $\left(\tilde{J}_t\right)$ and (J'_t) are computed by an imperfect and exact RTRL algorithm from a sequence of parameters (θ_t) in the control ball and an initialization (s_0, J_0) in the stable tube, they both belong to the stable tube, and are therefore bounded by Corollary 6.24. Thus $J_t - J'_t$ is bounded and

$$O\left(S_t \left\|\tilde{J}_{t-1} - J'_{t-1}\right\|\right) = O\left(S_t\right)$$

so that

$$\tilde{J}_t - J'_t = \frac{\partial \mathbf{T}_t}{\partial s}\left(s''_{t-1},\, \theta_0\right)\, \left(\tilde{J}_{t-1} - J'_{t-1}\right) + E_t + r_t$$

for some remainder $r_t = O\left(S_t\right)$. From this we obtain, by induction,

$$\begin{aligned}\tilde{J}_t - J'_t = \sum_{s \leq t} \left(\prod_{p=s+1}^{t} \frac{\partial \mathbf{T}_p}{\partial s}\left(s''_{p-1},\, \theta_0\right)\right) E_s \\ + \sum_{s \leq t} \left(\prod_{p=s+1}^{t} \frac{\partial \mathbf{T}_p}{\partial s}\left(s''_{p-1},\, \theta_0\right)\right) r_s.\end{aligned}$$

Since $s''_t = \mathbf{T}_{0:t}(s_0, \theta_0)$, the claim will be proved if we prove that the remainder term is $O\left(S_t\right)$. The norm of the remainder term is

$$\left\|\sum_{s \leq t} \left(\prod_{p=s+1}^{t} \frac{\partial \mathbf{T}_p}{\partial s}\left(s''_{p-1},\, \theta_0\right)\right) r_s\right\| \leq \sum_{s \leq t} \left\|\prod_{p=s+1}^{t} \frac{\partial \mathbf{T}_p}{\partial s}\left(s''_{p-1},\, \theta_0\right)\right\|_{\mathrm{op}} \|r_s\|.$$

Since s''_{t-1} belongs to the stable tube, we can apply Corollary 6.16: the product $\prod_{p=s+1}^{t} \frac{\partial \mathbf{T}_p}{\partial s}\left(s''_{p-1},\, \theta_0\right)$ has operator norm at most $M(1-\alpha/2)^{(t-s)/h}$ for some constant $M > 0$.

As a result,

$$\sum_{s\le t} \left\| \prod_{p=s+1}^{t} \frac{\partial \mathbf{T}_p}{\partial s}\left(s''_{p-1},\, \theta_0\right) \right\|_{\mathrm{op}} \|r_s\| \le \left(\sup_{s\le t} \|r_s\|\right) \left(\sum_{s\le t} M(1-\alpha/2)^{(t-s)/h}\right)$$
$$\le (2hM/\alpha) \sup_{s\le t} \|r_s\| = O\left(S_t\right)$$

since $r_s = O\left(S_t\right)$. □

Remember that, for imperfect RTRL algorithms, we assume that $\mathcal{U}_t$ has the form $\mathcal{U}_t(v, s, \theta) = P_t(s, \theta) \cdot v$ for some linear operator P_t (Assumption 2.22, in addition to Assumption 2.14).

Lemma 7.13 (Bounds on $\mathcal{U}_t$ in the linear case)**.** *Under Assumptions 2.14 and 2.22, the operator P_t and its derivative with respect to (s, θ) are bounded on a ball containing the stable tube. Namely,*

$$\sup_{t\ge 1} \sup_{B_{\mathcal{S}_t}(s_t^*, r_{\mathcal{S}})\times B_\Theta(\theta^*, r_\Theta)} \|P_t(s,\theta)\|_{\mathrm{op}} < \infty$$

and

$$\sup_{t\ge 1} \sup_{B_{\mathcal{S}_t}(s_t^*, r_{\mathcal{S}})\times B_\Theta(\theta^*, r_\Theta)} \left\|\partial_{(s,\theta)} P_t(s,\theta)\right\|_{\mathrm{op}} < \infty,$$

where the balls are those appearing in Assumption 2.14.

Proof. This is a direct rewriting of Assumption 2.14 for the particular case $\mathcal{U}_t(v, s, \theta) = P_t(s, \theta) \cdot v$. Indeed, $\partial_v \mathcal{U}_t = P_t(s, \theta)$ so that the first point of Assumption 2.14 gives the first statement of the lemma.

For the second statement, consider the second point of Assumption 2.14. Here $\partial_{(s,\theta)} \mathcal{U}_t\left(v,\, s,\, \theta\right)$ is $\partial_{(s,\theta)}(P_t(s,\theta) \cdot v)$. Its operator norm is $O(1 + \|v\|)$ by Assumption 2.14. Remember that $P_t(s,\theta) \in \mathrm{L}(\mathrm{L}(\Theta, \mathbb{R}), \Theta)$. Let us now compute the operator norm of $\partial_{(s,\theta)} P_t(s,\theta)$. Let $(u_s, u_\theta) \in \mathcal{S} \times \Theta$ be a unit vector that realizes this operator norm, namely, $\left\|\partial_{(s,\theta)} P_t(s,\theta)\right\|_{\mathrm{op}} = \left\|\partial_{(s,\theta)} P_t(s,\theta) \cdot (u_s, u_\theta)\right\|_{\mathrm{op}}$; here the second operator norm is as an operator on $v \in \mathrm{L}(\Theta, \mathbb{R})$, since $\partial_{(s,\theta)} P_t(s,\theta) \cdot (u_s, u_\theta) \in \mathrm{L}(\mathrm{L}(\Theta, \mathbb{R}), \Theta)$. Let now $v \in \mathrm{L}(\Theta, \mathbb{R})$ be a unit vector that realizes the operator norm of $\partial_{(s,\theta)} P_t(s,\theta) \cdot (u_s, u_\theta)$, so that $\left\|\partial_{(s,\theta)} P_t(s,\theta)\right\|_{\mathrm{op}} = \left\|\left(\partial_{(s,\theta)} P_t(s,\theta) \cdot (u_s, u_\theta)\right) \cdot v\right\|$ (the last norm is in Θ). Since $P_t(s,\theta) \cdot v$ is linear in v, we have $\left(\partial_{(s,\theta)} P_t(s,\theta) \cdot (u_s, u_\theta)\right) \cdot v = \partial_{(s,\theta)}\left(P_t(s,\theta) \cdot v\right) \cdot (u_s, u_\theta)$ (indeed, both are the limit of $\frac{1}{\varepsilon}(P_t(s+\varepsilon u_s, \theta+\varepsilon u_\theta) \cdot v - P_t(s,\theta) \cdot v)$ when $\varepsilon \to 0$). Therefore, $\left\|\partial_{(s,\theta)} P_t(s,\theta)\right\|_{\mathrm{op}} = \left\|\partial_{(s,\theta)}\left(P_t(s,\theta) \cdot v\right) \cdot (u_s, u_\theta)\right\|$. However, since (u_s, u_θ) is a unit vector, the latter is at most $\left\|\partial_{(s,\theta)}\left(P_t(s,\theta) \cdot v\right)\right\|_{\mathrm{op}}$. This is $O(1 + \|v\|)$ by Assumption 2.14, but v is a unit vector so this is $O(1)$.

Finally, by Lemma 6.18, the stable tube for s and θ is included in the balls of Assumption 2.14. □

We now introduce an operator that represents the first-order change in the computed gradients v_t, with respect to a change of state at a previous time s; this linearization is computed along a trajectory defined by some s_{t_0} and θ.

Definition 7.14 (Product of the differentials with respect to the states of the transition operators)**.** *Let $t_0 \geq 1$. Let $\theta \in B^*_\Theta$, and $s_{t_0} \in B^*_{\mathcal{S}_{t_0}}$. For $t \geq t_0$, abbreviate $s_t = \mathbf{T}_{t_0:t}(s_{t_0}, \theta)$, using the notation from Definition 6.2. We define, for $t \geq s \geq t_0$, the linear operator $\Pi^{t_0}_{s,t}(s_{t_0}, \theta)$ from $\mathrm{L}(\Theta, \mathcal{S}_s)$ to Θ, which sends $E \in \mathrm{L}(\Theta, \mathcal{S}_s)$ to*

$$\Pi^{t_0}_{s,t}(s_{t_0}, \theta)\, E := P_t(s_t, \theta) \cdot \left(\frac{\partial \mathcal{L}_t}{\partial s}(s_t) \cdot \left(\prod_{p=s+1}^{t} \frac{\partial \mathbf{T}_p}{\partial s}(s_{p-1}, \theta) \right) E \right).$$

(Note that $E \in \mathrm{L}(\Theta, \mathcal{S}_s)$ so that $\left(\prod_{p=s+1}^{t} \frac{\partial \mathbf{T}_p}{\partial s}(s_{p-1}, \theta)\right) E$ belongs to $\mathrm{L}(\Theta, \mathcal{S}_t)$; multiplying this by $\partial_s \mathcal{L}_t(s_t) \in \mathrm{L}(\mathcal{S}_t, \mathbb{R})$ produces an element of $\mathrm{L}(\Theta, \mathbb{R})$, from which P_t produces an element of Θ.)

Lemma 7.15 (Expressing the imperfect tangent vectors)**.** *Under the exact same assumptions and notations as in Lemma 7.12, set*

$$v_t = \mathcal{U}_t\left(\frac{\partial \mathcal{L}_t(s_t)}{\partial s} \cdot \tilde{J}_t,\, s_t,\, \theta_{t-1} \right), \qquad v'_t = \mathcal{U}_t\left(\frac{\partial \mathcal{L}_t(s_t)}{\partial s} \cdot J'_t,\, s_t,\, \theta_{t-1} \right)$$

where $\mathcal{U}_t$ satisfies Assumption 2.22, namely, $\mathcal{U}_t(v, s, \theta) = P_t(s, \theta) \cdot v$ for some linear operator P_t.

Then, for all $t \geq 1$,

$$v_t - v'_t = \sum_{s \leq t} \Pi^0_{s,t}(s_0, \theta_0)\, E_s + O\left(m(t) \sup_{s \leq t} d(\theta_s, \theta_0) \right).$$

Proof. Let $s''_t := \mathbf{T}_{0:t}(s_0, \theta_0)$ as in Lemma 7.12. Thanks to the assumption on $\mathcal{U}_t$,

$$\begin{aligned} v_t - v'_t &= P_t(s_t, \theta_{t-1}) \cdot \frac{\partial \mathcal{L}_t}{\partial s}(s_t) \cdot \left(\tilde{J}_t - J'_t \right) \\ &= P_t(s''_t, \theta_0) \cdot \frac{\partial \mathcal{L}_t}{\partial s}(s''_t) \cdot \left(\tilde{J}_t - J'_t \right) \\ &\quad + O\left(\left(P_t(s_t, \theta_{t-1}) \cdot \frac{\partial \mathcal{L}_t}{\partial s}(s_t) - P_t(s''_t, \theta_0) \cdot \frac{\partial \mathcal{L}_t}{\partial s}(s''_t) \right) \left\| \tilde{J}_t - J'_t \right\| \right). \end{aligned}$$

By the expression for $\tilde{J}_t - J'_t$ in Lemma 7.12, and by definition of $\Pi^0_{s,t}$ and of s''_t, we have

$$P_t(s''_t, \theta_0) \cdot \frac{\partial \mathcal{L}_t}{\partial s}(s''_t) \cdot \left(\tilde{J}_t - J'_t \right) = \sum_{s \leq t} \Pi^0_{s,t}(s_0, \theta_0)\, E_s + O\left(\left\| P_t(s''_t, \theta_0) \right\|_{\mathrm{op}} \left\| \frac{\partial \mathcal{L}_t}{\partial s}(s''_t) \right\| S_t \right)$$

where $S_t := \sup_{s \leq t-1} d(\theta_s, \theta_0)$ as in Lemma 7.12.

As in Lemma 7.12, s''_t belongs to the stable tube. By Assumption 2.24, $\left\| \frac{\partial \mathcal{L}_t}{\partial s}(s''_t) \right\| = O(m(t))$. Lemma 7.13 shows that P_t is bounded on the table tube. So $\|P_t(s''_t, \theta_0)\| \left\| \frac{\partial \mathcal{L}_t}{\partial s}(s''_t) \right\| S_t = O(m(t)\, S_t)$.

Thus, to reach our expression for $v_t - v'_t$, we only have to prove that

$$O\left(\left(P_t(s_t, \theta_{t-1}) \cdot \frac{\partial \mathcal{L}_t}{\partial s}(s_t) - P_t(s''_t, \theta_0) \cdot \frac{\partial \mathcal{L}_t}{\partial s}(s''_t) \right) \left\| \tilde{J}_t - J'_t \right\| \right) = O(m(t)\, S_t).$$

As in Lemma 7.12, $\left\| \tilde{J}_t - J'_t \right\|$ is uniformly bounded because both belong to the stable tube. So we have to prove that $P_t(s_t, \theta_{t-1}) \cdot \frac{\partial \mathcal{L}_t}{\partial s}(s_t) - P_t(s''_t, \theta_0) \cdot \frac{\partial \mathcal{L}_t}{\partial s}(s''_t) = O(m(t)\, S_t)$.

By Assumption 2.14, $\mathcal{U}_t$ is C^1 so that P_t is C^1. By Definition 2.7, $\mathcal{L}_t$ is C^2. Therefore,

$$P_t(s_t, \theta_{t-1}) \cdot \frac{\partial \mathcal{L}_t}{\partial s}(s_t) - P_t(s_t'', \theta_0) \cdot \frac{\partial \mathcal{L}_t}{\partial s}(s_t'') = O\left(\left(d(s_t,\, s_t'') + d(\theta_{t-1},\, \theta_0) \right) \sup_{(s,\theta)} \left\| \partial_{(s,\theta)} \left(P_t(s,\theta) \cdot \frac{\partial \mathcal{L}_t}{\partial s}(s) \right) \right\| \right)$$

where the supremum over (s, θ) is on a ball where all assumptions hold, since s_t, s_t'', θ_{t-1} and θ_0 all belong to the stable tube.

We want to prove that this quantity is $O(m(t)\, S_t)$. We proved in Lemma 7.12 that $d(s_t,\, s_t'')$ is $O(S_t)$, and $d(\theta_{t-1},\, \theta_0)$ is $O(S_t)$ by definition of S_t.

So we have to prove that the derivative of $P_t(s,\theta) \cdot \frac{\partial \mathcal{L}_t}{\partial s}(s)$ with respect to (s,θ) is $O(m(t))$. By differentiating the product, we have to bound $P_t(s,\theta)$ and its derivative, as well as the first and second derivatives of $\mathcal{L}_t$ with respect to s.

Thanks to Lemma 7.13, $P_t(s,\theta)$ is bounded on the stable tube, together with its derivative $\partial_{(s,\theta)} P_t(s,\theta)$.

By Assumption 2.24, the first and second derivatives of $\mathcal{L}_t$ are controlled by $O(m(t))$ on the stable tube.

This proves that $\partial_{(s,\theta)} \left(P_t(s,\theta) \cdot \partial_s \mathcal{L}_t(s) \right)$ is $O(m(t))$ on the stable tube. This ends the proof. □

Lemma 7.16 (Operator norm of $\Pi_{s,\,t}^{t_0}$)**.** *There is a constant $M > 0$ such that, for any $t_0 > 1$, for any $\theta_{t_0} \in B_\Theta^*$ and $s_{t_0} \in \mathbb{T}_{t_0}$, for any $t \geq s \geq t_0$, the operator $\Pi_{s,\,t}^{t_0}(s_{t_0}, \theta_{t_0})$ has operator norm at most*

$$\left\| \Pi_{s,\,t}^{t_0}(s_{t_0}, \theta_{t_0}) \right\|_{\mathrm{op}} \leq M\, m(t)\, (1 - \alpha/2)^{(t-s)/h},$$

where α and h are the spectral radius constants of Assumption 2.13, and where the operator norm of $\Pi_{s,\,t}^{t_0}(s_{t_0}, \theta_{t_0}) \in \mathrm{L}(\mathrm{L}(\Theta, \mathcal{S}_s), \Theta)$ is defined with respect to the operator norm on $\mathrm{L}(\Theta, \mathcal{S}_s)$ and the usual norm on Θ.

Proof. Let $E \in \mathrm{L}(\Theta, \mathcal{S}_s)$. Then by Definition 7.14,

$$\Pi_{s,\,t}^{t_0}(s_{t_0}, \theta_{t_0})\, E = P_t(s_t, \theta_{t_0}) \cdot \left(\frac{\partial \mathcal{L}_t}{\partial s}(s_t) \cdot \left(\prod_{p=s+1}^{t} \frac{\partial \mathbf{T}_p}{\partial s}(s_{p-1},\, \theta_{t_0}) \right) E \right)$$

where $s_t := \mathbf{T}_{t_0:t}(s_{t_0}, \theta_{t_0})$ for $t \geq t_0$. Since s_{t_0} belongs to the stable tube and $\theta_{t_0} \in B_\Theta^*$, these states belong to the stable tube. Therefore, we can apply Corollary 6.16: the product $\prod_{p=s+1}^{t} \frac{\partial \mathbf{T}_p}{\partial s}(s_{p-1},\, \theta_{t_0})$ has operator norm at most $M(1 - \alpha/2)^{(t-s)/h}$ for some constant $M > 0$.

By Assumption 2.24, the first derivative of $\mathcal{L}_t$ has operator norm $O(m(t))$ on the stable tube. Finally, by Lemma 7.13, the operator P_t has bounded operator norm on the stable tube. This proves the claim. □

Let $(\mathfrak{m}_t)_{t \geq t_0}$ be the sequence produced by the imperfect RTRL algorithm starting at $\mathfrak{m}_{t_0}$, and with a parameter θ_{t_0}. Remember the deviation $D_{t_0:t}\left(\theta_{t_0},\, (\mathfrak{m}_t)_{t \geq t_0},\, \boldsymbol{\eta}\right)$ introduced in Definition 4.24, that measures the effect on θ_t of the difference between the imperfect RTRL trajectory $\mathfrak{m}_t$ and a trajectory closer to RTRL.

Lemma 7.17 (Expression of the noise)**.** *Let* $\theta_{t_0} \in \Theta$ *with* $d(\theta, \theta^*) \leq \frac{r^*_\Theta}{3}$. *Let* $(s_{t_0}, J_{t_0}) \in \mathbb{T}_{t_0} \times \mathbb{T}^J_{t_0}$. *Assume* $\overline{\eta} \leq \overline{\eta}_{\mathcal{V}}$.

Let θ_t *and* $\mathfrak{m}_t = (s_t, \tilde{J}_t)$ *be the trajectory computed by an imperfect RTRL algorithm (Def. 2.10) starting at* $\mathfrak{m}_{t_0} = (s_{t_0}, J_{t_0})$ *at time* t_0.

Then, for all $t_0 + 1 \leq t < T^{r^*_\Theta}_{t_0}$,

$$D_{t_0:t}\left(\theta_{t_0}, (\mathfrak{m}_t), \boldsymbol{\eta}\right) \leq \left\| \sum_{s=t_0+1}^{t} \left(\sum_{p=s}^{t} \eta_p \, \Pi^{t_0}_{s,\,p}(s_{t_0}, \theta_{t_0}) \right) E_s \right\| + O\left(\left(\sum_{s=t_0+1}^{t} \eta_s \, m(s) \right)^2 \right).$$

Note that, since we assume $\bar{\eta} \leq \overline{\eta}_{\mathcal{V}}$, the O term is uniform with respect to $\bar{\eta} \leq \overline{\eta}_{\mathcal{V}}$.

Moreover, thanks to Lemma 7.2, we know that, for t large enough, we have $T^{r^*_\Theta}_t > L(t)$. As a result, by definition of T_{k+1}, for k large enough, all results apply to times t in the interval $[T_k; T_{k+1}]$.

Proof. Without loss of generality, we conduct the proof with $t_0 = 0$.

First expression of the noise. From Definition 2.10, the imperfect RTRL trajectory $(\mathfrak{m}_t) = (s_t, \tilde{J}_t)$ and (θ_t) satisfies

$$\begin{cases} s_t = \mathbf{T}_t\left(s_{t-1}, \theta_{t-1}\right) \\ \tilde{J}_t = \dfrac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial s} \tilde{J}_{t-1} + \dfrac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial \theta} + E_t \\ v_t = \mathcal{U}_t\left(\dfrac{\partial \mathcal{L}_t(s_t)}{\partial s} \cdot \tilde{J}_t, \, s_t, \, \theta_{t-1} \right) \\ \theta_t = \Phi\left(\theta_{t-1}, \, \eta_t \, v_t\right). \end{cases}$$

Remember the definition 4.24 of the deviation $D_{t_0:t}\left(\theta_{t_0}, (\mathfrak{m}_t)_{t \geq t_0}, \boldsymbol{\eta}\right)$. To compute the deviation, we have to compare this to the regularized trajectory initialized likewise, but satisfying the recurrence equations $\bar{\mathfrak{m}}_t = \mathcal{A}_t(\theta_{t-1}, \bar{\mathfrak{m}}_{t-1})$, $\bar{v}_t = \boldsymbol{V}_t\left(\theta_{t-1}, \bar{\mathfrak{m}}_t\right)$, and $\bar{\theta}_t = \Phi_t(\bar{\theta}_{t-1}, \eta_t \bar{v}_t)$ with $\mathcal{A}_t$ and $\boldsymbol{V}_t$ the operators of the RTRL algorithm (Def. 6.1). The equation on $\bar{\mathfrak{m}}_t$ amounts to $\bar{\mathfrak{m}}_t = (\bar{s}_t, \bar{J}_t)$ with

$$\begin{cases} \bar{s}_t = \mathbf{T}_t\left(\bar{s}_{t-1}, \theta_{t-1}\right) \\ \bar{J}_t = \dfrac{\partial \mathbf{T}_t}{\partial s}\left(\bar{s}_{t-1}, \theta_{t-1}\right) \cdot \bar{J}_{t-1} + \dfrac{\partial \mathbf{T}_t}{\partial \theta}\left(\bar{s}_{t-1}, \theta_{t-1}\right) \end{cases}$$

but the equation for $\bar{s}_t$ is the same as for s_t, so $\bar{s}_t = s_t$ for all $t \geq 0$, and thus

$$\bar{J}_t = \frac{\partial \mathbf{T}_t}{\partial s}\left(s_{t-1}, \theta_{t-1}\right) \cdot \bar{J}_{t-1} + \frac{\partial \mathbf{T}_t}{\partial \theta}\left(s_{t-1}, \theta_{t-1}\right)$$

and the evolution equations of $\bar{v}_t$ and $\bar{\theta}_t$ become

$$\bar{v}_t = \mathcal{U}_t\left(\frac{\partial \mathcal{L}_t(s_t)}{\partial s} \cdot \bar{J}_t, \, s_t, \, \theta_{t-1} \right), \qquad \bar{\theta}_t = \Phi\left(\bar{\theta}_{t-1}, \, \eta_t \, \bar{v}_t\right),$$

by the definition of $\boldsymbol{V}_t$ (Def. 6.1). Then, thanks to Definition 4.24, for all $t \geq 1$,

$$D_{0:t}\left(\theta_0, (\mathfrak{m}_t), \boldsymbol{\eta}\right) = d\left(\theta_t, \bar{\theta}_t\right).$$

Expressing the noise through a Taylor expansion. Thanks to Corollary 6.24, we may find a stable tube $\left(\mathbb{T}_t \times \mathbb{T}_t^J\right)$ on $\left(s_t,\, \tilde{J}_t\right)$ suitable for both the RTRL algorithm, and any imperfect RTRL algorithm admitting ϕ as error gauge. As a result, thanks to Lemma 5.9, for $0 \leq t < T_0^{r^*_\Theta}$, θ_t belongs to B^*_Θ, and $\left(s_t,\, \tilde{J}_t\right)$ belongs to $\mathbb{T}_t \times \mathbb{T}_t^J$.

Since all θ_t's belong to B^*_Θ, and since $\mathbb{T}_t \times \mathbb{T}_t^J$ is also a stable tube for (non-imperfect) RTRL, for all $t \geq 0$, $\left(s_t,\, \bar{J}_t\right)$ also belongs to $\mathbb{T}_t \times \mathbb{T}_t^J$.

As a consequence, for every $t \geq 1$, v_t and $\bar{v}_t$ belong to $B_{\mathcal{V}_t}$, which was introduced just below the statement of Lemma 6.28. As a consequence, thanks to Lemma 7.3 we have, for all $1 \leq t < T_0^{r^*_\Theta}$,

$$\theta_t - \bar{\theta}_t = \sum_{s=1}^{t} \eta_s \, (v_s - \bar{v}_s) + O\left(\sum_{s=1}^{t} \eta_s^2 \, m(s)^2\right),$$

so that

$$d\left(\theta_t,\, \bar{\theta}_t\right) \leq \left\| \sum_{s=1}^{t} \eta_s \, (v_s - \bar{v}_s) \right\| + O\left(\sum_{s=1}^{t} \eta_s^2 \, m(s)^2\right). \tag{18}$$

Control of the noise. The sequences s_t, $\tilde{J}_t$, $\bar{J}_t$, v_t and $\bar{v}_t$ exactly satisfy the recurrence equations appearing in Lemmas 7.12 and 7.15. Therefore, thanks to Lemma 7.15, for all $1 \leq t < T_0^{r^*_\Theta}$,

$$v_t - \bar{v}_t = \sum_{s=1}^{t} \Pi^0_{s,\,t}(s_0, \theta_0) \, E_s + O\left(m(t) \sup_{s \leq t} d\left(\theta_s,\, \theta_0\right)\right).$$

Now, thanks to the last assertion of Lemma 5.13 (which holds both for the exact RTRL algorithm and for an imperfect RTRL algorithm with the same stable tube), we have

$$\sup_{s \leq t} d\left(\theta_s,\, \theta_0\right) \leq \kappa_6 \sum_{s=1}^{t} \eta_s \, m(s),$$

so that

$$v_t - \bar{v}_t = \sum_{s=1}^{t} \Pi^0_{s,\,t}(s_0, \theta_0) \, E_s + O\left(m(t) \sum_{s=1}^{t} \eta_s \, m(s)\right).$$

As a consequence, for all $1 \leq t < T_0^{r^*_\Theta}$,

$$\sum_{s=1}^{t} \eta_s \, (v_s - \bar{v}_s) = \sum_{s=1}^{t} \eta_s \sum_{p=1}^{s} \Pi^0_{p,\,s}(s_0, \theta_0) \, E_p + O\left(\sum_{s=1}^{t} \eta_s \, m(s) \sum_{p=1}^{s} \eta_p \, m(p)\right). \tag{19}$$

For the O term, we can bound the sum for p up to s by the sum for p up to t, so that the O term is $O\left(\left(\sum_{s=1}^{t} \eta_s \, m(s)\right)^2\right)$. Finally, for all $1 \leq t < T_0^{r^*_\Theta}$,

$$\sum_{s=1}^{t} \eta_s \sum_{p=1}^{s} \Pi^0_{p,\,s}(s_0, \theta_0) \, E_p = \sum_{p=1}^{t} \left(\sum_{s=p}^{t} \eta_s \, \Pi^0_{p,\,s}(s_0, \theta_0)\right) E_p.$$

This concludes the proof. □

The next two statements concern the measurability of algorithm variables along imperfect RTRL trajectories; namely, we track their dependencies with respect to the noise terms E_t.

Lemma 7.18 (Measurability for imperfect RTRL algorithms)**.** *For $t \geq 1$, we denote by $\mathcal{F}_t$ the σ–algebra generated by $\mathcal{F}_0$ and the $(E_s)_{1 \leq s \leq t}$, where $\mathcal{F}_0$ is defined in Assumption 2.18.*

Then $\{\mathcal{F}_t\}_{t \geq 0}$ is a filtration, and Assumption 2.18 rewrites as $\mathbb{E}\left[E_t \middle| \mathcal{F}_{t-1}\right] = 0$, for every $t \geq 1$.

Moreover, for all $t \geq 1$, $\mathcal{F}_t$ contains all the variables s_t, θ_t, $\tilde{J}_t$ and E_t, as well as all the operators $(\mathbf{T}_s)_{s \geq 1}$, $(\mathcal{L}_s)_{s \geq 1}$, $(\mathcal{U}_s)_{s \geq 1}$ and $(\Phi_s)_{s \geq 1}$.

Proof. $\{\mathcal{F}_t\}_{t \geq 0}$ is a filtration by its construction as an increasing sequence of σ-algebras.

Moreover, by Assumption 2.18, $\mathcal{F}_0$ contains θ_0, s_0, $\tilde{J}_0$, as well as all the operators $(\mathbf{T}_t)_{t \geq 1}$, $(\mathcal{L}_t)_{t \geq 1}$, $(\mathcal{U}_t)_{t \geq 1}$ and $(\Phi_t)_{t \geq 1}$.

Then by Definition 2.10, s_1 is $\mathcal{F}_0$–measurable, and therefore, $\mathcal{F}_1$–measurable. Moreover, by Definition 2.10 and Assumption 2.18, $\tilde{J}_1$ is $\mathcal{F}_1$–measurable since E_1 is $\mathcal{F}_1$–measurable. Finally, v_1 and θ_1 are also $\mathcal{F}_1$–measurable by Definition 2.10.

By induction on $t \geq 1$, the property then holds for all $t \geq 1$. □

Corollary 7.19 (Measurability along imperfect trajectories)**.** *Let $k \geq 0$ and let s_{T_k} and θ_{T_k} be the state and parameter obtained by the imperfect RTRL algorithm at time T_k.*

For each $T_k + 1 \leq t \leq T_{k+1}$, define

$$c_t := \sum_{p=t}^{T_{k+1}} \eta_p \, \Pi_{t,\,p}^{T_k}(s_{T_k}, \theta_{T_k}).$$

Then, for all $T_k + 1 \leq t \leq T_{k+1}$, c_t is $\mathcal{F}_{T_k}$–measurable and $\mathcal{F}_{t-1}$–measurable, where $\{\mathcal{F}_t\}_{t \geq 0}$ is the filtration from Assumption 2.18.

Proof. Let $T_k + 1 \leq t \leq T_{k+1}$. As we saw in Lemma 7.18, s_{T_k} and θ_{T_k} are $\mathcal{F}_{T_k}$–measurable.

By Definition 7.14, for each $T_k \leq t \leq p$, the operator $\Pi_{t,\,p}^{T_k}(s_{T_k}, \theta_{T_k})$ is computed from s_{T_k}, θ_{T_k}, $\mathcal{L}_p$, P_p, the states $s_l = \mathbf{T}_{T_k:l}\left(s_{T_k}, \theta_{T_k}\right)$, and the family of operators $(\mathbf{T}_u)_{u \geq 1}$.

Again thanks to the Lemma 7.18, the operators $\mathcal{L}_p$, P_p (defined by $\mathcal{U}_p$), and all the operators $(\mathbf{T}_u)_{u \geq 1}$, are $\mathcal{F}_{T_k}$–measurable. The compound operator $\mathbf{T}_{T_k:l}$ is $\mathcal{F}_{T_k}$–measurable too, as a composition of operators $(\mathbf{T}_u)$.

Since s_{T_k}, θ_{T_k}, and $\mathbf{T}_{T_k:l}$ are $\mathcal{F}_{T_k}$–measurable, the states $s_l = \mathbf{T}_{T_k:l}\left(s_{T_k}, \theta_{T_k}\right)$ are $\mathcal{F}_{T_k}$–measurable. This proves that all objects defining c_t are $\mathcal{F}_{T_k}$–measurable, hence $\mathcal{F}_{t-1}$–measurable since $t - 1 \geq T_k$. □

Here we prove that, under the unbiased noise of Assumption 2.18, the (random) trajectory of an imperfect RTRL algorithm has negligible noise in the sense of Definition 4.25, with arbitrarily high probability.

Lemma 7.20 (Noise control for the imperfect RTRL algorithm)**.** *For all $k \geq 0$, denote $\mathcal{E}_k = \left\{\theta \in \Theta \,\middle|\, d\left(\theta,\, \theta^*\right) \leq \frac{r_\Theta^*}{3}\right\} \times \mathbb{T}_{T_k} \times \mathbb{T}_{T_k}^J$.*

Any imperfect RTRL algorithm (under the unbiasedness and error gauge assumptions of Section 2.4.4) satisfies the following. There exists a sequence (δ_k) tending

to 0, such that, for every $\varepsilon > 0$, there exists $K \geq 0$ such that, for every $\bar{\eta} \leq \overline{\eta}_{\mathcal{V}}$, for any trajectory $(\theta_t)_{t\geq 0}$, $(\mathfrak{m}_t)_{t\geq 0}$, with $\mathfrak{m}_t = \left(s_t,\, \tilde{J}_t\right)$ for all $t \geq 0$, of the imperfect RTRL algorithm, we have

$$P\left(1_{(\theta_{T_k},\, s_{T_k},\, \tilde{J}_{T_k}) \in \mathcal{E}_k}\, D_{T_k:T_{k+1}}\left(\theta_{T_k},\, (\mathfrak{m}_t),\, \boldsymbol{\eta}\right) \leq \delta_k\, \eta_{T_k}\, L(T_k), \quad \forall k \geq K\right) \geq 1 - \varepsilon.$$

Therefore, for any $\varepsilon > 0$, for any $\bar{\eta} \leq \overline{\eta}_{\mathcal{V}}$, with probability at least $1 - \varepsilon$, trajectories of the imperfect RTRL algorithm have negligible noise starting at $K = K(\varepsilon)$, at speed (δ_k) (Def. 4.25).

Proof. Let us fix some $\bar{\eta} \leq \overline{\eta}_{\mathcal{V}}$. Let (θ_t) and $(\mathfrak{m}_t)$ be a learning trajectory of the imperfect RTRL algorithm, with $m_t = (s_t, \tilde{J}_t)$.

Thanks to Corollary 6.11, we may apply Lemma 5.21 to obtain that, for all $k \geq 0$, we have $T_{k+1} \leq T_{T_k}^{r^*_\Theta}(\boldsymbol{\eta})$. Therefore, thanks to Lemma 7.17, for any $k \geq 0$ such that

$$\left(\theta_{T_k},\, s_{T_k},\, \tilde{J}_{T_k}\right) \in \left\{\theta \in \Theta \,\middle|\, d\left(\theta,\, \theta^*\right) \leq \frac{r^*_\Theta}{3}\right\} \times \mathbb{T}_{T_k} \times \mathbb{T}^J_{T_k},$$

we have

$$\begin{aligned} D_{T_k:T_{k+1}}\left(\theta_{T_k},\, (\mathfrak{m}_t),\, \boldsymbol{\eta}\right) &\leq \left\| \sum_{s=T_k+1}^{T_{k+1}} \left(\sum_{p=s}^{T_{k+1}} \eta_p\, \Pi^{T_k}_{s,p}(s_{T_k}, \theta_{T_k}) \right) E_s \right\| + O\left(\left(\sum_{s=T_k+1}^{T_{k+1}} \eta_s\, m(s) \right)^2 \right) \\ &= N_k + O\left(\eta^2_{T_k}\, L(T_k)^2\, m(T_k)^2\right) \end{aligned}$$

by the fifth point of Corollary 5.18 (remembering we use $m_{\mathrm{H}}(\cdot) = m(\cdot)$), and where we have introduced

$$N_k := \left\| \sum_{t=T_k+1}^{T_{k+1}} \left(\sum_{p=t}^{T_{k+1}} \eta_p\, \Pi^{T_k}_{t,p}(s_{T_k}, \theta_{T_k}) \right) E_t \right\|.$$

We will now bound N_k in probability via its second moment, conditioned on the event that $\left(\theta_{T_k},\, s_{T_k},\, \tilde{J}_{T_k}\right) \in \mathcal{E}_k$.

Expressing the noise term of the upper bound. For $T_k + 1 \leq t \leq T_{k+1}$, define

$$c_t := \sum_{p=t}^{T_{k+1}} \eta_p\, \Pi^{T_k}_{t,p}(s_{T_k}, \theta_{T_k})$$

so that $N_k = \left\| \sum_{t=T_k+1}^{T_{k+1}} c_t\, E_t \right\|$. Note that $c_t\, E_t \in \Theta$ by definition of c_t and of the operators $\Pi^{T_k}_{t,p}$ (Def. 7.14). Then

$$N_k^2 = \left\| \sum_{t=T_k+1}^{T_{k+1}} c_t\, E_t \right\|^2 = \sum_{t=T_k+1}^{T_{k+1}} \sum_{s=T_k+1}^{T_{k+1}} \langle c_t\, E_t, c_s\, E_s \rangle.$$

As a result,

$$\mathbb{E}\left[N_k^2 \middle| \mathcal{F}_{T_k} \right] = \sum_{t=T_k+1}^{T_{k+1}} \sum_{s=T_k+1}^{T_{k+1}} \mathbb{E}\left[\langle c_t\, E_t, c_s\, E_s \rangle \middle| \mathcal{F}_{T_k} \right].$$

Now, thanks to the unbiasedness assumption for E_t, the cross-terms $s \neq t$ vanish: indeed, for every $T_k + 1 \leq s < t \leq T_{k+1}$, c_t, c_s and E_s are $\mathcal{F}_{t-1}$–measurable by Corollary 7.19 and because $t - 1 \geq s$, and therefore,

$$\begin{aligned}\mathbb{E}\left[\langle c_t\, E_t, c_s\, E_s\rangle \middle| \mathcal{F}_{T_k}\right] &= \mathbb{E}\left[\mathbb{E}\left[\langle c_t\, E_t, c_s\, E_s\rangle \middle| \mathcal{F}_{t-1}\right] \middle| \mathcal{F}_{T_k}\right] \\ &= \mathbb{E}\left[\langle c_t\, \mathbb{E}\left[E_t \middle| \mathcal{F}_{t-1}\right], c_s\, E_s\rangle \middle| \mathcal{F}_{T_k}\right] \\ &= 0,\end{aligned}$$

thanks to Assumption 2.18. As a consequence,

$$\mathbb{E}\left[N_k^2 \middle| \mathcal{F}_{T_k}\right] = \sum_{t=T_k+1}^{T_{k+1}} \mathbb{E}\left[\|c_t\, E_t\|^2 \middle| \mathcal{F}_{T_k}\right] \leq \sum_{t=T_k+1}^{T_{k+1}} \|c_t\|_{\mathrm{op}}^2\, \mathbb{E}\left[\|E_t\|_{\mathrm{op}}^2 \middle| \mathcal{F}_{T_k}\right],$$

again because c_t is $\mathcal{F}_{T_k}$-measurable (Corollary 7.19), and because $\|c_t\, E_t\| \leq \|c_t\|_{\mathrm{op}}\, \|E_t\|_{\mathrm{op}}$ for $E_t \in \mathrm{L}(\Theta, \mathcal{S}_t)$ and $c_t \in \mathrm{L}(\mathrm{L}(\Theta, \mathcal{S}_t), \Theta)$.

Control of the $\|E_t\|_{\mathrm{op}}$'s. By Corollary 6.24, trajectories of the imperfect RTRL algorithm preserve the stable tube for s and $\tilde{J}$. Therefore, if $\left(\theta_{T_k}, s_{T_k}, \tilde{J}_{T_k}\right) \in \mathcal{E}_k$, then by Lemmas 5.10 and 5.21, θ_t, s_t and $\tilde{J}_t$ stay in the stable tube at least up to time T_{k+1}. Since the stable tube is uniformly bounded, the sequence $\tilde{J}_t$ is bounded for $T_k \leq t \leq T_{k+1}$. As a result, thanks to Assumption 2.21 and Lemma 6.12, the sequence (E_t) is bounded in this interval (again, conditioned on $\left(\theta_{T_k}, s_{T_k}, \tilde{J}_{T_k}\right) \in \mathcal{E}_k$). Therefore,

$$\mathbb{E}\left[N_k^2 \middle| \mathcal{F}_{T_k}\right] = O\left(\sum_{t=T_k+1}^{T_{k+1}} \|c_t\|_{\mathrm{op}}^2\right).$$

Control of the c_t's. Now, for all $T_k + 1 \leq t \leq T_{k+1}$, by definition of c_t,

$$\|c_t\|_{\mathrm{op}} \leq \sum_{p=t}^{T_{k+1}} \eta_p \left\|\Pi_{t,p}^{T_k}(s_{T_k}, \theta_{T_k})\right\|_{\mathrm{op}},$$

so that

$$\|c_t\|_{\mathrm{op}}^2 \leq \left(\sum_{p=t}^{T_{k+1}} \eta_p \left\|\Pi_{t,p}^{T_k}(s_{T_k}, \theta_{T_k})\right\|_{\mathrm{op}}\right)^2.$$

Thanks to Lemma 7.16,

$$\left\|\Pi_{t,p}^{T_k}(s_{T_k}, \theta_{T_k})\right\|_{\mathrm{op}} \leq M\, m(t)\, (1 - \alpha/2)^{(p-t)/h}$$

for some constant $M > 0$, where α and h are the spectral radius constants of Assumption 2.13. Therefore,

$$\|c_t\|_{\mathrm{op}}^2 \leq M^2 \left(\sum_{p=t}^{T_{k+1}} \eta_p\, m(p)\, (1 - \alpha/2)^{\frac{p-t}{h}}\right)^2.$$

Thanks to Schwarz's inequality, writing $\kappa_\alpha = \sum_{m \geq 0} (1 - \alpha/2)^{\frac{m}{h}} < \infty$, we have

$$\begin{aligned}\|c_t\|_{\mathrm{op}}^2 &\leq M^2 \sum_{n=t}^{T_{k+1}} (\eta_n\, m(n))^2\, (1 - \alpha/2)^{\frac{n-t}{h}} \sum_{m=t}^{T_{k+1}} (1 - \alpha/2)^{\frac{m-t}{h}} \\ &\leq M^2\, \kappa_\alpha \sum_{n=t}^{T_{k+1}} (\eta_n\, m(n))^2\, (1 - \alpha/2)^{\frac{n-t}{h}},\end{aligned}$$

so that

$$\begin{aligned}\sum_{t=T_k+1}^{T_{k+1}} \|c_t\|_{\mathrm{op}}^2 &\leq M^2\,\kappa_\alpha \sum_{t=T_k+1}^{T_{k+1}} \sum_{n=t}^{T_{k+1}} (\eta_n\, m(n))^2\, (1-\alpha/2)^{\frac{n-t}{h}} \\ &= M^2\,\kappa_\alpha \sum_{n=T_k+1}^{T_{k+1}} (\eta_n\, m(n))^2 \sum_{t=T_k+1}^{n} (1-\alpha/2)^{\frac{n-t}{h}} \\ &\leq M^2\,\kappa_\alpha^2 \sum_{n=T_k+1}^{T_{k+1}} (\eta_n\, m(n))^2\,.\end{aligned}$$

Upper-bounding the noise term, and applying Borel-Cantelli's lemma. As a consequence,

$$\mathbb{E}\left[N_k^2 \middle| \mathcal{F}_{T_k}\right] = O\left(\sum_{t=T_k+1}^{T_{k+1}} \eta_t^2\, m(t)^2\right) = O\left(\eta_{T_k}^2\, m(T_k)^2\, L(T_k)\right),$$

as k tends to infinity, thanks to the sixth point of Corollary 5.18. Moreover, this bound is uniform over $\left\{\theta \in \Theta \,\middle|\, d(\theta,\, \theta^*) \leq \frac{r_\Theta^*}{3}\right\} \times \mathbb{T}_{T_k} \times \mathbb{T}_{T_k}^J$ and over $\overline{\eta} \leq \overline{\eta}_{\mathcal{V}}$, because all of the intermediate results we used are.

Let us now define a sequence (δ_k) by

$$\delta_k := k^{-\frac{A-\gamma-1/2}{2(1-A)}}$$

where the exponents are those appearing in Lemma 6.6. By this lemma, $A > \gamma + 1/2$, so that δ_k tends to 0.

Then, for any $k \geq 0$, thanks to Bienaymé–Chebyshev's inequality, we have

$$\begin{aligned}&P\left(1_{(\theta_{T_k},\, s_{T_k},\, \tilde{J}_{T_k}) \in \mathcal{E}_k}\, N_k \geq \delta_k\, \eta_{T_k}\, L(T_k)\right) \\ &\qquad \leq \frac{1}{\delta_k^2\, \eta_{T_k}^2\, L(T_k)^2}\, \mathbb{E}\left[1_{(\theta_{T_k},\, s_{T_k},\, \tilde{J}_{T_k}) \in \mathcal{E}_k}\, N_k^2\right] \\ &\qquad = \frac{1}{\delta_k^2\, \eta_{T_k}^2\, L(T_k)^2}\, \mathbb{E}\left[1_{(\theta_{T_k},\, s_{T_k},\, \tilde{J}_{T_k}) \in \mathcal{E}_k}\, \mathbb{E}\left[N_k^2 \middle| \mathcal{F}_{T_k}\right]\right] \\ &\qquad \leq O\left(\frac{m(T_k)^2}{\delta_k^2\, L(T_k)}\right) = O\left(\frac{{T_k}^{2\gamma}}{\delta_k^2\, {T_k}^A}\right),\end{aligned}$$

since $\left(\theta_{T_k},\, s_{T_k},\, \tilde{J}_{T_k}\right)$ is $\mathcal{F}_{T_k}$–measurable as we saw in Lemma 7.18, and by the definitions of $m(T)$ and $L(T)$ (Section 6.3).

Now, thanks to Lemma 6.9, we know that $T_k \sim c\, k^{1/(1-A)}$, for some $c > 0$, as k tends to infinity. As a result, we have

$$p_k := \frac{m(T_k)^2}{\delta_k^2\, L(T_k)} = \delta_k^{-2}\, {T_k}^{2\gamma - A} \sim c^{2\gamma - A}\, k^{\frac{\gamma - 1/2}{1-A}}$$

by our choice of δ_k. By Lemma 6.6, $\gamma < A - 1/2$ so that $(\gamma - 1/2)/(1-A) < -1$. Therefore, the series $\sum p_k$ converges.

As a result, for all $K \geq 0$, we have

$$P\left(\exists k \geq K \text{ such that } 1_{(\theta_{T_k}, s_{T_k}, \tilde{J}_{T_k}) \in \mathcal{E}_k} N_k \geq \delta_k \eta_{T_k} L(T_k)\right) \leq O\left(\sum_{k \geq K} p_k\right).$$

Let then $\varepsilon > 0$, and let K_ε such that the upper bound $O\left(\sum_{k \geq K_\varepsilon} p_k\right)$ is less than ε. Since all the bounds are uniform over $\bar{\eta} \leq \overline{\eta}_{\mathcal{V}}$, K_ε is independent of $\bar{\eta}$. Thus, for any $\bar{\eta} \leq \overline{\eta}_{\mathcal{V}}$,

$$P\left(1_{(\theta_{T_k}, s_{T_k}, \tilde{J}_{T_k}) \in \mathcal{E}_k} N_k \leq \delta_k \eta_{T_k} L(T_k), \quad \forall k \geq K_\varepsilon\right) \geq 1 - \varepsilon.$$

Remember that $D_{T_k:T_{k+1}}(\theta_{T_k}, (\mathfrak{m}_t), \boldsymbol{\eta}) = N_k + O\left(\eta_{T_k}^2 L(T_k)^2 m(T_k)^2\right)$. Thus, on the set $\left\{1_{(\theta_{T_k}, s_{T_k}, \tilde{J}_{T_k}) \in \mathcal{E}_k} N_k \leq \delta_k \eta_{T_k} L(T_k), \ \forall k \geq K_\varepsilon\right\}$ we have, for every $k \geq K_\varepsilon$,

$$1_{(\theta_{T_k}, s_{T_k}, \tilde{J}_{T_k}) \in \mathcal{E}_k} D_{T_k:T_{k+1}}(\theta_{T_k}, (\mathfrak{m}_t), \boldsymbol{\eta}) \leq \left(\delta_k + O\left(\eta_{T_k} L(T_k) m(T_k)^2\right)\right) \eta_{T_k} L(T_k).$$

Lemma 6.6 shows that $\eta_{T_k} L(T_k) m(T_k)^2$ converges to 0, as k tends to infinity. Moreover, as before, the constants of the O term only depend on the constants of the problem, and are uniform with respect to $\bar{\eta} \leq \overline{\eta}_{\mathcal{V}}$. Let us write, for $k \geq 0$, $\tilde{\delta}_k := \delta_k + O\left(\eta_{T_k} L(T_k) m(T_k)^2\right)$. Therefore, for every $\bar{\eta} \leq \overline{\eta}_{\mathcal{V}}$, we have

$$P\left(1_{(\theta_{T_k}, s_{T_k}, \tilde{J}_{T_k}) \in \mathcal{E}_k} D_{T_k:T_{k+1}}(\theta_{T_k}, (\mathfrak{m}_t), \boldsymbol{\eta}) \leq \tilde{\delta}_k \eta_{T_k} L(T_k), \quad \forall k \geq K_\varepsilon\right) \geq 1 - \varepsilon.$$

We have therefore established the claim. □

7.5 Convergence of the RTRL Algorithm, Imperfect RTRL Algorithms, and of the TBPTT Algorithm

Corollary 7.21 (Convergence of RTRL, extended RTRL algorithms, and imperfect RTRL algorithms)**.** *Theorem 2.28 holds.*

Proof. Remember that Definition 6.1 casts the operators associated with any extended RTRL algorithm in the abstract framework of Section 4. Thus, the proof consists in showing that extended RTRL algorithms and imperfect RTRL algorithms satisfy all the assumptions of Theorem 4.27 and Theorem 4.28 respectively, albeit on a smaller ball $B_\Theta^* \subset \Theta$ than the ball on which the assumptions of Section 2 hold (because the ball B_Θ^* has been reduced several times in the proofs above).

In Section 4, we have a (non-imperfect) algorithm $\mathcal{A}$ that has to satisfy a series of assumptions. For extended RTRL algorithms, we check these assumptions directly.

Imperfect RTRL algorithms are treated somewhat differently. For every imperfect RTRL algorithm, there is a corresponding (non-imperfect) extended RTRL algorithm obtained by setting $E_t = 0$ in the definition (namely, Defs. 2.9 and 2.10 with the same system and update operators, respectively without or with noise E_t). Imperfect RTRL algorithms are random trajectories in the sense of Definition 4.23; then we apply Theorem 4.28. For this, imperfect RTRL algorithms do not need to satisfy all the assumptions of Section 4, only to respect the stable tube (Def. 4.23), and to have negligible noise (Def. 4.25) with respect to the corresponding non-imperfect RTRL algorithm. The corresponding non-imperfect algorithm does have to satisfy the assumptions of Section 4.

We now turn to each of the assumptions of Section 4.

Thanks to Corollary 6.24, both extended RTRL algorithms and imperfect RTRL algorithms admit a stable tube, thus satisfying Assumption 4.11. In particular, imperfect RTRL algorithms with random noise produce random trajectories which respect the stable tube in the sense of Definition 4.23.

Thanks to Lemma 6.26 and Corollary 6.27, the exponential forgetting and Lipschitz Assumptions 4.14 and 4.13 are satisfied for extended RTRL algorithms.

Lemma 6.28 proves that the gradient computation operators are bounded on the stable tube, and Lipschitz, so that Assumptions 4.15 and 4.16 are satisfied.

The parameter update operators are Lipschitz, thanks to Proposition 7.1, so that Assumption 4.17 is satisfied.

Thanks to Assumption 2.26, Lemma 6.6 and Lemma 6.10, Assumption 4.19 on the stepsize sequence is satisfied.

Thanks to Lemmas 7.7 and 7.11, θ^* satisfies the local optimality conditions of Assumption 4.18.

Thanks to Lemma 7.20, there exists a sequence (δ_k) tending to 0 such that, for any $\varepsilon > 0$, there exists $K(\varepsilon)$ such that, for any $\bar{\eta} \leq \overline{\eta}_{\mathcal{V}}$ ($\overline{\eta}_{\mathcal{V}} > 0$ is introduced in Proposition 7.1), with probability greater than $1 - \varepsilon$, imperfect RTRL algorithms have negligible noise starting at $K(\varepsilon)$, at speed (δ_k).

Set $\mathcal{N}_{\theta^*} := \left\{ \theta \in \Theta \,\middle|\, d(\theta, \theta^*) \leq \frac{r^*_\Theta}{4} \right\}$ where r^*_Θ is the radius of B^*_Θ, $\mathcal{N}_{s_0^*} := \mathbb{T}_0$, and $\mathcal{N}_0^J := \mathbb{T}_0^J$. Then, thanks to Theorem 4.27, there exists $\overline{\eta}_{\text{conv}} > 0$ such that an extended RTRL algorithm initialized any parameter $\theta_0 \in \mathcal{N}_{\theta^*}$, any state $s_0 \in \mathcal{N}_{s_0^*}$ and any differential $\tilde{J}_0 \in \mathcal{N}_0^J$, produces a sequence of parameters θ_t that converges to θ^*. Moreover, by Theorem 4.28, for any $\varepsilon > 0$, there exists $\overline{\eta}_{\text{conv}} > 0$ such that, for any overall learning rate $\overline{\eta} \leq \overline{\eta}_{\text{conv}}$, any initial parameter $\theta_0 \in \mathcal{N}_{\theta^*}$, any initial state $s_0 \in \mathcal{N}_{s_0^*}$ and any initial differential $\tilde{J}_0 \in \mathcal{N}_0^J$, with probability at least $1 - \varepsilon$, an imperfect RTRL algorithm produces a sequence of parameters θ_t converging to θ^*. □

We now turn to truncated backpropagation through time (TBPTT), which uses essentially the same approach using the open-loop algorithm.

We start with the following classical result: on a finite interval $[T_k; T_{k+1}]$, TBPTT is equivalent to an RTRL algorithm that updates the parameter only at the end of the interval and initializes J to 0 at the beginning of the interval. However, technically we cannot define J_{T_k} to 0 at the start of each interval, because its value is used to compute the gradient at the end of the previous interval. So we just define the next value of J as if $J_{T_k} = 0$ in the formula below.

Proposition 7.22 (TBPTT as RTRL on intervals)**.** *The TBPTT algorithm (Def. 3.13) is equivalent to RTRL with the parameter updated on steps T_k, and the influence of J_t on J_{t+1} cut at time T_k, namely:*

$$\begin{cases} s_t = \mathbf{T}_t(s_{t-1}, \theta_{T_k}), & T_k + 1 \leq t \leq T_{k+1} \\ J_{T_k+1} = \dfrac{\partial \mathbf{T}_{T_k+1}(s_{T_k}, \theta_{T_k})}{\partial \theta}, & \\ J_t = \dfrac{\partial \mathbf{T}_t(s_{t-1}, \theta_{T_k})}{\partial s} J_{t-1} + \dfrac{\partial \mathbf{T}_t(s_{t-1}, \theta_{T_k})}{\partial \theta}, & T_k + 2 \leq t \leq T_{k+1} \\ v_t = \dfrac{\partial \mathcal{L}_t(s_t)}{\partial s} \cdot J_t, & T_k + 1 \leq t \leq T_{k+1}, \end{cases}$$

and parameter update

$$\theta_{T_{k+1}} = \theta_{T_k} - \eta_{T_{k+1}} \sum_{t=T_k+1}^{T_{k+1}} v_t.$$

Proof. The proof is classical (as long as the parameter is not updated, RTRL and backpropagation through time compute the same gradient, by equivalence of forward and backward gradient computations), and we omit it. □

Corollary 7.23 (Convergence of the TPBTT algorithm)**.** *Theorem 3.14 holds.*

Proof. The proof consists in showing the TBPTT algorithm satisfies the assumptions of Theorem 4.29.

Remember that Definition 6.1 defines the abstract operators $\mathcal{A}_t$ and $\boldsymbol{V}_t$ for the RTRL algorithm.

Then one checks that the algorithm in Proposition 7.22 is equivalent to the open-loop algorithm studied in Theorem 4.29, with $\mathcal{A}_t$ the RTRL update on $\mathfrak{m} = (s, J)$ from Definition 6.1, with $\boldsymbol{V}_t$ as in Definition 6.1 using $\mathcal{U}_t(v, s, \theta) = v$, with $\Phi_t(\theta, v) = \theta - v$, and with

$$\mathfrak{m}'_{T_k} := (s_{T_k}, 0).$$

Indeed, in that case, the update

$$\mathfrak{m}_{T_k} \leftarrow \mathfrak{m}'_{T_k} \in \mathbb{T}_{\mathcal{M}_{T_k}}$$

at the beginning of each interval, is equivalent to resetting J to 0 at the beginning of each interval. (Note that resetting the value of J to 0 stays in the stable tube, thanks to Corollary 6.24.) Moreover, since $\Phi_t(\theta, v) = \theta - v$, the iterated update $\theta_t = \Phi_t\left(\theta_{t-1}, \eta_{T_{k+1}} v_t\right)$ in Theorem 4.29 is equivalent to

$$\theta_{T_{k+1}} = \theta_{T_k} - \sum_{t=T_k+1}^{T_{k+1}} \eta_{T_{k+1}} v_t,$$

which is indeed the parameter update of Proposition 7.22.

Moreover, the interval lengths $T_{k+1} - T_k = {T_k}^A$ in the assumptions of Theorem 3.14 match the definition of T_k in Lemma 6.9, with the constraint on A from Lemma 6.6.

The assumptions of Theorem 4.29 are the same as those of Theorem 4.27; we have checked above in the proof of Corollary 7.21 that these assumptions are satisfied for RTRL.

Thus, by Theorem 4.29 and Proposition 7.22, the TBPTT algorithm produces a sequence of parameters θ_{T_k} converging to θ^*. □

7.6 NoBackTrack and UORO as Imperfect RTRL Algorithms

We now prove Lemma 3.22: NoBackTrack and UORO satisfy the unbiasedness and bounded noise assumptions.

The notation for NoBackTrack and UORO is introduced in Section 3.3.1. Remember that the noise E_t is defined via random signs $\varepsilon(t)$ at each time t. For every $t \geq 1$, we write $\mathcal{F}'_t$ the σ-algebra generated by $\mathcal{F}_0$ (defined in Assumption 2.18) together with the $\varepsilon(s)$ for $1 \leq s \leq t$. Since E_t is computed from the $\varepsilon(t)$'s, the

σ-algebra $\mathcal{F}_t$ generated by the $(E_s)_{s\leq t}$ is contained in $\mathcal{F}'_t$: $\mathcal{F}_t \subset \mathcal{F}'_t$. For Assumption 2.18, we want to prove that $\mathbb{E}\left[E_t \middle| \mathcal{F}_{t-1}\right] = 0$; we will prove the stronger result that $\mathbb{E}\left[E_t \middle| \mathcal{F}'_{t-1}\right] = 0$.

Computing the conditional expectation with respect to $\mathcal{F}'_t$ means integrating with respect to the laws of the $\varepsilon(s)$'s, for $s > t$.

We are going to study the error E_t at some time $t \geq 1$. In this section, we fix the following abbreviations. For $t \geq 1$, let the values of the NoBackTrack objects at time $t-1$ be $\theta = \theta_{t-1} \in \Theta$, $s = s_{t-1} \in \mathcal{S}_{t-1}$ and $J = \tilde{J}_{t-1} \in \mathrm{Rk}_1(\Theta, \mathcal{S}_{t-1})$ with $J = v^{\mathcal{S}} \otimes v^{\Theta}$, and further abbreviate

$$a := \frac{\partial \mathbf{T}_t}{\partial s}(s, \theta) \quad \text{and} \quad b_i := \frac{\partial \mathbf{T}_t^i}{\partial \theta}(s, \theta), \quad 1 \leq i \leq \dim \mathcal{S}_t$$

where $\mathbf{T}_t^i$ is the i-th component of $\mathbf{T}_t$ in the basis of $\mathcal{S}_t$ used to define NoBackTrack. Finally, abbreviate ε_i for $\varepsilon_i(t)$.

7.6.1 NoBackTrack as an Imperfect RTRL Algorithm

Let us first express the NoBackTrack update. Let

$$\left(w^{\mathcal{S}}, w^{\Theta}\right) := \mathcal{R}_t\left(v^{\mathcal{S}}, v^{\Theta}, s, \theta\right),$$

where $\mathcal{R}_t$ is the reduction operator for NoBackTrack defined in Section 3.3.1. Then abbreviate

$$\rho := \sqrt{\frac{\|v^{\Theta}\|}{\|a\left(v^{\mathcal{S}}\right)\|}} \quad \text{and} \quad \rho_i := \sqrt{\frac{\|b_i\|}{\|\mathfrak{e}_i\|}}, \quad 1 \leq i \leq \dim \mathcal{S}_t,$$

with $\mathfrak{e}_i$ the basis vectors used to define NoBackTrack. If $a\left(v^{\mathcal{S}}\right) = 0$, we define $\rho := 1$.

Then by definition of the NoBackTrack reduction operator $\mathcal{R}_t$, we have

$$\begin{cases} w^{\mathcal{S}} = \rho\, a\left(v^{\mathcal{S}}\right) + \displaystyle\sum_{i=1}^{\dim \mathcal{S}_t} \varepsilon_i\, \rho_i\, \mathfrak{e}_i \\ w^{\Theta} = \rho^{-1}\, v^{\Theta} + \displaystyle\sum_{i=1}^{\dim \mathcal{S}_t} \varepsilon_i\, \rho_i^{-1}\, b_i. \end{cases}$$

By induction and by Lemma 7.18, $\theta = \theta_{t-1}$, $s = s_{t-1}$, $J = \tilde{J}_{t-1}$ and thus a, b, ρ, and ρ_i are $\mathcal{F}_{t-1}$-measurable hence $\mathcal{F}'_{t-1}$-measurable. (We assume the basis $\mathfrak{e}_i$ is deterministic.) Thus $w^{\mathcal{S}}$ and w^{Θ} are $\mathcal{F}'_t$-measurable, since they also use the ε_i's at time t. Note that in NoBackTrack, $v^{\mathcal{S}}$ and v^{Θ} at time t are $w^{\mathcal{S}}$ and w^{Θ} from the previous step $t-1$, so that $v^{\mathcal{S}}$ and v^{Θ} are $\mathcal{F}'_{t-1}$-measurable.

As a result, from the definition of NoBackTrack (Def. 3.20), the imperfect Jacobian computed by the NoBackTrack update is

$$\begin{aligned} \tilde{J}_t = w^{\mathcal{S}} \otimes w^{\Theta} = a\left(v^{\mathcal{S}}\right) \otimes v^{\Theta} + \sum_{i=1}^{\dim \mathcal{S}_t} \underbrace{\varepsilon_i^2}_{=1}\, \mathfrak{e}_i \otimes b_i \\ + \rho\, a\left(v^{\mathcal{S}}\right) \otimes \sum_{i=1}^{\dim \mathcal{S}_t} \varepsilon_i\, \rho_i^{-1}\, b_i + \rho^{-1} v^{\Theta} \otimes \sum_{i=1}^{\dim \mathcal{S}_t} \varepsilon_i\, \rho_i\, \mathfrak{e}_i \\ + \sum_{i,j=1,\, i\neq j}^{\dim \mathcal{S}_t} \varepsilon_i\, \rho_i\, \mathfrak{e}_i \otimes \varepsilon_j\, \rho_j^{-1}\, b_j, \end{aligned}$$

and the error term E_t is

$$\begin{aligned} E_t &= \tilde{J}_t - \left(\frac{\partial \mathbf{T}_t}{\partial s}(s,\,\theta) \cdot J + \frac{\partial \mathbf{T}_t}{\partial \theta}(s,\,\theta) \right) \\ &= w^{\mathcal{S}} \otimes w^{\Theta} - \left(a\left(v^{\mathcal{S}}\right) \otimes v^{\Theta} + b \right) \\ &= w^{\mathcal{S}} \otimes w^{\Theta} - a\left(v^{\mathcal{S}}\right) \otimes v^{\Theta} - \sum_{i=1}^{\dim \mathcal{S}_t} \mathfrak{e}_i \otimes b_i, \end{aligned}$$

namely

$$\begin{aligned} E_t = \rho\, a\left(v^{\mathcal{S}}\right) \otimes \sum_{i=1}^{\dim \mathcal{S}_t} \varepsilon_i\, \rho_i^{-1}\, b_i + \rho^{-1} v^{\Theta} \otimes \sum_{i=1}^{\dim \mathcal{S}_t} \varepsilon_i\, \rho_i\, \mathfrak{e}_i \\ + \sum_{i,j=1,\, i\neq j}^{\dim \mathcal{S}_t} \varepsilon_i\, \rho_i\, \mathfrak{e}_i \otimes \varepsilon_j\, \rho_j^{-1}\, b_j. \end{aligned} \tag{20}$$

NoBackTrack – unbiasedness of the Jacobian update rule. Let us show that Assumption 2.18 is satisfied with the filtration $\{\mathcal{F}'_t\}$, hence a fortiori with $\{\mathcal{F}_t\}$.

By construction, at each time t the ε_i's are random Bernoulli variables that are independent from $\mathcal{F}'_{t-1}$. Thus, $\mathbb{E}\left[\varepsilon_i \middle| \mathcal{F}'_{t-1}\right] = 0$ and $\mathbb{E}\left[\varepsilon_i \varepsilon_j \middle| \mathcal{F}'_{t-1}\right] = 0$ for $i \neq j$.

On the other hand, we have seen that all the other variables in the expression (20) for E_t are $\mathcal{F}'_{t-1}$-measurable. Then, taking the conditional expectation with respect to $\mathcal{F}'_{t-1}$ in (20), we obtain $\mathbb{E}\left[E_t \middle| \mathcal{F}'_{t-1}\right] = 0$ as needed. □

NoBackTrack – size of the error. Next, we deal with Assumption 2.21 (bounded errors) for NoBackTrack.

From the expression (20) for E_t, since for vectors v_1, v_2, we have $\|v_1 \otimes v_2\|_{\mathrm{op}} = \|v_1\|\,\|v_2\|$, it holds that

$$\begin{aligned} \|E_t\|_{\mathrm{op}} &\leq \rho \left\|a\left(v^{\mathcal{S}}\right)\right\| \sum_{i=1}^{\dim \mathcal{S}_t} \rho_i^{-1}\, \|b_i\| + \rho^{-1} \left\|v^{\Theta}\right\| \sum_{i=1}^{\dim \mathcal{S}_t} \rho_i\, \|\mathfrak{e}_i\| \\ &\quad + \sum_{i,j=1,\, i\neq j}^{\dim \mathcal{S}_{t+1}} \rho_i\, \|\mathfrak{e}_i\| \rho_j^{-1}\, \|b_j\| \\ &= \sqrt{\|a\left(v^{\mathcal{S}}\right)\|\,\|v^{\Theta}\|} \sum_{i=1}^{\dim \mathcal{S}_t} \sqrt{\|\mathfrak{e}_i\|\,\|b_i\|} + \sqrt{\|a\left(v^{\mathcal{S}}\right)\|\,\|v^{\Theta}\|} \sum_{i=1}^{\dim \mathcal{S}_t} \sqrt{\|\mathfrak{e}_i\|\,\|b_i\|} \\ &\quad + \sum_{i,j=1,\, i\neq j}^{\dim \mathcal{S}_{t+1}} \sqrt{\|\mathfrak{e}_i\|\,\|b_i\|}\sqrt{\|\mathfrak{e}_j\|\,\|b_j\|} \\ &\leq 2\sqrt{\|a\|_{\mathrm{op}}}\sqrt{\|v^{\mathcal{S}}\|\,\|v^{\Theta}\|} \sum_{i=1}^{\dim \mathcal{S}_t} \sqrt{\|\mathfrak{e}_i\|\,\|b_i\|} + \sum_{i,j=1,\, i\neq j}^{\dim \mathcal{S}_t} \sqrt{\|\mathfrak{e}_i\|\,\|b_i\|}\sqrt{\|\mathfrak{e}_j\|\,\|b_j\|}. \end{aligned}$$

Now, $J = v^{\mathcal{S}} \otimes v^{\Theta}$, so $\|J\|_{\mathrm{op}} = \left\|v^{\mathcal{S}}\right\| \left\|v^{\Theta}\right\|$. As a result,

$$\sqrt{\|v^{\mathcal{S}}\|\,\|v^{\Theta}\|} = \|J\|_{\mathrm{op}}^{1/2}.$$

Next, since the $\mathfrak{e}_i$'s form and orthonormal basis of vectors of $\mathcal{S}_t$ according to Definition 3.17, we have $\|\mathfrak{e}_i\| = 1$ and $\|b_i\| \leq \|b\|_{\mathrm{op}}\,\|\mathfrak{e}_i\| = \|b\|_{\mathrm{op}}$. Finally, by definition of a and b,

$$\|a\|_{\mathrm{op}},\, \|b\|_{\mathrm{op}} \leq \left\| \frac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial(s,\theta)} \right\|_{\mathrm{op}}.$$

Plugging this into the bound for $\|E_t\|_{\text{op}}$, we find

$$\|E_t\|_{\text{op}} \leq (2 \dim \mathcal{S}_t)\, y\, \|J\|_{\text{op}}^{1/2} + (\dim \mathcal{S}_t)^2\, y$$

where $y = \left\| \frac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial(s,\theta)} \right\|_{\text{op}}$.

Since $\dim \mathcal{S}_t$ is bounded by Assumption 3.21, this shows the size of the error in NoBackTrack is compliant with the requirements of Definition 2.19 and Assumption 2.21, with $\phi(x, y) = C(1 + x^{1/2})(1 + y)$.

Note that we have only proved that $\|E_t\|$ is controlled by $\sqrt{\|J\|_{\text{op}}}$ provided J is rank-one. This is true by construction at all times along the NoBackTrack trajectory, so the bound above holds at all times on any NoBackTrack trajectory (with probability 1); this is all that is needed for Assumption 2.21. □

7.6.2 UORO as an Imperfect RTRL Algorithm

The analysis of UORO is very similar to NoBackTrack. Let us first express the UORO update. Let

$$\left(w^{\mathcal{S}},\, w^{\Theta}\right) := \mathcal{R}_t\left(v^{\mathcal{S}},\, v^{\Theta},\, s,\, \theta\right),$$

where $\mathcal{R}_t$ is the reduction operator for UORO defined in Section 3.3.1. Then, abbreviate

$$\rho_0 := \sqrt{\frac{\|v^{\Theta}\|}{\|a\,(v^{\mathcal{S}})\|}} \quad \text{and} \quad \rho_1 := \sqrt{\frac{\left\|\sum_{i=1}^{\dim \mathcal{S}_t} \varepsilon_i\, b_i\right\|}{\left\|\sum_{i=1}^{\dim \mathcal{S}_t} \varepsilon_i\, \mathfrak{e}_i\right\|}},$$

where the $\mathfrak{e}_i$'s are the basis vectors used to define UORO. If $a\left(v^{\mathcal{S}}\right) = 0$, we define $\rho := 1$.

Then, by definition of the UORO reduction operator $\mathcal{R}_t$, we have

$$\begin{cases} w^{\mathcal{S}} = \rho_0\, a\left(v^{\mathcal{S}}\right) + \rho_1 \displaystyle\sum_{i=1}^{\dim \mathcal{S}_t} \varepsilon_i\, \mathfrak{e}_i \\ w^{\Theta} = \rho_0^{-1}\, v^{\Theta} + \rho_1^{-1} \displaystyle\sum_{i=1}^{\dim \mathcal{S}_t} \varepsilon_i\, b_i. \end{cases}$$

By induction and by Lemma 7.18, $\theta = \theta_{t-1}$, $s = s_{t-1}$, $J = \tilde{J}_{t-1}$ and thus a, b, ρ_0, and ρ_1 are $\mathcal{F}_{t-1}$-measurable hence $\mathcal{F}'_{t-1}$-measurable. (We assume the basis $\mathfrak{e}_i$ is deterministic.) Thus $w^{\mathcal{S}}$ and w^{Θ} are $\mathcal{F}'_t$-measurable, since they also use the ε_i's at time t. Note that in UORO, $v^{\mathcal{S}}$ and v^{Θ} at time t are $w^{\mathcal{S}}$ and w^{Θ} from the previous step $t-1$, so that $v^{\mathcal{S}}$ and v^{Θ} are $\mathcal{F}'_{t-1}$-measurable.

As a result, from the definition of UORO (Def. 3.20), the imperfect Jacobian computed by the NoBackTrack update is

$$\begin{aligned} \tilde{J}_t = w^{\mathcal{S}} \otimes w^{\Theta} = a\left(v^{\mathcal{S}}\right) \otimes v^{\Theta} + \sum_{i=1}^{\dim \mathcal{S}_t} \underbrace{\varepsilon_i^2}_{=1}\, \mathfrak{e}_i \otimes b_i \\ + \rho_0\, a\left(v^{\mathcal{S}}\right) \otimes \rho_1^{-1} \sum_{i=1}^{\dim \mathcal{S}_t} \varepsilon_i\, b_i + \rho_0^{-1}\, v^{\Theta} \otimes \rho_1 \sum_{i=1}^{\dim \mathcal{S}_t} \varepsilon_i\, \mathfrak{e}_i \\ + \sum_{i,j=1,\, i \neq j}^{\dim \mathcal{S}_t} \varepsilon_i\, \mathfrak{e}_i \otimes \varepsilon_j\, b_j, \end{aligned}$$

and the error term E_t is

$$\begin{aligned} E_t &= \tilde{J}_t - \left(\frac{\partial \mathbf{T}_t}{\partial s}(s,\,\theta) \cdot J + \frac{\partial \mathbf{T}_t}{\partial \theta}(s,\,\theta) \right) \\ &= w^{\mathcal{S}} \otimes w^{\Theta} - \left(a\left(v^{\mathcal{S}}\right) \otimes v^{\Theta} + b \right) \\ &= w^{\mathcal{S}} \otimes w^{\Theta} - a\left(v^{\mathcal{S}}\right) \otimes v^{\Theta} - \sum_{i=1}^{\dim \mathcal{S}_t} \mathfrak{e}_i \otimes b_i, \end{aligned}$$

namely

$$\begin{aligned} E_t = \rho_0\, a\left(v^{\mathcal{S}}\right) \otimes \rho_1^{-1} \sum_{i=1}^{\dim \mathcal{S}_t} \varepsilon_i\, b_i + \rho_0^{-1}\, v^{\Theta} \otimes \rho_1 \sum_{i=1}^{\dim \mathcal{S}_t} \varepsilon_i\, \mathfrak{e}_i \\ + \sum_{i,j=1,\, i\neq j}^{\dim \mathcal{S}_t} \varepsilon_i\, \mathfrak{e}_i \otimes \varepsilon_j\, b_j. \end{aligned} \tag{21}$$

UORO – unbiasedness of the Jacobian update rule. Let us show that Assumption 2.18 is satisfied with the filtration $\{\mathcal{F}'_t\}$, hence a fortiori with $\{\mathcal{F}_t\}$.

By construction, at each time t the ε_i's are random Bernoulli variables that are independent from $\mathcal{F}'_{t-1}$. Thus, $\mathbb{E}\left[\varepsilon_i \middle| \mathcal{F}'_{t-1}\right] = 0$ and $\mathbb{E}\left[\varepsilon_i \varepsilon_j \middle| \mathcal{F}'_{t-1}\right] = 0$ for $i \neq j$.

On the other hand, we have seen that all the other variables in the expression (21) for E_t are $\mathcal{F}'_{t-1}$-measurable. Then, taking the conditional expectation with respect to $\mathcal{F}'_{t-1}$ in (21), we obtain $\mathbb{E}\left[E_t \middle| \mathcal{F}'_{t-1}\right] = 0$ as needed. □

UORO – size of the error. Next, we deal with Assumption 2.21 (bounded errors) for UORO.

From the expression (21) for E_t, since for vectors v_1, v_2, we have $\|v_1 \otimes v_2\|_{\text{op}} = \|v_1\|\,\|v_2\|$, it holds that

$$\begin{aligned} \|E_t\|_{\text{op}} &\leq \rho_0 \left\|a\left(v^{\mathcal{S}}\right)\right\| \rho_1^{-1} \left\| \sum_{i=1}^{\dim \mathcal{S}_{t+1}} \varepsilon_i\, b_i \right\| + \rho_0^{-1} \left\|v^{\Theta}\right\| \rho_1 \left\| \sum_{i=1}^{\dim \mathcal{S}_t} \varepsilon_i\, \mathfrak{e}_i \right\| \\ &\quad + \sum_{i,j=1,\, i\neq j}^{\dim \mathcal{S}_t} \|\mathfrak{e}_i\|\,\|b_j\| \\ &= 2\sqrt{\|a\left(v^{\mathcal{S}}\right)\|\,\|v^{\Theta}\|} \sqrt{\left\| \sum_{i=1}^{\dim \mathcal{S}_{t+1}} \varepsilon_i\, \mathfrak{e}_i \right\| \left\| \sum_{i=1}^{\dim \mathcal{S}_t} \varepsilon_i\, b_i \right\|} \\ &\quad + \sum_{i,j=1,\, i\neq j}^{\dim \mathcal{S}_{t+1}} \|\mathfrak{e}_i\|\,\|b_j\| \\ &\leq 2\sqrt{\|a\|_{\text{op}}}\sqrt{\|v^{\mathcal{S}}\|\,\|v^{\Theta}\|} \sqrt{\left\| \sum_{i=1}^{\dim \mathcal{S}_t} \varepsilon_i\, \mathfrak{e}_i \right\| \left\| \sum_{i=1}^{\dim \mathcal{S}_t} \varepsilon_i\, b_i \right\|} \\ &\quad + \sum_{i,j=1,\, i\neq j}^{\dim \mathcal{S}_t} \|\mathfrak{e}_i\|\,\|b_j\|\,. \end{aligned}$$

Now, $J = v^{\mathcal{S}} \otimes v^{\Theta}$, so $\|J\|_{\text{op}} = \left\|v^{\mathcal{S}}\right\| \left\|v^{\Theta}\right\|$. As a result,

$$\sqrt{\|v^{\mathcal{S}}\|\,\|v^{\Theta}\|} = \|J\|_{\text{op}}^{1/2}\,.$$

Next, since the $\mathfrak{e}_i$'s form and orthonormal basis of vectors of $\mathcal{S}_t$ according to Definition 3.17, we have $\|\mathfrak{e}_i\| = 1$ and $\|b_i\| \leq \|b\|_{\text{op}} \|\mathfrak{e}_i\| = \|b\|_{\text{op}}$. Finally, by definition of a and b,

$$\|a\|_{\text{op}}, \|b\|_{\text{op}} \leq \left\| \frac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial(s, \theta)} \right\|_{\text{op}}.$$

Plugging this into the bound for $\|E_t\|_{\text{op}}$, we find

$$\|E_t\|_{\text{op}} \leq (2 \dim \mathcal{S}_t)\, y\, \|J\|_{\text{op}}^{1/2} + (\dim \mathcal{S}_t)^2\, y$$

where $y = \left\| \frac{\partial \mathbf{T}_t(s_{t-1}, \theta_{t-1})}{\partial(s, \theta)} \right\|_{\text{op}}$.

Since $\dim \mathcal{S}_t$ is bounded by Assumption 3.21, this shows the size of the error in UORO is compliant with the requirements of Definition 2.19 and Assumption 2.21, with $\phi(x, y) = C(1 + x^{1/2})(1 + y)$.

Note that we have only proved that $\|E_t\|$ is controlled by $\sqrt{\|J\|_{\text{op}}}$ provided J is rank-one. This is true by construction at all times along the UORO trajectory, so the bound above holds at all times on any UORO trajectory (with probability 1); this is all that is needed for Assumption 2.21. □

A Positive-Stable Matrices

We recall several equivalent definitions of positive-stable matrices (also known, with signs reversed, as Hurwitz matrices).

Definition A.1 (Positive-stable matrix)**.** *A real matrix A is* positive-stable *if one of the following equivalent conditions is satisfied:*

1. *All the eigenvalues of A have positive real part.*
2. *The solution of the differential equation $\theta' = -A\theta$ converges to 0 for any initial value.*
3. *There exists a symmetric, positive definite matrix B (Lyapunov function) such that $\theta^\top B\theta$ is decreasing along the solutions of the differential equation $\theta' = -A\theta$.*
4. *There exists a symmetric positive definite matrix B such that*

   $$\theta^\top BA\theta > 0$$

   for all $\theta \neq 0$. (This is the same B as in the previous condition.)
5. *There exists a symmetric positive definite matrix B such that $BA + A^\top B$ is positive definite.*

Stability is invariant by matrix similarity $A \leftarrow C^{-1}AC$, since this preserves eigenvalues.

Since the solution of $\theta_t' = -A\theta_t$ is $\theta_t = e^{-tA}\theta_0$, a Lyapunov function that works is $\theta_0^\top B\theta_0 := \int_{t\geq 0} \|\theta_t\|^2$. Indeed this is decreasing, because $\theta_t^\top B\theta_t$ is just the same integral starting at t instead of 0. Explicitly this is $B = \int_{t\geq 0} (e^{-tA})^\top e^{-tA}$. It satisfies $BA + A^\top B = \text{Id}$.

The proofs are classical, and therefore we omit them.

Proposition A.2. *1. A symmetric positive definite matrix is positive-stable. (Take $B = \mathrm{Id}$ above.)*

2. If $A + A^\top$ is positive definite and H is symmetric positive definite, then AH is positive-stable. (Take $B = H$ above.)

B Equicontinuity of the Extended Hessians in the C^3 Case

In this section, which has no bearing on the rest of the proof, we note that equicontinuity of the extended Hessians around θ^* (Assumption 2.25.b), may be deduced in a straightforward way under simpler additional assumptions.

Lemma B.1 (Controlling the derivatives of $\mathbf{s}_t$ and $\mathcal{L}_t$)**.** *Assume that all first, second and third order derivatives of the $\mathbf{T}_t$'s and of the $\mathcal{L}_t$'s exist, and are bounded on a ball around the target trajectory (θ^*, s_t^*). (In particular, we can take $\gamma = 0$.)*

Then all derivatives up to third order of the $\mathbf{s}_t(s_0^*, \cdot)$'s are bounded on B_Θ^*, where B_Θ^* is defined in Lemma 6.18. Likewise, all derivatives up to third order of the $\mathcal{L}_{\rightsquigarrow t}(s_0^*, \cdot)$'s are bounded as well, so that the family of Hessians $\frac{\partial^2 \mathcal{L}_{\rightsquigarrow t}}{\partial \theta^2}(s_0^*, \cdot)$ is equicontinuous on B_Θ^*. In other words, Assumption 2.25.a is satisfied on the smaller ball B_Θ^*.

Proof. Let us first bound the derivatives of the $\mathbf{s}_t$'s.

Bounding the derivatives of the $\mathbf{s}_t$'s. For all $t \geq 1$ and $\theta \in \Theta$, we have

$$\mathbf{s}_t(s_0^*, \theta) = \mathbf{T}_t(\mathbf{s}_{t-1}(s_0^*, \theta), \theta).$$

Let us write, for all $t \geq 1$ and $\theta \in B_\Theta^*$, $s_t = \mathbf{s}_t(s_0^*, \theta)$. Then, for all $t \geq 1$ and $\theta \in B_\Theta^*$, s_t is in the stable tube $\mathbb{T}_t$ of Lemma 6.18. (Note that Assumption 2.25.b is not used for the proof of Lemma 6.18.)

Let $t \geq 1$, and $\theta \in B_\Theta^*$. Then, we have

$$\frac{\partial \mathbf{s}_t}{\partial \theta}(s_0^*, \theta) = \frac{\partial \mathbf{T}_t}{\partial s}(s_{t-1}, \theta) \cdot \frac{\partial \mathbf{s}_{t-1}}{\partial \theta}(s_0^*, \theta) + \frac{\partial \mathbf{T}_t}{\partial \theta}(s_{t-1}, \theta)$$

$$\begin{aligned}\frac{\partial^2 \mathbf{s}_t}{\partial \theta^2}(s_0^*, \theta) = {} & \frac{\partial \mathbf{T}_t}{\partial s}(s_{t-1}, \theta) \cdot \frac{\partial^2 \mathbf{s}_{t-1}}{\partial \theta^2}(s_0^*, \theta) + \frac{\partial^2 \mathbf{T}_t}{\partial s^2}(s_{t-1}, \theta) \cdot \left(\frac{\partial \mathbf{s}_{t-1}}{\partial \theta}(s_0^*, \theta)\right)^{\otimes 2} \\ & + 2\frac{\partial^2 \mathbf{T}_t}{\partial \theta \partial s}(s_{t-1}, \theta) \cdot \left(\frac{\partial \mathbf{s}_{t-1}}{\partial \theta}(s_0^*, \theta), \mathrm{Id}\right) + \frac{\partial^2 \mathbf{T}_t}{\partial \theta^2}(s_{t-1}, \theta)\end{aligned}$$

$$
\begin{aligned}
\frac{\partial^3 \mathbf{s}_t}{\partial \theta^3}(s_0^*, \theta) = & \frac{\partial \mathbf{T}_t}{\partial s}(s_{t-1}, \theta) \cdot \frac{\partial^3 \mathbf{s}_{t-1}}{\partial \theta^3}(s_0^*, \theta) \\
& + \frac{\partial^2 \mathbf{T}_t}{\partial s^2}(s_{t-1}, \theta) \cdot \left(\frac{\partial^2 \mathbf{s}_{t-1}}{\partial \theta^2}(s_0^*, \theta), \frac{\partial \mathbf{s}_{t-1}}{\partial \theta}(s_0^*, \theta) \right) \\
& + \frac{\partial^2 \mathbf{T}_t}{\partial \theta \partial s}(s_{t-1}, \theta) \cdot \left(\frac{\partial^2 \mathbf{s}_{t-1}}{\partial \theta^2}(s_0^*, \theta), \mathrm{Id} \right) \\
& + \frac{\partial^3 \mathbf{T}_t}{\partial s^3}(s_{t-1}, \theta) \cdot \left(\frac{\partial \mathbf{s}_{t-1}}{\partial \theta}(s_0^*, \theta) \right)^{\otimes 3} \\
& + \frac{\partial^3 \mathbf{T}_t}{\partial \theta \partial s^2}(s_{t-1}, \theta) \cdot \left(\left(\frac{\partial \mathbf{s}_{t-1}}{\partial \theta}(s_0^*, \theta) \right)^{\otimes 2}, \mathrm{Id} \right) \\
& + 2 \frac{\partial^2 \mathbf{T}_t}{\partial s^2}(s_{t-1}, \theta) \cdot \left(\frac{\partial^2 \mathbf{s}_{t-1}}{\partial \theta^2}(s_0^*, \theta), \frac{\partial \mathbf{s}_{t-1}}{\partial \theta}(s_0^*, \theta) \right) \\
& + 2 \frac{\partial^3 \mathbf{T}_t}{\partial \theta \partial s^2}(s_{t-1}, \theta) \cdot \left(\left(\frac{\partial \mathbf{s}_{t-1}}{\partial \theta}(s_0^*, \theta) \right)^{\otimes 2}, \mathrm{Id} \right) \\
& + 3 \frac{\partial^3 \mathbf{T}_t}{\partial \theta^2 \partial s}(s_{t-1}, \theta) \cdot \left(\frac{\partial \mathbf{s}_{t-1}}{\partial \theta}(s_0^*, \theta), \mathrm{Id}^{\otimes 2} \right) \\
& + 2 \frac{\partial^3 \mathbf{T}_t}{\partial \theta^2 \partial s}(s_{t-1}, \theta) \cdot \left(\frac{\partial^2 \mathbf{s}_{t-1}}{\partial \theta^2}(s_0^*, \theta), \mathrm{Id} \right) \\
& + \frac{\partial^3 \mathbf{T}_t}{\partial \theta^3}(s_{t-1}, \theta).
\end{aligned}
$$

By Corollary 6.23 with $E_t = 0$ for all $t \geq 1$, the $\frac{\partial \mathbf{s}_t}{\partial \theta}(s_0^*, \cdot)$'s are bounded uniformly in t over B_Θ^*.

Now, we see the update equation on the $\frac{\partial^2 \mathbf{s}_t}{\partial \theta^2}(s_0^*, \cdot)$'s has the form:

$$
\frac{\partial^2 \mathbf{s}_t}{\partial \theta^2}(s_0^*, \theta) = \frac{\partial \mathbf{T}_t}{\partial s}(s_{t-1}, \theta) \cdot \frac{\partial^2 \mathbf{s}_{t-1}}{\partial \theta^2}(s_0^*, \theta) + B_t(\theta),
$$

where $B_t(\theta)$ is made up of terms bounded uniformly in t over B_Θ^*. Moreover, the $\frac{\partial \mathbf{T}_t}{\partial s}(s_{t-1}, \theta)$'s have spectral radius $1 - \alpha$. Then, thanks to Proposition 6.22, we obtain that the $\frac{\partial^2 \mathbf{s}_t}{\partial \theta^2}(s_0^*, \cdot)$'s are uniformly bounded in t on B_Θ^*.

Finally, the same reasoning applies to the $\frac{\partial^3 \mathbf{s}_t}{\partial \theta^3}(s_0^*, \cdot)$'s.

Bounding the derivatives of the $\mathcal{L}_{\rightsquigarrow t}$'s. For all $t \geq 1$ and $\theta \in \Theta$, we have

$$
\mathcal{L}_{\rightsquigarrow t}(s_0^*, \theta) = \mathcal{L}_t(\mathbf{s}_t(s_0^*, \theta)).
$$

Let us write as before, for all $t \geq 1$, $s_t = \mathbf{s}_t(s_0^*, \theta)$. Let $t \geq 1$, and $\theta \in B_\Theta^*$. Then

$$
\begin{aligned}
\frac{\partial \mathcal{L}_{\rightsquigarrow t}}{\partial \theta}(s_0^*, \theta) &= \frac{\partial \mathcal{L}_t}{\partial s}(s_t) \cdot \frac{\partial \mathbf{s}_t}{\partial \theta}(s_0^*, \theta) \\
\frac{\partial^2 \mathcal{L}_{\rightsquigarrow t}}{\partial \theta^2}(s_0^*, \theta) &= \frac{\partial \mathcal{L}_t}{\partial s}(s_t) \cdot \frac{\partial^2 \mathbf{s}_t}{\partial \theta^2}(s_0^*, \theta) + \frac{\partial^2 \mathcal{L}_t}{\partial s^2}(s_t) \cdot \left(\frac{\partial \mathbf{s}_t}{\partial \theta}(s_0^*, \theta) \right)^{\otimes 2} \\
\frac{\partial^3 \mathcal{L}_{\rightsquigarrow t}}{\partial \theta^3}(s_0^*, \theta) &= \frac{\partial \mathcal{L}_t}{\partial s}(s_t) \cdot \frac{\partial^3 \mathbf{s}_t}{\partial \theta^3}(s_0^*, \theta) \\
&\quad + 3 \frac{\partial^2 \mathcal{L}_t}{\partial s^2}(s_t, \theta) \cdot \left(\frac{\partial^2 \mathbf{s}_t}{\partial \theta^2}(s_0^*, \theta), \frac{\partial \mathbf{s}_t}{\partial \theta}(s_0^*, \theta) \right) \\
&\quad + \frac{\partial^3 \mathcal{L}_t}{\partial s^3}(s_t, \theta) \cdot \left(\frac{\partial \mathbf{s}_t}{\partial \theta}(s_0^*, \theta) \right)^{\otimes 3}.
\end{aligned}
$$

Thanks to the first part of the proof, all the derivatives of the $\mathbf{s}_t\left(s_0^*,\,\cdot\right)$'s are uniformly bounded in t on B_Θ^*. By assumption, all the derivatives up to third order of the $\mathcal{L}_t$'s are bounded on the stable tube. Therefore, all the derivatives up to third order of the $\mathcal{L}_{\rightsquigarrow t}\left(s_0^*,\,\cdot\right)$'s are uniformly bounded in t on B_Θ^*. Finally, this shows the family of functions $\frac{\partial^2 \mathcal{L}_{\rightsquigarrow t}}{\partial \theta^2}\left(s_0^*,\,\cdot\right)$ is equicontinuous on B_Θ^*. □

Lemma B.2 (Equicontinuity satisfied for the extended Hessians with all derivatives bounded)**.** *Assume that all first, second and third order derivatives of the $\mathbf{T}_t$'s and of the $\mathcal{L}_t$'s exist, and are bounded on a ball around the target trajectory (θ^*, s_t^*). (In particular, we can take $\gamma = 0$.)*

Assume moreover that the second derivatives of the extended Hessians $\mathcal{U}_t$ are controlled as follows on the balls of Assumption 2.14. Namely, we assume that there is a constant $\kappa_{\mathcal{U}} > 0$ such that for all $t \geq 1$, for all $v \in \mathrm{L}(\Theta, \mathbb{R})$, $s \in B_{\mathcal{S}_t}$, and $\theta \in B_\Theta$, we have

$$\left\| \frac{\partial^2 \mathcal{U}_t}{\partial v^2}\left(v,\, s,\, \theta\right) \right\|_{\mathrm{op}} < \kappa_{\mathcal{U}}, \qquad \left\| \frac{\partial^2 \mathcal{U}_t}{\partial v\, \partial\left(s,\, \theta\right)}\left(v,\, s,\, \theta\right) \right\|_{\mathrm{op}} < \kappa_{\mathcal{U}},$$

and

$$\left\| \frac{\partial^2 \mathcal{U}_t}{\partial\left(s,\, \theta\right)^2}\left(v,\, s,\, \theta\right) \right\|_{\mathrm{op}} \leq \kappa_{\mathcal{U}}\left(1 + \|v\|\right).$$

Then, Assumption 2.25.b is satisfied on B_Θ^*.

Notably, these assumptions on $\mathcal{U}_t$ cover the preconditioned case $\mathcal{U}_t(v, s, \theta) = P_t(s, \theta) \cdot v$ with smooth enough P_t.

Proof. Thanks to Lemma B.1 above, we know the $\frac{\partial^2 \mathcal{L}_{\rightsquigarrow t}}{\partial \theta^2}\left(s_0^*,\,\cdot\right)$'s are equicontinuous on B_Θ^*.

For $t \geq 0$ and $\theta \in \Theta$, let us write $g_t\left(\theta\right) = \left(\frac{\partial}{\partial \theta}\mathcal{L}_{\rightsquigarrow t}(s_0^*, \theta),\, \mathbf{s}_t(s_0^*, \theta),\, \theta\right)$. Then, for $\theta \in \Theta$, we have $\mathcal{H}_t\left(\theta\right) = \frac{\partial \mathcal{U}_t \circ g_t}{\partial \theta}\left(\theta\right)$, and (writing $g = (v,\, s,\, \theta)$)

$$\begin{aligned}
\mathcal{H}_t\left(\theta\right) - \mathcal{H}_t\left(\theta^*\right) &= \frac{\partial \mathcal{U}_t}{\partial g}\left(g_t\left(\theta\right)\right) \cdot \frac{\partial g_t}{\partial \theta}\left(\theta\right) - \frac{\partial \mathcal{U}_t}{\partial g}\left(g_t\left(\theta^*\right)\right) \cdot \frac{\partial g_t}{\partial \theta}\left(\theta^*\right) \\
&= \left(\frac{\partial \mathcal{U}_t}{\partial g}\left(g_t\left(\theta\right)\right) - \frac{\partial \mathcal{U}_t}{\partial g}\left(g_t\left(\theta^*\right)\right)\right) \cdot \frac{\partial g_t}{\partial \theta}\left(\theta\right) \\
&\quad - \frac{\partial \mathcal{U}_t}{\partial g}\left(g_t\left(\theta^*\right)\right) \cdot \left(\frac{\partial g_t}{\partial \theta}\left(\theta^*\right) - \frac{\partial g_t}{\partial \theta}\left(\theta\right)\right).
\end{aligned} \tag{22}$$

Moreover, for all $\theta \in \Theta$, we have

$$\frac{\partial g_t}{\partial \theta}\left(\theta\right) = \left(\frac{\partial^2}{\partial \theta^2}\mathcal{L}_{\rightsquigarrow t}(s_0^*, \theta),\, \frac{\partial \mathbf{s}_t}{\partial \theta}(s_0^*, \theta),\, \mathrm{Id}\right). \tag{23}$$

Let us first control the second term of (22). We have

$$\begin{aligned}\left\|\frac{\partial \mathcal{U}_t}{\partial g}\left(g_t\left(\theta^*\right)\right)\cdot\left(\frac{\partial g_t}{\partial \theta}\left(\theta^*\right)-\frac{\partial g_t}{\partial \theta}\left(\theta\right)\right)\right\| &\leq \left\|\frac{\partial \mathcal{U}_t}{\partial v}\left(g_t\left(\theta^*\right)\right)\cdot\left(\frac{\partial^2}{\partial \theta^2}\mathcal{L}_{\rightsquigarrow t}(s_0^*,\theta)-\frac{\partial^2}{\partial \theta^2}\mathcal{L}_{\rightsquigarrow t}(s_0^*,\theta^*)\right)\right\| \\ &+\left\|\frac{\partial \mathcal{U}_t}{\partial\left(s,\,\theta\right)}\left(g_t\left(\theta^*\right)\right)\cdot\left(\frac{\partial \mathbf{s}_t}{\partial \theta}(s_0^*,\theta)-\frac{\partial \mathbf{s}_t}{\partial \theta}(s_0^*,\theta^*),\,0\right)\right\| \\ &\leq \left\|\frac{\partial \mathcal{U}_t}{\partial v}\left(g_t\left(\theta^*\right)\right)\right\|_{\mathrm{op}}\left\|\frac{\partial^2}{\partial \theta^2}\mathcal{L}_{\rightsquigarrow t}(s_0^*,\theta)-\frac{\partial^2}{\partial \theta^2}\mathcal{L}_{\rightsquigarrow t}(s_0^*,\theta^*)\right\| \\ &+O\left(1+\left\|\frac{\partial \mathcal{L}_{\rightsquigarrow t}}{\partial \theta}\left(s_0^*,\,\theta^*\right)\right\|\right)\left\|\frac{\partial \mathbf{s}_t}{\partial \theta}(s_0^*,\theta)-\frac{\partial \mathbf{s}_t}{\partial \theta}(s_0^*,\theta^*)\right\|.\end{aligned}$$

by the control of $\partial \mathcal{U}_t/\partial(s,\theta)$ in Assumption 2.14. Now, the operator norm of $\frac{\partial \mathcal{U}_t}{\partial v}$ is bounded on the stable tube thanks to Assumption 2.14. Moreover, thanks to Lemma B.1, the third derivatives of $\mathcal{L}_{\rightsquigarrow t}$ and the second derivatives of $\mathbf{s}_t$ are bounded, so that the differences between θ and θ^* in this expression are $O\left(\theta-\theta^*\right)$. As a result, on the stable tube, the second term of Equation (22) is bounded by some $\rho_2\left(\|\theta-\theta^*\|\right)$.

Let us now control the first term of Equation (22). We have

$$\left\|\left(\frac{\partial \mathcal{U}_t}{\partial g}\left(g_t\left(\theta\right)\right)-\frac{\partial \mathcal{U}_t}{\partial g}\left(g_t\left(\theta^*\right)\right)\right)\cdot\frac{\partial g_t}{\partial \theta}\left(\theta\right)\right\| \leq \left\|\frac{\partial \mathcal{U}_t}{\partial g}\left(g_t\left(\theta\right)\right)-\frac{\partial \mathcal{U}_t}{\partial g}\left(g_t\left(\theta^*\right)\right)\right\|_{\mathrm{op}}\left\|\frac{\partial g_t}{\partial \theta}\left(\theta\right)\right\|.$$

Now, Equation (23) together with Lemma B.1 show that the gradients of the g_t's are bounded on B_Θ^*. Next, thanks to our assumptions on the second derivatives of $\mathcal{U}_t$, by decomposing g_t, we have (with suprema taken on the stable tube)

$$\begin{aligned}&\left\|\frac{\partial \mathcal{U}_t}{\partial g}\left(g_t\left(\theta\right)\right)-\frac{\partial \mathcal{U}_t}{\partial g}\left(g_t\left(\theta^*\right)\right)\right\|_{\mathrm{op}} \\ &\leq \left(\sup\left\|\frac{\partial^2 \mathcal{U}_t}{\partial v^2}\right\|_{\mathrm{op}}\right)\left\|g_t\left(\theta\right)-g_t\left(\theta^*\right)\right\|+\left(\sup_{\theta'\in B_\Theta^*}\left\|\frac{\partial^2 \mathcal{U}_t}{\partial\left(s,\,\theta\right)^2}\left(g_t\left(\theta'\right)\right)\right\|_{\mathrm{op}}\right)\left\|g_t\left(\theta\right)-g_t\left(\theta^*\right)\right\| \\ &\qquad+2\sup_{\theta'\in B_\Theta^*}\left\|\frac{\partial^2 \mathcal{U}_t}{\partial v\,\partial\left(s,\,\theta\right)}\left(g_t\left(\theta'\right)\right)\right\|_{\mathrm{op}}\left\|g_t\left(\theta\right)-g_t\left(\theta^*\right)\right\| \\ &\leq \left(\sup\left\|\frac{\partial^2 \mathcal{U}_t}{\partial v^2}\right\|_{\mathrm{op}}\right)\left\|g_t\left(\theta\right)-g_t\left(\theta^*\right)\right\|+O\left(1+\sup_{B_\Theta^*}\left\|\frac{\partial \mathcal{L}_{\rightsquigarrow t}\left(s_0^*,\,\cdot\right)}{\partial \theta}\right\|\right)\left\|g_t\left(\theta\right)-g_t\left(\theta^*\right)\right\| \\ &\qquad+2\sup_{\theta'\in B_\Theta^*}\left\|\frac{\partial^2 \mathcal{U}_t}{\partial v\,\partial\left(s,\,\theta\right)}\left(g_t\left(\theta'\right)\right)\right\|_{\mathrm{op}}\left\|g_t\left(\theta\right)-g_t\left(\theta^*\right)\right\|.\end{aligned}$$

Now, by assumption, the second derivative of $\mathcal{U}_t$ with respect to v, and its cross-derivative with respect to v and (s,θ), are both bounded on the stable tube. Moreover, by Lemma B.1, the first derivative of $\mathcal{L}_{\rightsquigarrow t}\left(s_0^*,\,\cdot\right)$ is bounded on B_Θ^*, while the g_t's are equicontinuous on the same ball. As a result, on the stable tube, the first term of Equation (22) is bounded by some $\rho_1\left(\|\theta-\theta^*\|\right)$.

Gathering the controls of the two terms of Equation (22) we have obtained, we see that the extended Hessians are indeed equicontinuous on B_Θ^*, so that Assumption 2.25.b is indeed satisfied on this ball. □

References

Leonard E Baum and Melvin Katz. Convergence rates in the law of large numbers. *Transactions of the American Mathematical Society*, 120(1):108–123, 1965. 42

Albert Benveniste, Michel Metivier, and Pierre Priouret. *Adaptive Algorithms and Stochastic Approximations*. Springer-Verlag Berlin Heidelberg, 1990. 6

Dimitri P Bertsekas and John N Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000. 6

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009. 6

Vivek S Borkar and Sean P Meyn. The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000. 6

Léon Bottou. Curiously fast convergence of some stochastic gradient descent algorithms. In *Proceedings of the symposium on learning and data science, Paris*, 2009. 7

Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. On the convergence of Adam and Adagrad. *arXiv preprint arXiv:2003.02395*, 2020. 6

Mert Gürbüzbalaban, Asu Ozdaglar, and Pablo Parrilo. Why random reshuffling beats stochastic gradient descent. *arXiv preprint arXiv:1510.08560*, 2015. 7

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *arXiv preprint arXiv:1609.05191*, 2016. 6

Herbert Jaeger. A tutorial on training recurrent neural networks, covering bppt, rtrl, ekf and the "echo state network" approach. Technical Report 159, German National Research Center for Information Technology, GMD, 2002. 1, 8, 15, 42

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL http://arxiv.org/abs/1412.6980. 7

Harold J. Kushner and George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer-Verlag New York, 2003. 6

Lennart Ljung. Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, 22:551–575, 1977. 6

Lennart Ljung and Torsten Söderström. *Theory and Practice of Recursive Identification*. MIT Press, 1984. 3, 6

James Martens. New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193*, 2014. 36

Asier Mujika, Florian Meier, and Angelika Steger. Approximating real-time recurrent learning with random kronecker factors. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 6594–6603. Curran Associates, Inc., 2018. URL

http://papers.nips.cc/paper/7894-approximating-real-time-recurrent-learning-with-ran
4, 15, 16, 44

Yann Ollivier. Riemannian metrics for neural networks i: feedforward networks. *Information and Inference: A journal of the IMA*, 4:108–153, 2015. 36

Yann Ollivier. Online natural gradient as a Kalman filter. *Electronic Journal of Statistics*, 12(2):2930–2961, 2018. 15, 36

Yann Ollivier, Corentin Tallec, and Guillaume Charpiat. Training recurrent networks online without backtracking. *arXiv preprint arXiv:1507.07680*, 2015. 1, 4, 15, 16, 44

B.A. Pearlmutter. Gradient calculations for dynamic recurrent neural networks: a survey. *IEEE Transactions on Neural Networks*, 6:1212–1228, September 1995. doi: 10.1109/72.410363. 1, 3, 6, 15, 42, 43

Boris Polyak and Anatoli Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30:838–855, 07 1992. doi: 10.1137/0330046. 6

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ryQu7f-RZ. 1, 5, 7, 28, 39

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. 5

Vladislav B Tadic. Almost sure convergence of two time-scale stochastic approximation algorithms. In *Proceedings of the 2004 American Control Conference*, volume 4, pages 3802–3807. IEEE, 2004. 7, 40

Corentin Tallec and Yann Ollivier. Unbiased Online Recurrent Optimization. In *International Conference on Learning Representations*, 2018. 1, 4, 15, 16, 42, 44, 46

Jacques Leopold Willems. *Stability Theory of Dynamical Systems*. Wiley Interscience Division, 1970. 23

Ronald J Williams and Jing Peng. An efficient gradient-based algorithm for online training of recurrent network trajectories. *Neural computation*, 2(4):490–501, 1990. 43

Ronald J Williams and David Zipser. Gradient-based learning algorithms for recurrent. *Backpropagation: Theory, architectures, and applications*, 433, 1995. 43

Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of Adam and Rmsprop. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11127–11135, 2019. 4, 6