LEARNING LOW DIMENSIONAL STATE SPACES WITH OVERPARAMETERIZED RECURRENT NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Overparameterization in deep learning typically refers to settings where a trained Neural Network (NN) has representational capacity to fit the training data in many ways, some of which generalize well, while others do not. In the case of Recurrent Neural Networks (RNNs), there exists an additional layer of overparameterization, in the sense that a model may exhibit many solutions that generalize well for sequence lengths seen in training, some of which *extrapolate* to longer sequences, while others do not. Numerous works studied the tendency of Gradient Descent (GD) to fit overparameterized NNs with solutions that generalize well. On the other hand, its tendency to fit overparameterized RNNs with solutions that extrapolate has been discovered only lately, and is far less understood. In this paper, we analyze the extrapolation properties of GD when applied to overparameterized linear RNNs. In contrast to recent arguments suggesting an implicit bias towards short-term memory, we provide theoretical evidence for learning low dimensional state spaces, which can also model long-term memory. Our result relies on a dynamical characterization which shows that GD (with small step size and near-zero initialization) strives to maintain a certain form of balancedness, as well as on tools developed in the context of the *moment problem* from statistics (recovery of a probability distribution from its moments). Experiments corroborate our theory, demonstrating extrapolation via learning low dimensional state spaces with both linear and non-linear RNNs.

1 INTRODUCTION

Neural Networks (NNs) are often *overparameterized*, in the sense that their representational capacity far exceeds what is necessary for fitting training data. Surprisingly, training overparameterized NNs via (variants of) Gradient Descent (GD) tends to produce solutions that generalize well, despite existence of many solutions that do not. This *implicit generalization* phenomenon attracted considerable scientific interest, resulting in various theoretical explanations (see, e.g., Woodworth et al. (2020); Yun et al. (2020); Zhang et al. (2017); Li et al. (2020); Ji & Telgarsky (2018); Lyu & Li (2019)).

Recent studies have surfaced a new form of implicit bias which arises in Recurrent Neural Networks (RNNs) and their variants (e.g., Long Short-Term Memory Hochreiter & Schmidhuber (1997) and Gated Recurrent Units Chung et al. (2014)). For such models, the length of sequences in training is often shorter than in testing, and it is not clear to what extent a learned solution will be able to *extrapolate* beyond the sequence lengths seen in training. In the overparameterized regime, where the representational capacity of the learned model exceeds what is necessary for fitting short sequences, there may exist solutions which generalize but do not extrapolate, meaning that their accuracy is high over short sequences but arbitrarily poor over long ones (see Cohen-Karlik et al. (2022)). In practice however, when training RNNs using GD, accurate extrapolation is often observed. We refer to this phenomenon as the *implicit extrapolation* of GD.

As opposed to the implicit generalization of GD, little is formally known about its implicit extrapolation. Existing theoretical analyses of the latter focus on linear RNNs — also known as *Linear Dynamical Systems* (LDS) — and either treat infinitely wide models (Emami et al., 2021), or models of finite width that learn from a memoryless teacher (Cohen-Karlik et al., 2022). In these regimes,

GD has been argued to exhibit an implicit bias towards short-term memory. While such results are informative, their generality remains in question, particularly since infinitely wide NNs are known to substantially differ from their finite-width counterparts, and since a memoryless teacher essentially neglects the main characteristic of RNNs (memory).

In this paper, we theoretically investigate the implicit extrapolation of GD when applied to overparameterized finite-width linear RNNs learning from a teacher with memory. We consider models with symmetric transition matrices, in the case where a student (learned model) with state space dimension d is trained on sequences of length k generated by a teacher with state space dimension $\hat{d}$. Our interest lies in the overparameterized regime, where d is greater than both k and $\hat{d}$, meaning that the student has state space dimension large enough to fully agree with the teacher on sequences of length k, while potentially disagreeing with it on longer sequences. As a necessary assumption on initialization, we follow prior work and focus on a certain balancedness condition, which is known (see experiments in Cohen-Karlik et al. (2022), as well as our theoretical analysis) to capture near-zero initialization as commonly employed in practice.

Our main theoretical result states that GD originating from a balanced initialization leads the student to extrapolate, *irrespective of how large its state space dimension is*. Key to the result is a surprising connection to a moment matching theorem from Cohen & Yeredor (2011), whose proof relies on ideas from compressed sensing (Elad, 2010; Eldar & Kutyniok, 2012) and neighborly polytopes (Gale, 1963). This connection may be of independent interest, and in particular may prove useful in deriving other results concerning implicit properties of GD. We corroborate our theory with experiments, which demonstrate extrapolation via learning low dimensional state spaces in both the analyzed setting and ones involving non-linear RNNs.

The implicit extrapolation of GD is a new and exciting area of inquiry. Our results suggest that short-term memory is not enough for explaining it as previously believed. We hope the techniques developed in this paper will contribute to a further understanding of this phenomenon.

2 Related Work

The study of linear RNNs, or LDS, has a rich history dating back to at least the early works of Kalman (Kalman, 1960; 1963). An extensively studied question relevant to extrapolation is that of *system identification*, which explores when the parameters of a teacher LDS can be recovered (see Ljung (1999)). Another related topic concerns finding compact realizations of systems, i.e. realizations of the same input-output mapping as a given LDS, with a state space dimension that is lower (see Antoulas (2005)). Despite the relation, our focus is fundamentally different from the above — we ask what happens when one learns an LDS using GD. Since GD is not explicitly designed to find a low dimensional state space, it is not clear that the application of GD to an overparameterized student allows system identification through a compact realization. The fact that it does relates to the implicit properties of GD, and to our knowledge has not been investigated in the classic LDS literature.

The implicit generalization of GD in training RNNs has been a subject of theoretical study for at least several years (see, e.g., Hardt et al. (2016); Allen-Zhu et al. (2019); Lim et al. (2021)). In contrast, works analyzing the implicit extrapolation of GD have surfaced only recently, specifically in Emami et al. (2021) and Cohen-Karlik et al. (2022).[1] Emami et al. (2021) analyzes linear RNNs in the infinite width regime, suggesting that in this case GD is implicitly biased towards impulse responses corresponding to short-term memory. Cohen-Karlik et al. (2022) studies finite-width linear RNNs (as we do), showing that when the teacher is memoryless (has state space dimension zero), GD emanating from a balanced initialization successfully extrapolates. Our work tackles an arguably more realistic and challenging setting — we analyze the regime in which the teacher has memory. Our results suggest that the implicit extrapolation of GD does not originate from a bias towards short-term memory, but rather a tendency to learn low dimensional state spaces.

On the empirical side, extrapolation of NNs to sequence lengths beyond those seen in training has been experimentally demonstrated in numerous recent works, covering both modern language

[1] We note that there have been works studying extrapolation in the context of non-recurrent NNs, e.g. Xu et al. (2020). This type of extrapolation deals with the behavior of learned functions outside the support of the training distribution, and is therefore fundamentally different from the type of extrapolation we consider (which deals with the behavior of learned functions over sequences longer than those seen in training).

and attention models (Press et al., 2022; Anil et al., 2022; Zhang et al., 2022), and RNNs with transition matrices of particular forms (Gu et al., 2022; 2021; 2020; Gupta, 2022). The current paper is motivated by these findings, and takes a step towards theoretically explaining them.

3 Linear Recurrent Neural Networks

Our theoretical analysis applies to single-input single-output (SISO) linear RNNs with symmetric transition matrices. Given a state space dimension $d \in \mathbb{N}$, this model is defined by the update rule:

$$\boldsymbol{s}_{t+1} = \boldsymbol{A}\boldsymbol{s}_t + \boldsymbol{B}x_t, \quad y_t = \boldsymbol{C}\boldsymbol{s}_t, \quad t = 0, 1, 2, \ldots, \tag{3.1}$$

where $\boldsymbol{A} \in \mathbb{R}^{d\times d}$, $\boldsymbol{B} \in \mathbb{R}^{d\times 1}$ and $\boldsymbol{C} \in \mathbb{R}^{1\times d}$ are configurable parameters, with the *transition matrix* $\boldsymbol{A}$ satisfying $\boldsymbol{A} = \boldsymbol{A}^\top$; $x_0, x_1, \ldots \in \mathbb{R}$ form an input sequence; $y_0, y_1, \ldots \in \mathbb{R}$ form the corresponding output sequence; and $\boldsymbol{s}_t \in \mathbb{R}^{d\times 1}$ represents the internal state at time t, assumed to be equal to zero at the outset (i.e. it is assumed that $\boldsymbol{s}_0 = \boldsymbol{0}$). As with any linear time-invariant system (Porat, 1996), the input-output mapping realized by the RNN is determined by its impulse response.

Definition 1. *The **impulse response** of the RNN is the output sequence corresponding to the input sequence $(x_0, x_1, x_2, \ldots) = (1, 0, 0, \ldots)$. Namely, it is the sequence $(\boldsymbol{CB}, \boldsymbol{CAB}, \boldsymbol{CA}^2\boldsymbol{B}, \ldots)$.*

For brevity, we employ the shorthand $\Theta := (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$. The $d \times d$ symmetric transition matrix $\boldsymbol{A}$ is parameterized through a $d(d+1)/2$-dimensional vector holding its upper triangular elements, and with a slight overloading of notation, the symbol $\boldsymbol{A}$ is also used to refer to this parameterization.

We note that our theory readily extends to multiple-input multiple-output (MIMO) networks, and the focus on the SISO case is merely for simplicity of presentation. Note also that the restriction to symmetric transition matrices is customary in both theory (Hazan et al., 2018) and practice (Gupta, 2022), and represents a generalization of the *canonical modal form*, which under mild non-degeneracy conditions does not limit generality (Boyd & Lessard, 2006).

Given a length k input sequence, $\mathbf{x} = (x_0, \ldots, x_{k-1}) \in \mathbb{R}^k$, consider the output at the last time step, i.e. $y := y_k \in \mathbb{R}$, and denote it by $RNN(\boldsymbol{x})$. Using this output as a label, we define an empirical loss induced by a training set $S = \left\{\left(\boldsymbol{x}^{(1)}, y^{(1)}\right), \ldots, \left(\boldsymbol{x}^{(N)}, y^{(N)}\right)\right\} \subset \mathbb{R}^k \times \mathbb{R}$:

$$\mathcal{L}_S(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = \frac{1}{N}\sum_{i=1}^{N} \ell\left(RNN\left(\boldsymbol{x}^{(i)}\right), y^{(i)}\right), \tag{3.2}$$

where $\ell(y, \hat{y}) = (y - \hat{y})^2$ is the square loss. By the update rule of the RNN (Equation 3.1), we have:

$$\mathcal{L}_S(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=0}^{k-1} \boldsymbol{CA}^{k-1-j}\boldsymbol{B}x_j^{(i)} - y^{(i)}\right)^2. \tag{3.3}$$

Suppose that ground truth labels are generated by an RNN as defined in Equation 3.1, and denote the state space dimension and parameters of this *teacher* network by $\hat{d}$ and $\hat{\Theta} = (\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{C}})$ respectively. We employ the common assumption (e.g., see Hardt et al. (2016)) by which input sequences are drawn from a whitened distribution, i.e. a distribution where $\mathbb{E}\left[x_j x_{j'}\right]$ equals 1 if $j = j'$ and 0 otherwise. The population loss over length k sequences can then be written as (see Lemma C.1):

$$\mathcal{L}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = \sum_{j=0}^{k-1}\left(\boldsymbol{CA}^j\boldsymbol{B} - \hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^j\hat{\boldsymbol{B}}\right)^2. \tag{3.4}$$

Equation 3.4 implies that a solution $\Theta = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ achieves zero population loss over length k sequences if and only if $\boldsymbol{CA}^j\boldsymbol{B} = \hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^j\hat{\boldsymbol{B}}$ for $j = 0, \ldots, k-1$. To what extent does such a solution imply that the *student* (i.e., the learned RNN) extrapolates to longer sequences? This depends on how close $\boldsymbol{CA}^j\boldsymbol{B}$ is to $\hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^j\hat{\boldsymbol{B}}$ for $j \geqslant k$.

Definition 2. *For $\epsilon \geqslant 0$ and $q \in \mathbb{N}$, we say that the student **ϵ-extrapolates with horizon** q with respect to (w.r.t) the teacher if:*

$$|\boldsymbol{CA}^j\boldsymbol{B} - \hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^j\hat{\boldsymbol{B}}| \leqslant \epsilon, \quad \forall j \in \{0, 1, \ldots, q-1\}. \tag{3.5}$$

*If the above holds for all $q \in \mathbb{N}$ then the student is said to **ϵ-extrapolate** w.r.t the teacher, and if it holds for all $q \in \mathbb{N}$ with $\epsilon = 0$ then the student is simply said to **extrapolate** w.r.t the teacher.*

Per Definition 2, ϵ-extrapolation with horizon q is equivalent to the first q elements of the student's impulse response being ϵ-close to those of the teacher's, whereas extrapolation means that the student's impulse response fully coincides with the teacher's. The latter condition implies that the student realizes the same input-output mapping as the teacher, for any sequence length (this corresponds to the notion of system identification; see Section 2). When the student is *overparameterized*, in the sense that d is greater than k and $\hat{d}$, it may perfectly generalize, i.e. lead the population loss over length k sequences (Equation 3.4) to equal zero, and yet fail to extrapolate as stated in the following proposition.

Proposition 3. *Assume $d > k$, and let $\epsilon \geqslant 0$ and $q \in \{k+1, k+2, \ldots\}$. Then, for any teacher parameters $\hat{\Theta}$, there exist student parameters Θ with which the population loss in Equation 3.4 equals zero, and yet the student does* not *ϵ-extrapolate with horizon q.*

Proof sketch (for complete proof see Appendix C.1.2). The result follows from the fact that the first d elements of the student's impulse response can be assigned freely via a proper choice of Θ. □

We are interested in the extent to which student parameters learned by GD extrapolate in the overparameterized regime. Proposition 3 implies that, regardless of how many (length k) sequences are used in training, if GD leads to any form of extrapolation, it must be a result of some implicit bias induced by the algorithm. Note that in our setting, extrapolation cannot be explained via classic tools from statistical learning theory, as evaluation over sequences longer than those seen in training violates the standard assumption of train and test data originating from the same distribution.

To decouple the question of extrapolation from that of generalization, we consider the case where the training set S is large, or more formally, where the empirical loss $\mathcal{L}_S(\cdot)$ (Equation 3.3) is well represented by the population loss $\mathcal{L}(\cdot)$ (Equation 3.4). We model GD with small step size via Gradient Flow (GF), as customary in the theory of NNs — see Saxe et al. (2013); Gunasekar et al. (2017); Arora et al. (2018; 2019b); Lyu & Li (2019); Li et al. (2020); Azulay et al. (2021) for examples and Elkabetz & Cohen (2021) for theoretical justification. That is, we analyze the following dynamics:

$$\dot{\alpha}(\tau) := \frac{d}{d\tau}\alpha(\tau) = -\frac{\partial}{\partial \alpha}\mathcal{L}\big(\boldsymbol{A}(\tau), \boldsymbol{B}(\tau), \boldsymbol{C}(\tau)\big) \ , \ \tau \geqslant 0 \, , \tag{3.6}$$

where $\alpha \in \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}\}$. If no assumption on initialization is made, no form of extrapolation can be established (indeed, the initial point may be a global minimizer of $\mathcal{L}(\cdot)$ that fails to extrapolate, and GF will stay there). Following prior work (see Cohen-Karlik et al. (2022)), we assume that initialization adheres to the following balancedness condition:

Definition 4. *An RNN with parameters $\Theta = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ is said to be* ***balanced*** *if $\boldsymbol{B} = \boldsymbol{C}^\top$.*

It was shown empirically in Cohen-Karlik et al. (2022) that the balancedness condition captures near-zero initialization as commonly employed in practice. We support this finding theoretically in Section 4.3. Aside from the initialization of the student, we will also assume that the teacher adheres to the balancedness condition.

4 Theoretical Analysis

We turn to our theoretical analysis. Section 4.1 proves that in the setting of Section 3, convergence of GF to a zero loss solution leads the student to extrapolate, *irrespective of how large its state space dimension is*. Section 4.2 extends this result by establishing that, under mild conditions, approximate convergence leads to approximate extrapolation. The results of sections 4.1 and 4.2 assume that GF emanates from a balanced initialization, which empirically is known to capture near-zero initialization as commonly employed in practice (see Section 3). Section 4.3 theoretically supports this empirical premise, by showing that with high probability, random near-zero initialization leads to balancedness.

We introduce notations that will be used throughout the analysis. For a matrix $\boldsymbol{Q} \in \mathbb{R}^{m \times n}$, we let $\|\boldsymbol{Q}\|_F$, $\|\boldsymbol{Q}\|_\infty$ and $\|\boldsymbol{Q}\|_2$ denote the Frobenius, ℓ_∞ and ℓ_2 (spectral) norms, respectively. For a vector $\boldsymbol{v} \in \mathbb{R}^m$, we use $\|\boldsymbol{v}\|$ to denote the Euclidean norm and v_i to denote its ith entry.

4.1 CONVERGENCE LEADS TO EXTRAPOLATION

Theoretical analyses of implicit generalization often assume convergence to a solution attaining zero loss (see, e.g., Azulay et al. (2021); Gunasekar et al. (2017); Lyu & Li (2019); Woodworth et al. (2020)). Under such an assumption, Theorem 5 below establishes implicit extrapolation, i.e. that the solution to which GF converges extrapolates, *irrespective of how large the student's state space dimension d is*. A condition posed by the theorem is that the training sequence length k is greater than two times the teacher's state space dimension $\hat{d}$. We empirically demonstrate the necessity of this condition in Section 5, by showing that smaller values for k lead to failure in extrapolation.

Theorem 5. *Assume that $d > k > 2\hat{d}$, the teacher parameters $\hat{\Theta}$ are balanced (Definition 4), and the student parameters Θ are learned by applying GF to the loss $\mathcal{L}(\cdot)$ (Equation 3.4) starting from a balanced initialization. Then, if GF converges to a point Θ^* satisfying $\mathcal{L}(\Theta^*) = 0$, this point extrapolates (Definition 2).*

In order to prove Theorem 5, we introduce two lemmas: Lemma 6, which shows that balancedness is preserved under GF; and Lemma 7, which (through a surprising connection to a moment problem from statistics) establishes that a balanced solution attaining zero loss necessarily extrapolates. With Lemmas 6 and 7 in place, the proof of Theorem 5 readily follows.

Lemma 6. *Let $\Theta(\tau)$, with $\tau \geqslant 0$, be a curve brought forth by applying GF to the loss $\mathcal{L}(\cdot)$ starting from a balanced initialization. Then, $\Theta(\tau)$ is balanced for every $\tau \geqslant 0$.*

Proof sketch (for complete proof see Appendix C.1.5). The result follows from the symmetric role of $\boldsymbol{B}$ and $\boldsymbol{C}$ in the loss $\mathcal{L}(\cdot)$. □

Lemma 7. *Suppose that $d > k > 2\hat{d}$, the teacher is balanced, and that the student parameters Θ are balanced and satisfy $\mathcal{L}(\Theta) = 0$. Then Θ extrapolates.*

Proof sketch (for complete proof see Appendix C.2). The proof is based on a surprising connection to the *moment problem* from statistics (recovery of a probability distribution from its moments), which has been studied for decades (see, e.g., Schmüdgen (2017)).

Without loss of generality, we may assume that $\hat{\boldsymbol{A}}$ is diagonal (if this is not the case then we apply an orthogonal eigendecomposition to $\hat{\boldsymbol{A}}$ and subsequently absorb eigenvectors into $\hat{\boldsymbol{B}}$ and $\hat{\boldsymbol{C}}$). We may also assume that $\hat{\boldsymbol{C}}\hat{\boldsymbol{B}} = 1$ (otherwise we absorb a scaling factor into $\boldsymbol{B}$ and/or $\boldsymbol{C}$). Since $\hat{\boldsymbol{C}}^\top = \hat{\boldsymbol{B}}$ (teacher is balanced), we may define a probability vector (i.e. a vector with non-negative entries summing up to one) $\hat{\boldsymbol{p}} \in \mathbb{R}^{\hat{d}}$ via $\hat{p}_i = \hat{C}_i \hat{B}_i$, $i = 1, \ldots, \hat{d}$. We let $\hat{Z}$ denote the random variable whose support is $\{\hat{A}_{11}, \ldots, \hat{A}_{\hat{d}\hat{d}}\}$ and which assumes the value $\hat{A}_{ii}$ with probability $\hat{p}_i$, $i = 1, \ldots, \hat{d}$. Notice that for every $j \in \mathbb{N}$:

$$\hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^j\hat{\boldsymbol{B}} = \sum\nolimits_{i=1}^{\hat{d}} \hat{p}_i \hat{A}_{ii}^j = \mathbb{E}[\hat{Z}^j]\,,$$

meaning that the elements of the teacher's impulse response are precisely the moments of $\hat{Z}$.

Similarly to above we may assume $\boldsymbol{A}$ is diagonal, and since $\mathcal{L}(\Theta) = 0$ it holds that $\boldsymbol{C}\boldsymbol{B} = \hat{\boldsymbol{C}}\hat{\boldsymbol{B}} = 1$. We may thus define a probability vector $\boldsymbol{p} \in \mathbb{R}^d$ via $p_i = C_i B_i$, $i = 1, \ldots, d$, and a random variable Z which assumes the value A_{ii} with probability p_i, $i = 1, \ldots, d$. For every $j \in \mathbb{N}$:

$$\boldsymbol{C}\boldsymbol{A}^j\boldsymbol{B} = \sum\nolimits_{i=1}^{d} p_i A_{ii}^j = \mathbb{E}[Z^j]\,,$$

and so the elements of the student's impulse response are precisely the moments of Z.

The probabilistic formulation we set forth admits an interpretation of extrapolation as a moment problem. Namely, since $\mathcal{L}(\Theta) = 0$ (i.e. $\boldsymbol{C}\boldsymbol{A}^j\boldsymbol{B} = \hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^j\hat{\boldsymbol{B}}$ for $j = 0, \ldots, k-1$) the random variables Z and $\hat{Z}$ agree on their first $k-1$ moments, and the question is whether they agree on all higher moments as well. We note that this question is somewhat more challenging than that tackled in classic instances of the moment problem, since the support of the random variable whose moments we match (Z) is not known to coincide with the support of the random variable we seek to recover ($\hat{Z}$). Luckily, a recent powerful result allow addressing the question we face — Cohen & Yeredor (2011) showed that the first $2n$ moments of a discrete random variable X taking at most $n \in \mathbb{N}$ values uniquely define X, in the sense that *any* discrete random variable agreeing with these $2n$ moments must be identical to X. Translating this result to our setting, we have that if Z agrees with $\hat{Z}$ on

its first $2\hat{d}$ moments, it must be identical to $\hat{Z}$, and in particular it must agree with $\hat{Z}$ on all higher moments as well. The fact that $k - 1 \geqslant 2\hat{d}$ then concludes the proof.

To attain some intuition for the result we imported from Cohen & Yeredor (2011), consider the simple case where $\hat{d} = 1$. The transition matrix $\hat{\boldsymbol{A}}$ is then a scalar $\hat{a} \in \mathbb{R}$, the random variable $\hat{Z}$ is deterministically equal to $\hat{a}$, and the teacher's impulse response is given by the moments $\mathbb{E}[\hat{Z}^j] = \hat{a}^j$, $j = 0, 1, \ldots$. Since by assumption $k > 2\hat{d}$, the fact that $\mathcal{L}(\Theta) = 0$ means the random variable corresponding to the student, Z, agrees with the first two moments of $\hat{Z}$. That is, Z satisfies $\mathbb{E}[Z] = \hat{a}$ and $\mathbb{E}[Z^2] = \hat{a}^2$. This implies that $\mathrm{Var}[Z] = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 = 0$, and therefore Z is deterministically equal to $\hat{a}$, i.e. it is identical to $\hat{Z}$. The two random variables thus agree on all of their moments, meaning the impulse responses of the student and teacher are the same. □

Proof of Theorem 5. By Lemma 6 (as well as continuity considerations) Θ^* is balanced. Lemma 7 then implies that it extrapolates. □

4.2 Approximate Convergence Leads to Approximate Extrapolation

Theorem 5 in Section 4.1 proves extrapolation in the case where GF converges to a zero loss solution. Theorem 8 below extends this result by establishing that, as long as the teacher is *stable*, approximate convergence leads to approximate extrapolation.

Theorem 8. *Assume the conditions of Theorem 5, and that the teacher parameters $\hat{\Theta}$ are stable, i.e. the singular values of $\hat{\boldsymbol{A}}$ are in $[-1, 1]$. Assume also that $\hat{\Theta}$ are non-degenerate, in the sense that the input-output mapping they realize is not identically zero. Then, for any $\epsilon > 0$ and $q \in \mathbb{N}$, there exists $\delta(\epsilon, q) > 0$ which admits the following: whenever GF leads the student parameters Θ to within (Frobenius) distance $\delta(\epsilon, q)$ from a balanced point Θ^* satisfying $\mathcal{L}(\Theta^*) = 0$, the student ϵ-extrapolates with horizon q.*

Proof sketch (for complete proof see Appendix C.3). In order to show ϵ-extrapolation with horizon q we need to show that $\forall \epsilon > 0, q \in \mathbb{N}, \exists \delta(\epsilon, q) > 0$ such that for any Θ' satisfying $||\Theta' - \Theta^*|| < \delta$, we have for $n = 0, \ldots, q$ that

$$\boldsymbol{C}'(\boldsymbol{A}')^n \boldsymbol{B}' - \hat{\boldsymbol{C}} \hat{\boldsymbol{A}}^n \hat{\boldsymbol{B}} = O(\epsilon). \tag{4.1}$$

First, by Theorem 5 Θ^* is an extrapolating solution. Following the interpretation of discrete random variables of Lemma 7 proof, we can bound the Wasserstein-1 distance between the extrapolating distribution defined by Θ^* and the approximating distribution defined by Θ'. From Proposition 2 in Wu & Yang (2020), we have that if $|\boldsymbol{C}'(\boldsymbol{A}')^n \boldsymbol{B}' - \boldsymbol{C}^*(\boldsymbol{A}^*)^n \boldsymbol{B}^*| \leqslant \delta$ for $n \in \{0, \ldots, 2\hat{d}\}$, then

$$\mathcal{W}_1(\Theta', \Theta^*) = O\left(\delta^{\frac{1}{4\hat{d}}}\right). \tag{4.2}$$

The second part of the proof ties the error at step p with the $\mathcal{W}_p$ distance (Biswas & Mackey, 2021):

$$\boldsymbol{C}'(\boldsymbol{A}')^p \boldsymbol{B}' - \boldsymbol{C}^*(\boldsymbol{A}^*)^p \boldsymbol{B}^* \leqslant \mathcal{W}_p. \tag{4.3}$$

The final step of the proof is based on Panaretos & Zemel (2019) and bounds $\mathcal{W}_p$ in terms of $\mathcal{W}_1$, i.e.,

$$\mathcal{W}_p(\Theta, \Theta^*) = S^{p-1} \mathcal{W}_1(\Theta, \Theta^*), \tag{4.4}$$

where S is the difference between the maximal and minimal eigenvalues in the teacher and student support (due to our assumption on convergence, S is bounded).

Combining the three steps above, we have that if $|\boldsymbol{C}'(\boldsymbol{A}')^n \boldsymbol{B}' - \boldsymbol{C}^*(\boldsymbol{A}^*)^n \boldsymbol{B}^*| \leqslant \delta$ for $n \in \{0, \ldots, 2\hat{d}\}$, then $\forall p > n$

$$\boldsymbol{C}'(\boldsymbol{A}')^p \boldsymbol{B}' - \boldsymbol{C}^*(\boldsymbol{A}^*)^p \boldsymbol{B}^* \leqslant S^{p-1} O\left(\delta^{\frac{1}{4\hat{d}}}\right). \tag{4.5}$$

By setting $\delta < \left(\frac{\epsilon}{c_1 \cdot S^{q-1}}\right)^{4\hat{d}}$ the RHS of Equation 4.5 satisfies $c_1 S^{p-1} \delta^{\frac{1}{4\hat{d}}} < \epsilon$. □

4.3 BALANCEDNESS CAPTURES NEAR-ZERO INITIALIZATION

Theorems 5 and 8 assume that GF emanates from a balanced initialization, i.e. from a point $\Theta = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ satisfying $\boldsymbol{B} = \boldsymbol{C}^\top$. It was shown in Cohen-Karlik et al. (2022) that theoretical predictions derived assuming balanced initialization faithfully match experiments conducted with near-zero initialization (an initialization commonly used in practice). Proposition 9 below theoretically supports this finding, establishing that with high probability, random near-zero initialization leads GF to arrive at an approximately balanced point, i.e. a point $\Theta = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ for which the difference between $\boldsymbol{B}$ and $\boldsymbol{C}^\top$ is negligible compared to their size.

Proposition 9. *Suppose that:* (i) $d > 4$*;* (ii) *the teacher parameters* $\hat{\Theta}$ *are balanced and are non-degenerate, in the sense that the input-output mapping they realize is not identically zero; and* (iii) *the student parameters are learned by applying GF to the loss* $\mathcal{L}(\cdot)$*. Let* $\tilde{\Theta}$ *be a random point in parameter space, with entries drawn independently from the standard normal distribution. For* $\epsilon > 0$*, consider the case where GF emanates from the initialization* $\epsilon\tilde{\Theta}$*, and denote the resulting curve by* $\Theta_\epsilon(\tau) = (\boldsymbol{A}_\epsilon(\tau), \boldsymbol{B}_\epsilon(\tau), \boldsymbol{C}_\epsilon(\tau))$*, with* $\tau \geqslant 0$*. Then, w.p. at least* 0.75*, for every* $\epsilon > 0$ *there exists* $\tau_\epsilon \geqslant 0$ *such that:*

$$\lim_{\epsilon \to 0^+} \frac{||\boldsymbol{B}_\epsilon(\tau_\epsilon) - \boldsymbol{C}_\epsilon^\top(\tau_\epsilon)||_F}{||\boldsymbol{B}_\epsilon(\tau_\epsilon) + \boldsymbol{C}_\epsilon^\top(\tau_\epsilon)||_F} = 0\,. \tag{4.6}$$

Proof sketch (for complete proof see Appendix C.4). The idea behind the proof is as follows. Assume ϵ is sufficiently small. Then, when the entries of $\Theta = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ are on the order of ϵ, we have $\frac{\partial}{\partial \boldsymbol{B}}\mathcal{L}(\Theta) \approx -2\hat{\boldsymbol{C}}\hat{\boldsymbol{B}} \cdot \boldsymbol{C}^\top$ and $\frac{\partial}{\partial \boldsymbol{C}}\mathcal{L}(\Theta) \approx -2\hat{\boldsymbol{C}}\hat{\boldsymbol{B}} \cdot \boldsymbol{B}^\top$. This implies that during the first part of the curve $\Theta_\epsilon(\tau)$ it holds that $\frac{d}{d\tau}(\boldsymbol{B}_\epsilon(\tau) - \boldsymbol{C}_\epsilon^\top(\tau)) \approx -2\hat{\boldsymbol{C}}\hat{\boldsymbol{B}} \cdot (\boldsymbol{B}_\epsilon(\tau) - \boldsymbol{C}_\epsilon^\top(\tau))$ and similarly $\frac{d}{d\tau}(\boldsymbol{B}_\epsilon(\tau) + \boldsymbol{C}_\epsilon^\top(\tau)) \approx \quad 2\hat{\boldsymbol{C}}\hat{\boldsymbol{B}} \cdot (\boldsymbol{B}_\epsilon(\tau) + \boldsymbol{C}_\epsilon^\top(\tau))$. Since $\hat{\boldsymbol{C}}\hat{\boldsymbol{B}} > 0$ (follows from the teacher parameters being balanced and non-degenerate), the entries of $\boldsymbol{B}_\epsilon(\tau) - \boldsymbol{C}_\epsilon^\top(\tau)$ shrink exponentially fast while those of $\boldsymbol{B}_\epsilon(\tau) + \boldsymbol{C}_\epsilon^\top(\tau)$ grow at the same rate. This exponential shrinkage/growth leads $\|\boldsymbol{B}_\epsilon(\tau) - \boldsymbol{C}_\epsilon^\top(\tau)\| / \|\boldsymbol{B}_\epsilon(\tau) + \boldsymbol{C}_\epsilon^\top(\tau)\|$ to become extremely small, more so the smaller ϵ is. □

5 EXPERIMENTS

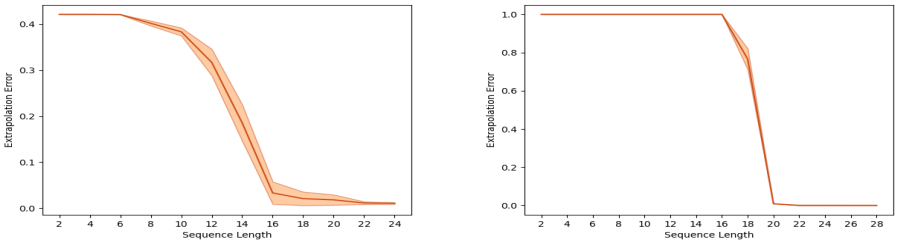
In this section we present experiments corroborating our theoretical analysis (Section 4). The latter establishes that, under certain conditions, a linear RNN with state space dimension d extrapolates when learning from a teacher network with state space dimension $\hat{d}$ via training sequences of length k, irrespective of how large d is compared to $\hat{d}$ and k. A key condition underlying the result is that k is larger than $2\hat{d}$. Section 5.1 below considers the theoretically analyzed setting, and empirically evaluates extrapolation as k varies. Its results demonstrate a phase transition, in the sense that extrapolation takes place when $k > 2\hat{d}$, in compliance with theory, but on the other hand fails when k falls below $2\hat{d}$, in which case the theory indeed does not guarantee extrapolation. Section 5.2 displays the same phenomenon with linear RNNs that do not adhere to some of the assumptions made by the theory (in particular the assumption of symmetric transition matrices, and those concerning balancedness). Finally, Section 5.3 considers non-linear RNNs (specifically, Gated Recurrent Unit networks Chung et al. (2014)), and shows that they too exhibit a phase transition in extrapolation as the training sequence length varies. For brevity, we defer some of the details of our implementation, as well as additional experiments, to Appendix A.

5.1 THEORETICALLY ANALYZED SETTING

Our first experiment considers the setting described in Section 3 and theoretically analyzed in Section 4. As representative values for the state space dimensions of the teacher and (overparameterized) student, we choose $\hat{d} = 5$ and $d = 40$ respectively (higher state space dimensions for the student, namely $d = 100$ and $d = 200$, yielded qualitatively identical results). For a given training sequence length k, the student is learned via GD applied directly to the population loss defined in Equation 3.4 (applying GD to the empirical loss defined in Equation 3.3, with $N = 10,000$ training examples, led to similar results). Figure 1(a) reports the extrapolation error (quantified by the ℓ_∞ distance between the impulse response of the learned student and that of the teacher) as a function of k. As can be seen, extrapolation exhibits a phase transition that accords with our theory: when $k > 2\hat{d}$ extrapolation error is low, whereas when k falls below $2\hat{d}$ extrapolation error is high.

5.2 Other Settings With Linear Recurrent Neural Networks

To assess the generality of our findings, we experiment with linear RNNs in settings that do not adhere to some of the assumptions made by our theory. Specifically, we evaluate settings in which: *(i)* the teacher is unbalanced, meaning $\hat{\boldsymbol{B}} \neq \hat{\boldsymbol{C}}^\top$, and its transition matrix $\hat{\boldsymbol{A}}$ is non-symmetric; *(ii)* the student's transition matrix $\boldsymbol{A}$ is not restricted to be symmetric; *(iii)* learning is implemented by optimizing the empirical loss defined in Equation 3.3 (rather than the population loss defined in Equation 3.4); and *(iv)* optimization is based on Adam Kingma & Ba (2014) (rather than GD), emanating from standard near-zero initialization which is generally unbalanced (namely, $\boldsymbol{B} \neq \boldsymbol{C}^\top$). Figure 1(b) reports the results of an experiment where the state space dimensions of the teacher and (overparameterized) student are $\hat{d} = 10$ and $d = 50$ respectively (higher state space dimensions for the student, namely $d = 100$ and $d = 200$, yielded qualitatively identical results), and where the teacher implements a delay line of $\hat{d}$ time steps (for details see Appendix B.2.2). Similar results obtained with randomly generated teachers are reported in Appendix A. As can be seen, despite the fact that our theory does not apply to the evaluated settings, its conclusions still hold — extrapolation error is low when the training sequence length k is greater than $2\hat{d}$, and high when k falls below $2\hat{d}$.



(a) Theoretically analyzed setting (b) Other setting with linear RNN

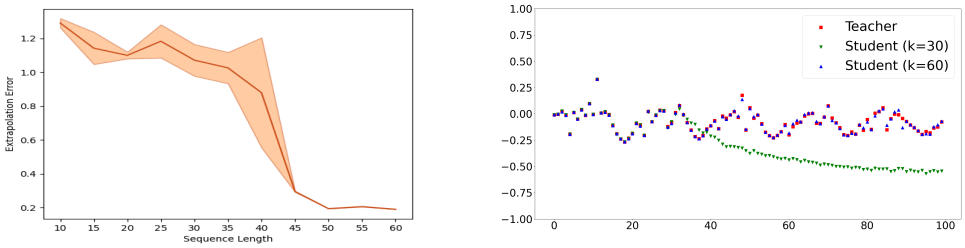
Figure 1: Demonstration of implicit extrapolation with linear RNNs. Plots show extrapolation error (average over three random seeds, with shaded region marking standard deviation) as a function of training sequence length k, for a student with state space dimension d learning from a teacher with state space dimension $\hat{d}$, where $d \gg \hat{d}$. (a) Models adhere to the setting described in Section 3 and theoretically analyzed in Section 4, with $\hat{d} = 5$, $d = 40$. (b) Models do not adhere to some of the assumptions made by the theory, and $\hat{d} = 10$, $d = 50$. Notice that extrapolation exhibits a phase transition that accords with theory — when $k > 2\hat{d}$ extrapolation error is low, and when k falls below $2\hat{d}$ extrapolation error is high. For further details see Sections 5.1 and 5.2 and Appendix A.

5.3 Non-Linear Recurrent Neural Networks

As a final experiment, we explore implicit extrapolation with non-linear RNNs, namely GRU networks. Specifically, we evaluate the extent to which a student GRU with state space dimension $d_g = 100$ extrapolates when learning from a teacher GRU with state space dimension $\hat{d}_g = 10$ (higher state space dimensions for the student, namely $d_g = 200$ and $d_g = 500$, yielded qualitatively identical results). The student is learned by optimizing an empirical loss comprising training sequences of length k_g, where k_g is predetermined. Optimization is based on Adam emanating from standard near-zero initialization. Figure 2(a) reports the extrapolation error (quantified by the ℓ_∞ distance between the response of the learned student and that of the teacher, averaged across randomly generated input sequences) for different choices of k_g. As can be seen, similarly to the case with linear RNNs (see Sections 5.1 and 5.2), there exists a critical threshold for the training sequence length k_g, above which extrapolation error is low and below which extrapolation error is high.[2] Figure 2(b) displays the average output response over different inputs of the teacher alongside those of two students — one trained with sequences of length $k_g = 30$, and the other with sequences of length $k_g = 60$.[3] As

[2]Note that this critical threshold is around four times the teacher's state space dimension, whereas with linear RNNs the critical threshold was around two times the teacher's state space dimension. Theoretically explaining this difference is viewed as an interesting direction for future work.

[3]Note that with GRU networks, in contrast to linear RNNs, the impulse response does not identify the input-output mapping realized by a network. It is presented in Figure 2(b) for demonstrative purposes.



(a) Extrapolation vs. training sequence length

(b) Average output over several inputs

Figure 2: Demonstration of implicit extrapolation with non-linear RNNs, namely GRU networks. Plots show results for a student with state space dimension $d_g = 100$ learning from a teacher with state space dimension $\hat{d}_g = 10$ using training sequences of length k_g, where k_g varies. (a) Extrapolation error (average over ten random seeds, with shaded region marking standard deviation) as a function of k_g. (b) Average output over several inputs of teacher and student for different choices of k_g. Notice that, similarly to the case with linear RNNs, there exists a critical threshold for k_g above which extrapolation error is low and below which extrapolation error is high. See more details in Appendix A.

expected, the impulse response of each student tracks that of the teacher for the first k_g time steps (where k_g is student-dependent). However, while the student for which $k_g = 30$ fails to track the teacher beyond k_g time steps, the student for which $k_g = 60$ succeeds, thereby exemplifying implicit extrapolation.

6 Conclusion

This paper studies the question of extrapolation in RNNs, and more specifically, of whether a student RNN trained on data generated by a teacher RNN can capture the behavior of the teacher over sequences longer than those seen in training. We focus on overparameterized students that can perfectly fit training sequences while producing a wide range of behaviors over longer sequences. Such a student will fail to extrapolate, unless the teacher possesses a certain structure, and the learning algorithm is biased towards solutions adhering to that structure. We show — theoretically for linear RNNs and empirically for both linear and non-linear RNNs — that such implicit extrapolation takes place when the teacher has a low dimensional state space and the learning algorithm is GD.

Existing studies of implicit extrapolation in (linear) RNNs (Emami et al., 2021; Cohen-Karlik et al., 2022) suggest that GD is biased towards solutions with short-term memory. While low dimensional state space and short-term memory may coincide in some cases, in general they do not, and a solution with low dimensional state space may entail long-term memory. Our theory and experiments show that in settings where low dimensional state space and short-term memory contradict each other, the implicit extrapolation chooses the former over the latter.

We note that linear RNNs fundamentally differ from the commonly studied model of linear (feed-forward) NNs (see, e.g., Arora et al. (2018); Ji & Telgarsky (2018); Arora et al. (2019a;b); Razin & Cohen (2020)). One of the key differences is that a linear RNN entails different powers of a parameter (transition) matrix, leading to a loss function which roughly corresponds to a sum of losses for multiple linear NNs having different architectures and shared weights. This makes the analysis of linear RNNs technically challenging, warranting introduction of new theoretical approaches such as the moment matching view employed in Section 4.1.

An important direction for future work is extending our theory to non-linear RNNs. We believe it is possible, in the same way that theories for linear (feed-forward) NNs were extended to account for non-linear NNs (see, e.g., Razin et al. (2021; 2022); Lyu & Li (2019)). An additional direction to explore is the applicability of our results to the recently introduced S4 model (Gu et al., 2022).

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. *Advances in neural information processing systems*, 32, 2019.

Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. Exploring length generalization in large language models. *arXiv preprint arXiv:2207.04901*, 2022.

Athanasios C Antoulas. *Approximation of large-scale dynamical systems*. SIAM, 2005.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning*, pp. 244–253. PMLR, 2018.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *International Conference on Learning Representations (ICLR)*, 2019a.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 7413–7424, 2019b.

Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In *International Conference on Machine Learning*, pp. 468–477. PMLR, 2021.

Niloy Biswas and Lester Mackey. Bounding wasserstein distance with couplings. *arXiv preprint arXiv:2112.03152*, 2021.

S Boyd and L Lessard. Ee263: Introduction to linear dynamical systems. *Online Lecture Notes, Stanford University, Spring Quarter*, 2007, 2006.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Anna Cohen and Arie Yeredor. On the use of sparsity for recovering discrete probability distributions from their moments. In *2011 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 753–756. IEEE, 2011.

Edo Cohen-Karlik, Avichai Ben David, Nadav Cohen, and Amir Globerson. On the implicit bias of gradient descent for temporal extrapolation. *International Conference on Artificial Intelligence and Statistics*, 2022.

Michael Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer Publishing Company, Incorporated, 2010.

Yonina C. Eldar and Gitta Kutyniok. *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2012.

Omer Elkabetz and Nadav Cohen. Continuous vs. discrete optimization of deep neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Melikasadat Emami, Mojtaba Sahraee-Ardakan, Parthe Pandit, Sundeep Rangan, and Alyson K. Fletcher. Implicit bias of linear rnns, 2021.

David Gale. Neighborly and cyclic polytopes. In *Proc. Sympos. Pure Math*, volume 7, pp. 225–232, 1963.

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. *Advances in Neural Information Processing Systems*, 33: 1474–1487, 2020.

Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in Neural Information Processing Systems*, 34, 2021.

Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in Neural Information Processing Systems*, 30, 2017.

Ankit Gupta. Diagonal state spaces are as effective as structured state spaces. *arXiv preprint arXiv:2203.14343*, 2022.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *arXiv preprint arXiv:1609.05191*, 2016.

Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Spectral filtering for general linear dynamical systems. *Advances in Neural Information Processing Systems*, 31, 2018.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2018.

Rudolf E Kalman. On the general theory of control systems. In *Proceedings First International Conference on Automatic Control, Moscow, USSR*, pp. 481–492, 1960.

Rudolf Emil Kalman. Mathematical description of linear dynamical systems. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 1(2):152–192, 1963.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pp. 1302–1338, 2000.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. *arXiv preprint arXiv:2012.09839*, 2020.

Soon Hoe Lim, N Benjamin Erichson, Liam Hodgkinson, and Michael W Mahoney. Noisy recurrent neural networks. *Advances in Neural Information Processing Systems*, 34:5124–5137, 2021.

Lennart Ljung. System identification. *Wiley encyclopedia of electrical and electronics engineering*, pp. 1–19, 1999.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2019.

Victor M Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6:405–431, 2019.

Boaz Porat. *A course in digital signal processing*. John Wiley & Sons, Inc., 1996.

Ofir Press, Noah Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022.

Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Noam Razin, Asaf Maman, and Nadav Cohen. Implicit regularization in tensor factorization. *International Conference on Machine Learning (ICML)*, 2021.

Noam Razin, Asaf Maman, and Nadav Cohen. Implicit regularization in hierarchical tensor factorization and deep convolutional neural networks. *International Conference on Machine Learning (ICML)*, 2022.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Konrad Schmüdgen. *The moment problem*, volume 9. Springer, 2017.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

Yihong Wu and Pengkun Yang. Optimal estimation of gaussian mixtures via denoised method of moments. *The Annals of Statistics*, 48(4):1981–2007, 2020.

Keyulu Xu, Mozhi Zhang, Jingling Li, Simon S Du, Ken-ichi Kawarabayashi, and Stefanie Jegelka. How neural networks extrapolate: From feedforward to graph neural networks. *arXiv preprint arXiv:2009.11848*, 2020.

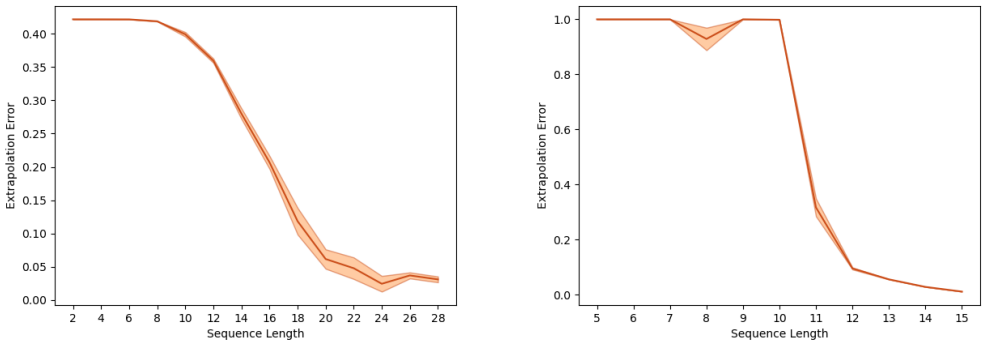
Chulhee Yun, Shankar Krishnan, and Hossein Mobahi. A unifying view on implicit bias in training linear neural networks. In *International Conference on Learning Representations*, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=Sy8gdB9xx`.

Yi Zhang, Arturs Backurs, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, and Tal Wagner. Unveiling transformers with lego: a synthetic reasoning task. *arXiv preprint arXiv:2206.04301*, 2022.

A FURTHER EXPERIMENTS

In this section we provide additional experiments that are not included in the main manuscript due to space constraints.



(a) Balanced teacher and general (unbalanced) student (b) Random (unbalanced) teacher and general (unbalanced) student

Figure 3: Extrapolation error as a function of the training sequence length k. (a) a balanced teacher with state dimensions $\hat{d} = 5$ and a general (unbalanced and non-diagonal) student with $d = 40$. (b) a random unbalanced teacher (see Section A.2) with dimension $\hat{d} = 5$, and a student that has a non-diagonal transition matrix and is trained with standard (small) initialization, with state dimension $d = 50$. In both plots results are averaged over 3 seeds.

A.1 BALANCED TEACHER

In Section 5.1, we have experimented with our proposed theoretical setup. In this section we provide additional figures and experiments.

A.1.1 UNBALANCED STUDENT

In this experiment we use the same balanced teacher with $\hat{d} = 5$ as done in Section 5.1. Instead of the diagonal student with balanced initialization, we use a general (non-diagonal) student with weights sampled from a Gaussian with scale 10^{-5} and $d = 40$. Results are depicted in Figure 3(a). A similar phase transition phenomenon to the one in Figure 1 is found also here.

A.1.2 EFFECT OF THE INITIALIZATION SCALE

Proposition 9 provides theoretical support for the fact that under near-zero initialization, the learned RNN tends to balancedness, which according to theorems 5 and 8 guarantees extrapolation. Below we empirically explore the impact of varying the initialization scale. We use the same setting as in Section A.1.1, and repeat the experiment with different initialization scales for the students' weights.

As can be seen in Figure 4, the extrapolation deteriorates for larger initialization scale, in the sense that it requires longer training sequences for getting good extrapolation error. This suggests that the condition of small initialization required by our theory is not an artifact of our proof technique, but rather a necessary condition for extrapolation to occur.

A.2 UNBALANCED TEACHER

In Section 5.2, we have tested the extrapolation with respect to a specific unbalanced teacher and have observed a similar phase transition as predicted by the theory of Section 4 and empirical evaluation of Section 5.1. Here we show that the phase transition is not limited to the specific teacher discussed by testing with respect to a randomly generated unbalanced (non-diagonal) teacher (see Section B.2.2).

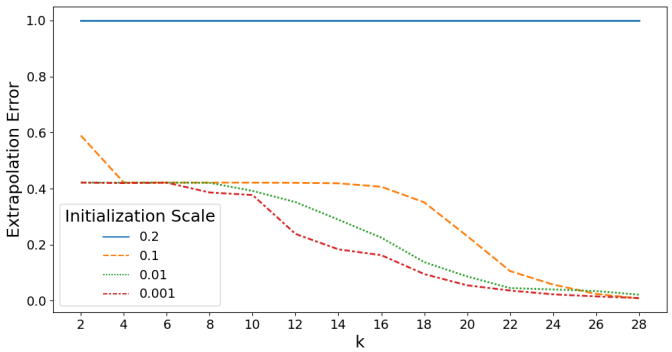


Figure 4: Extrapolation error as a function of training sequence length k for different initialization scales. Extrapolation error increases along with the scale of initialization.

The teacher is set to $\hat{d} = 5$ and student to $d = 50$. Results are presented in Figure 3 (b). Here too we can observe the phase transition phenomena.

A.3 Impulse Response Figures

In Section 5 we have presented the extrapolation performance in different settings. In order to better convey the meaning of extrapolating vs non-extrapolating solutions we present here figures of the impulse response of different models.

We start with the impulse response corresponding to the experiment described in Section 5.1. The figure depicts the balanced teacher with $\hat{d} = 5$ and two selected students (with $d = 40$), one trained with $k = 10$ and the other with $k = 20$.

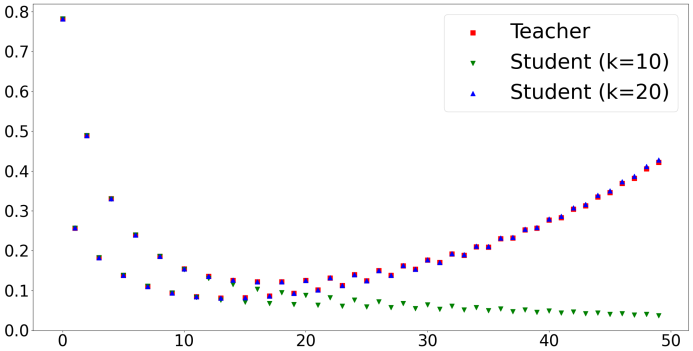


Figure 5: **Balanced teacher and student impulse response**. Students trained with: $k = 10, 20$ with respect to the balanced teacher described in Section B.2.1. As can be seen, both students track the teacher up to the k used in training, for $k = 10$ there is no extrapolation for larger values of k, whereas $k = 20$ tracks the teacher well beyond the sequence length used in training.

We can see that the student trained with $k = 10$ tracks the teacher several steps beyond the $10th$ time step and then decays to zero. For $k = 20$ we can see near perfect extrapolation for the horizon evaluated.

Next we turn to Section 5.2 and depict the average impulse responses of the "delay teacher" and the students trained with respect to the mentioned teacher.

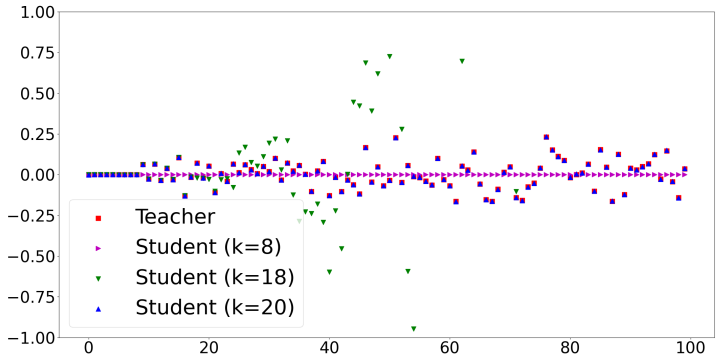


Figure 6: **Unbalanced teacher (delay) and student impulse response**. Students trained with: $k = 8, 18, 20$ with respect to the unbalanced delay teacher described in Section B.2.2. We can see that for $k = 18$ the student diverges for longer sequences while $k = 20$ which is trained for merely two additional time steps extrapolates and tracks the teacher almost perfectly.

Since the teacher here has $\hat{d} = 10$, a model trained with $k = 8$ is trained with respect to the zero impulse response (see Section B.2.2 for details on delay teacher), and as expected results with the 'zero' solution. we can see that for $k = 18$ the student diverges from the teacher shortly after the $18th$ time step. For $k = 20$ we can see near perfect extrapolation up to the horizon considered.

B Implementation Details

All the experiments are implemented using PyTorch.

B.1 Optimization

In Section 5.1 we optimize the population loss, which entails minimizing Equation 3.4 with respect to the parameters of the learned model. We use 15K optimization steps with Adam optimizer and a learning rate of 10^{-3}. In this experiment, the results were not sensitive to the initialization scale of the (balanced) student. In Section 5.2 and Section 5.3 in the experiments that involve minimizing the empirical loss, we use 50K optimization steps with early stopping (most experiments required less than 10K steps). The batch size is set to 100, data is sampled from a Gaussian with zero mean and scale of 1. Experiments were not sensitive to most hyper-parameters other than learning rate and initialization scale. The examination of the effect of initialization scale presented in Section A.1.2 is done with learning rate scheduler `torch.optim.lr_scheduler.MultiStepLR` using milestones at $[5000, 10000, 15000, 30000]$ and a decaying factor of $\gamma = 0.1$.

B.2 Teacher Generation

One of the main challenges in empirically evaluating extrapolation is that randomly sampling weights from a Gaussian distribution may result with an RNN of lower effective rank (i.e. the resulting RNN may be accurately approximated with another RNN with a smaller hidden dimension). We will now describe the teacher generation scheme for the different experiments.

B.2.1 Balanced Teacher Generation

A balanced teacher consists of d entries corresponding to the diagonal teacher and d entries representing $\hat{\boldsymbol{B}} = \hat{\boldsymbol{C}}^\top$. In order to avoid cases of rapid decay in the impulse response on the one hand, and exponential growth on the other, we set the eigenvalues to distribute uniformly between 0.6 and 1.05. The values of $\hat{\boldsymbol{B}}$ and $\hat{\boldsymbol{C}}$ are randomly sampled from a Gaussian around 0.5 and scale 1 and then normalized such that $\hat{\boldsymbol{C}}\hat{\boldsymbol{B}} = 1$.

B.2.2 UNBALANCED TEACHER GENERATION

In this experiment, the teacher has a general (non-symmetrid) matrix $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}} \neq \hat{\boldsymbol{C}}^\top$. We set the weights as described next.

Delay Teacher A 'delay' teacher has an impulse response of 1 at time step $i = \hat{d} - 1$, that is, the teacher has an impulse response of $(0, \dots, 0, 1, 0, \dots)$. In order to generate the mentioned impulse response we set the weights as follows,

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 & & 0 \\ & & \ddots & \\ 0 & 0 & & 1 \\ 0 & 0 & & 0 \end{pmatrix}, \ \boldsymbol{B} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}, \text{ and } \boldsymbol{C}^\top = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{B.1}$$

Note that $\boldsymbol{B}, \boldsymbol{C}$ above are set to extract the last entry of the first row of $\boldsymbol{A}^i$ and $\boldsymbol{A}$ is a Nilpotent shift matrix. It is straightforward to verify that $\boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B} = 1$ for $i = \hat{d} - 1$ and 0 otherwise.

Random Unbalanced Teacher The second unbalanced teacher is randomly generated. In order to avoid the caveats mentioned in Section A.1, we randomly sample the diagonal (from a Gaussian with zero mean and scale 0.1) and super diagonal (from a Gaussian with mean 0.7 and scale 0.1) of A. We set B, C as in equation B.1. The structure of $\boldsymbol{A}$ ensures similar properties to that of the delayed teacher, specifically, that the first entries of the impulse response is zero and the teacher is 'revealed' only after $\hat{d}$ time steps.

B.2.3 NON-LINEAR TEACHER GENERATION

As opposed to the linear teacher discussed in previous sections, when the teacher is a Gated Recurrent Units (GRU), it is unclear how to generate a non-trivial teacher. When randomly generating a teacher GRU the result is either a trivial model that quickly decays to zero or a teacher with an exploding impulse response (depending on the scale of the initialization). In order to produce a teacher with interesting extrapolation behaviour, we initialize a model with an initialization scale of 10^{-6} and train for 1000 step the model to mimic an arbitrarily chosen impulse response. The result of the mentioned procedure is a teacher GRU with non-trivial behaviour. Figure 2(b) shows that we get with this non-trivial teacher the phase transition phenomena as described in Section 5.3.

B.3 EXTRAPOLATION ERROR

The concept of extrapolation is very intuitive, and yet it does not admit any standard error measure. A proper extrapolation error measure should: (a) capture fine differences between two models with good extrapolation behaviour; and on the other hand, (b) be insensitive to the scale in which two non-extrapolating model explode. A natural approach which we take here is to report the ℓ_∞ norm difference on the tail of the impulse response. A model is considered non-extrapolating if the extrapolation error is worse than the extrapolation error of a trivial solution which has an impulse response of zeros.

C DEFERRED PROOFS

Here we provide complete proofs for the results in the paper.

C.1 AUXILARY PROOFS

In this section we provide missing proofs from the main paper and additional lemmas to be used in the main proofs.

C.1.1 POPULATION LOSS

Lemma C.1 (**Proof of Equation 3.4**). *Assume $\boldsymbol{x} \sim \mathcal{D}$ such that $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\boldsymbol{x}] = 0, \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\boldsymbol{x}\boldsymbol{x}^\top] = \boldsymbol{I}_k \in \mathbb{R}^{k,k}$, where $\boldsymbol{I}_k$ is the identity matrix. y is given by $y = \widehat{RNN}(\boldsymbol{x})$ where $\widehat{RNN}(\cdot)$ denotes the output of a teacher RNN, $\hat{\Theta} = (\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{C}})$. Denote $w_i = \hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^i\hat{\boldsymbol{B}}$, the loss for the student RNN satisfies:*

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\ell\left(RNN\left(\boldsymbol{x}\right), y\right)\right] = \sum_{i=0}^{k-1}\left(\boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B} - w_i\right)^2. \tag{C.1}$$

Proof of Lemma C.1. The population loss for training with sequences of length k is

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\ell\left(RNN\left(\boldsymbol{x}\right), y\right)\right] = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\left(\sum_{i=0}^{k-1}\boldsymbol{C}\boldsymbol{A}^{k-1-i}\boldsymbol{B}x_i - \sum_{j=0}^{k-1}w_{k-1-j}x_j\right)^2\right]. \tag{C.2}$$

Reversing the order of summation, expanding the terms,

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\ell\left(RNN\left(\boldsymbol{x}\right), y\right)\right] = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\left(\sum_{i=0}^{k-1}\boldsymbol{C}\boldsymbol{A}^{i}\boldsymbol{B}x_{k-1-i} - \sum_{j=0}^{k-1}w_jx_{k-1-j}\right)^2\right] \tag{C.3}$$

$$= \sum_{i,j=0}^{k-1}\left[\boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B}\boldsymbol{C}\boldsymbol{A}^j\boldsymbol{B} - 2\boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B}w_j + w_iw_j\right]\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[x_{k-1-i}x_{k-1-j}\right] \tag{C.4}$$

$$= \sum_{i,j=0}^{k-1}\left[\boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B}\boldsymbol{C}\boldsymbol{A}^j\boldsymbol{B} - 2\boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B}w_j + w_iw_j\right]\mathbf{1}_{\mathrm{k-1-i=k-1-j}} \tag{C.5}$$

$$= \sum_{i,j=0}^{k-1}\left[\boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B}\boldsymbol{C}\boldsymbol{A}^j\boldsymbol{B} - 2\boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B}w_j + w_iw_j\right]\mathbf{1}_{\mathrm{i=j}} \tag{C.6}$$

$$= \sum_{i=0}^{k-1}\left[(\boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B})^2 - 2\boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B}w_i + w_i^2\right] = \sum_{i=0}^{k-1}\left(\boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B} - w_i\right)^2. \tag{C.7}$$

where the transition from the second to third rows is by our assumption that $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[x_ix_j] = \mathbf{1}_{\mathrm{i=j}}$. Therefore we have,

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\ell\left(RNN\left(\boldsymbol{x}\right), y\right)\right] = \sum_{i=0}^{k-1}\left(\boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B} - w_i\right)^2. \tag{C.8}$$

concluding the proof. □

C.1.2 PERFECT GENERALIZATION AND FAILED EXTRAPOLATION

Proposition C.2 (Proposition 3 in main paper)**.** *Assume $d > k$, and let $\epsilon \geqslant 0$ and $q \in \{k+1, k+2, \ldots\}$. Then, for any teacher parameters $\hat{\Theta}$, there exist student parameters Θ with which the population loss in Equation 3.4 equals zero, and yet the student does* not *ϵ-extrapolate with horizon q.*

Proof. Consider a student, Θ, such that $\boldsymbol{A}$ is symmetric (and therefore has an orthogonal eigen-decomposition). Denote $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$. The impulse response at time step i can be expressed as $\boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B} = \boldsymbol{C}\boldsymbol{U}\boldsymbol{\Lambda}^i\boldsymbol{U}^\top\boldsymbol{B}$. The latter can be written compactly in matrix form as $\boldsymbol{V}\boldsymbol{g}$ where $\boldsymbol{V}$ is the

Vandermonde matrix with $diag(\boldsymbol{\Lambda})$ as its values,

$$\boldsymbol{V} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ \lambda_1 & \lambda_2 & \dots & \lambda_d \\ \lambda_1^2 & \lambda_2^2 & \dots & \lambda_d^2 \\ \vdots & \vdots & & \vdots \\ \lambda_1^{d-1} & \lambda_2^{d-1} & \dots & \lambda_d^{d-1} \end{pmatrix},$$

and $\boldsymbol{g}$ is defined as $\boldsymbol{g} \equiv (\boldsymbol{CU})^\top \circ \boldsymbol{U}^\top \boldsymbol{B}$.[4] A known result on square Vandermonde matrices is that they are invertible if and only if $\lambda_i \neq \lambda_j,\ \forall i \neq j$. Given a fixed set of distinct values $(\lambda_1, \dots, \lambda_d)$ and an arbitrary impulse response $\boldsymbol{r} \in \mathbb{R}^d$, in order for the student to generate the impulse response $\boldsymbol{r}$ (i.e. $\boldsymbol{Vg} = \boldsymbol{r}$), one can set the coefficient vector, $\boldsymbol{g} = \boldsymbol{V}^{-1}\boldsymbol{r}$ and end up with a symmetric student with $\boldsymbol{r}$ as its impulse response of length d.

Consider a teacher RNN, $\hat{\Theta} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$, we can set and the first k entries of $\boldsymbol{r}$ to $r_i = \hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^{i-1}\hat{\boldsymbol{B}},\ \forall i = \{1, \dots, k\}$. We are therefore left with $d - k$ degrees of freedom which yields many different students that correspond to the first k entries of the teacher while fitting arbitrary values beyond the k. □

C.1.3 Equivalence Between Balanced RNNs with Symmetric and Diagonal Transition Matrices

Lemma C.3. *A balanced RNN, $\Theta = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$, with a symmetric transition matrix (i.e. $\boldsymbol{B} = \boldsymbol{C}^\top$ and $\boldsymbol{A} = \boldsymbol{A}^\top$) has an equivalent (i.e. generating the same impulse response) RNN, $\Theta' = (\boldsymbol{A}', \boldsymbol{B}', \boldsymbol{C}')$, which is balanced and its transition matrix is diagonal.*

Lemma C.3 allows alternating between systems with symmetric and diagonal matrices. This is useful to simplify the analysis in Section 4.

Proof of Lemma C.3. Any symmetric matrix admits an orthogonal eigendecomposition with real (non-imaginary) eigenvalues. Denote $\boldsymbol{A} = \boldsymbol{U\Lambda U}^\top$. We can define

$$\boldsymbol{A}' = \boldsymbol{\Lambda}, \quad \boldsymbol{B}' = \boldsymbol{U}^\top \boldsymbol{B} \quad \text{and} \quad \boldsymbol{C}' = \boldsymbol{CU},$$

The ith index of the impulse response is given by

$$\boldsymbol{CA}^i\boldsymbol{B} = \boldsymbol{CU\Lambda}^i\boldsymbol{U}^\top\boldsymbol{B} = \boldsymbol{C}'\left(\boldsymbol{A}'\right)^i\boldsymbol{B}'$$

concluding that Θ and Θ' have the same impulse response of any length. □

C.1.4 Gradient Derivation

For completeness and Section C.4, we compute the gradients for the general setting.

Lemma C.4. *Given the population loss*

$$\mathcal{L}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = \sum_{j=0}^{k-1} \left(\boldsymbol{CA}^j\boldsymbol{B} - \hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^j\hat{\boldsymbol{B}}\right)^2. \tag{3.4 revisited}$$

Denote $\nabla\ell_i = \boldsymbol{CA}^i\boldsymbol{B} - w_i$, the derivatives of the loss with respect to $\boldsymbol{B}$, and $\boldsymbol{C}$ satisfy:

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{B}} = \sum_{i=0}^{k-1} \nabla\ell_i(\boldsymbol{A}^i)^\top\boldsymbol{C}^\top, \tag{C.9}$$

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{C}} = \sum_{i=0}^{k-1} \nabla\ell_i\boldsymbol{B}^\top\left(\boldsymbol{A}^i\right)^\top. \tag{C.10}$$

[4] Here $\circ$ denotes the Hadamard (elementwise) product.

Proof of Lemma C.4. Here, we will compute the gradient of the population loss.

Note that for $j \geqslant 0$, the derivative of $\boldsymbol{C}\boldsymbol{A}^j\boldsymbol{B}$ w.r.t to $\boldsymbol{B}$ is given by

$$\frac{\partial(\boldsymbol{C}\boldsymbol{A}^j\boldsymbol{B})}{\partial\boldsymbol{B}} = (\boldsymbol{A}^j)^\top\boldsymbol{C}^\top. \tag{C.11}$$

Similarly, the derivative of $\boldsymbol{C}\boldsymbol{A}^j\boldsymbol{B}$ w.r.t to $\boldsymbol{C}$ is given by

$$\frac{\partial(\boldsymbol{C}\boldsymbol{A}^j\boldsymbol{B})}{\partial\boldsymbol{C}} = \boldsymbol{B}^\top(\boldsymbol{A}^j)^\top. \tag{C.12}$$

Using these derivatives, we can calculate the derivative of the population loss, (assigning $w_i = \hat{\boldsymbol{B}}\hat{\boldsymbol{A}}^i\hat{\boldsymbol{C}}$),

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\ell\left(RNN\left(\boldsymbol{x}\right), y\right)\right] = \sum_{i=0}^{k-1}\left(\boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B} - w_i\right)^2. \tag{C.13}$$

Denoting $\nabla\ell_i = \boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B} - w_i$, and noting that w_i is constant (depends on $\hat{\Theta}$), we have for $\boldsymbol{X} \in \{\boldsymbol{B}, \boldsymbol{C}\}$:

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{X}} = \sum_{i=0}^{k-1}\frac{\partial\left(\boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B} - w_i\right)^2}{\partial\boldsymbol{X}} = \sum_{i=0}^{k-1}\nabla\ell_i\frac{\partial\left(\boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B} - w_i\right)}{\partial\boldsymbol{X}} = \sum_{i=0}^{k-1}\nabla\ell_i\frac{\partial\left(\boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B}\right)}{\partial\boldsymbol{X}}. \tag{C.14}$$

Plugging in Equation C.11 and Equation C.12, we have:

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{B}} = \sum_{i=0}^{k-1}\nabla\ell_i\frac{\partial\left(\boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B}\right)}{\partial\boldsymbol{B}} = \sum_{i=0}^{k-1}\nabla\ell_i(\boldsymbol{A}^i)^\top\boldsymbol{C}^\top, \tag{C.15}$$

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{C}} = \sum_{i=0}^{k-1}\nabla\ell_i\frac{\partial\left(\boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B}\right)}{\partial\boldsymbol{C}} = \sum_{i=0}^{k-1}\nabla\ell_i\boldsymbol{B}^\top(\boldsymbol{A}^i)^\top, \tag{C.16}$$

□

C.1.5 Lemma 6 (Conservation of Balancedness)

Lemma C.5. ***[Lemma 6 in main paper]*** *When optimizing equation 3.4 with GF emenating from a balanced initialization $\Theta(0)$, the parameters $\Theta(\tau)$ are balanced for all $\tau \in \mathbb{R}_+$.*

We prove the above result by first showing it for gradient descent and then translating the result to gradient flow. The gradient descent result is stated below, and generalizes a result that was shown in Cohen-Karlik et al. (2022) for the memoryless case.

Lemma C.6. *When optimizing equation 3.4 with GD with balanced initial conditions, then $\forall t \in \mathbb{N}$, Θ has a balanced weight configuration, i.e. $\boldsymbol{B}_t = \boldsymbol{C}_t^\top$.*

Proof of Lemma C.6. We prove by induction. By our assumption, the condition holds for $t = 0$. Assume $\boldsymbol{B}_t = \boldsymbol{C}_t^\top$, our goal is to show the conditions hold for $(\boldsymbol{B}_{t+1}, \boldsymbol{C}_{t+1})$. In order to show that $\boldsymbol{B}_{t+1} = \boldsymbol{C}_{t+1}^\top$, we only need to show that $\frac{\partial\mathcal{L}}{\partial\boldsymbol{B}_t} = \left(\frac{\partial\mathcal{L}}{\partial\boldsymbol{C}_t}\right)^\top$. Writing the gradients (Lemma C.4), we have

$$\left(\frac{\partial\mathcal{L}}{\partial\boldsymbol{C}_t}\right)^\top = \sum_{i=0}^{k-1}\nabla\ell_i\boldsymbol{A}_t^i\boldsymbol{B}_t = \sum_{i=0}^{k-1}\nabla\ell_i(\boldsymbol{A}_t^\top)^i\boldsymbol{C}_t^\top = \frac{\partial\mathcal{L}}{\partial\boldsymbol{B}_t}, \tag{C.17}$$

where the inequality follows from the induction assumption and the symmetric structure of $\boldsymbol{A}_t$. To conclude, the gradients at time t are the same and $\boldsymbol{B}_t = \boldsymbol{C}_t^\top$ by the induction assumption, arriving at

$$\boldsymbol{B}_{t+1} = \boldsymbol{B}_t - \eta\frac{\partial\mathcal{L}}{\partial\boldsymbol{B}_t} = \boldsymbol{C}_t^\top - \eta\left(\frac{\partial\mathcal{L}}{\partial\boldsymbol{C}_t}\right)^\top = \boldsymbol{C}_{t+1}^\top \tag{C.18}$$

□

The proof of Lemma 6 follows from Lemma C.6 and the fact that for sufficiently small step size GD approximates GF with arbitrary precision (see Theorem 3 in Elkabetz & Cohen, 2021).

C.1.6 CONSERVATION OF DIFFERENCE OF NORMS

Appendix C.1.5 shows that if weights are initialized to be balanced, this property is conserved throughout optimization. Here we show under standard initialization schemes, the difference between the norms of $\boldsymbol{B}$ and $\boldsymbol{C}$ is also conserved.

Lemma C.7. *The difference between the norms of $\boldsymbol{B}$, $\boldsymbol{C}$ is conserved throughout GF, i.e.,*

$$\frac{d}{dt}\left(\|\boldsymbol{B}\|^2 - \|\boldsymbol{C}\|^2\right) = 0. \tag{C.19}$$

Proof of Lemma C.7. We wish to prove that the difference between the norms is conserved over time. Consider the following expression:

$$\alpha \equiv \|\boldsymbol{B}\|^2 - \|\boldsymbol{C}\|^2 = \boldsymbol{B}^\top \boldsymbol{B} - \boldsymbol{C}\boldsymbol{C}^\top. \tag{C.20}$$

With this notation, we just need to prove that $\dot{\alpha} = 0$. The derivative of $\boldsymbol{B}, \boldsymbol{C}$ with respect to time is given by,

$$\dot{\boldsymbol{B}} = -\sum_{i=0}^{k-1} \nabla \ell_i (\boldsymbol{A}^\top)^i \boldsymbol{C}^\top, \tag{C.21}$$

$$\dot{\boldsymbol{C}} = -\sum_{i=0}^{k-1} \nabla \ell_i \boldsymbol{B}^\top (\boldsymbol{A}^\top)^i. \tag{C.22}$$

Using the interchangeability of derivative and transpose, we have:

$$\dot{\alpha} = \dot{\boldsymbol{B}}^\top \boldsymbol{B} + \boldsymbol{B}^\top \dot{\boldsymbol{B}} - \dot{\boldsymbol{C}}\boldsymbol{C}^\top - \boldsymbol{C}\dot{\boldsymbol{C}}^\top = 2\boldsymbol{B}^\top \dot{\boldsymbol{B}} - 2\dot{\boldsymbol{C}}\boldsymbol{C}^\top. \tag{C.23}$$

Plugging equation C.21 and equation C.22, we get

$$\dot{\alpha} = 2\boldsymbol{B}^\top \left(-\sum_{i=0}^{k-1} \nabla \ell_i (\boldsymbol{A}^\top)^i \boldsymbol{C}^\top\right) - 2\left(-\sum_{i=0}^{k-1} \nabla \ell_i \boldsymbol{B}^\top (\boldsymbol{A}^\top)^i\right)\boldsymbol{C}^\top \tag{C.24}$$

$$= -2\left[\boldsymbol{B}^\top \left(\sum_{i=0}^{k-1} \nabla \ell_i (\boldsymbol{A}^\top)^i\right)\boldsymbol{C}^\top - \boldsymbol{B}^\top \left(\sum_{i=0}^{k-1} \nabla \ell_i (\boldsymbol{A}^\top)^i\right)\boldsymbol{C}^\top\right] = 0. \tag{C.25}$$

establishing that $\frac{d}{dt}\left(\|\boldsymbol{B}\|^2 - \|\boldsymbol{C}\|^2\right) = 0$.

□

C.2 LEMMA 7 (EXACT EXTRAPOLATION)

Lemma C.8. ***[Lemma 7 in main paper]*** *Suppose that $d > k > 2\hat{d}$, the teacher is balanced, and that the student parameters Θ are balanced and satisfy $\mathcal{L}(\Theta) = 0$. Then Θ extrapolates.*

Proof of Lemma C.8. By Lemma C.3, a balanced RNN with symmetric transition matrix has an equivalent (generating the same impulse response) balanced RNN with a diagonal transition matrix. We will continue under the assumption of diagonal transition matrices.

Without loss of generality we assume $\hat{\boldsymbol{C}}\hat{\boldsymbol{B}} = 1$. Otherwise, the problem can be rescaled by $\hat{\boldsymbol{C}}\hat{\boldsymbol{B}}$, which is equivalent to rescaling the initial conditions, and providing no additional information.[5]

From the balanced assumption, we have $\hat{\boldsymbol{C}}^\top = \hat{\boldsymbol{B}}$. Denote $\hat{\boldsymbol{p}} = \hat{\boldsymbol{C}}^\top \odot \hat{\boldsymbol{B}} = \hat{\boldsymbol{B}} \odot \hat{\boldsymbol{B}}$, and we get $\hat{p}_i \geqslant 0$ and $\sum_i \hat{p}_i = 1$, and therefore $\hat{\boldsymbol{p}}$ may be interpreted as a distribution over a random variable with $\hat{d}$ possible values. We shall assume that these values are $\hat{A}_1, \ldots, \hat{A}_{\hat{d}}$, and denote the corresponding random variable by Z.

Furthermore, we can also interpret elements of the impulse response of $\hat{\Theta}$ as moments of this distribution. Let us write the nth element of the impulse response as:

$$\hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^n\hat{\boldsymbol{B}} = \sum_i \hat{p}_i \hat{a}_i^n = \mathbb{E}_{\hat{\boldsymbol{p}}}[\hat{Z}^n], \tag{C.26}$$

[5]The case for which $\hat{\boldsymbol{C}}\hat{\boldsymbol{B}} = 0$ is handled separately.

where $\mathbb{E}_{\boldsymbol{p}}[Z]$ is the expected value of a random variable Z under the distribution $\boldsymbol{p}$. In the same way, we can define for the learned model Θ, a distribution $p_i = C_i B_i$, and write the learned impulse response as:

$$\boldsymbol{C}\boldsymbol{A}^n\boldsymbol{B} = \sum_i p_i a_i^n = \mathbb{E}_{\boldsymbol{p}}[Z^n]. \tag{C.27}$$

This view provides us with a moment matching interpretation of the learning problem. Namely, the fact that Θ matches the first k elements of the teacher impulse response, is the same as saying they agree on the first $k-1$ moments $\mathbb{E}_{\boldsymbol{p}}[Z^j]$ for $j \in \{1, \dots, k-1\}$.[6] The question of extrapolation is whether equality in the first $k-1$ moments implies an equality in all other moments.

In (Cohen & Yeredor, 2011, Theorem 1) and in (Wu & Yang, 2020, Lemma 4) it is shown that the first $2\hat{d}$ moments of a discrete random variable taking at most $\hat{d}$ different values uniquely define this random variable. Therefore, any other discrete random variable identifying with the teacher on $2\hat{d}$ moments must be the same random variable and therefore identifies on higher moments as well. Since we assumed $k > 2\hat{d}$, this result immediately implies that equality in the first $k-1$ moments implies equality in all other moments.

For the case $\hat{\boldsymbol{C}}\hat{\boldsymbol{B}} = 0$, from our assumption that the teacher is balanced, we have that the condition is met only if $\hat{C}_i = \hat{B}_i = 0$ for $i = 1, \dots, \hat{d}$. Such a teacher has an impulse response of zeros, for $k \geqslant 1$, a student minimizing the loss must also satisfy $\boldsymbol{C}\boldsymbol{B} = 0$ and therefore has the zeros as its impulse response thus extrapolating with respect to the said teacher.

□

C.3 Theorem 8 (Approximate Extrapolation)

This section is devoted to the proof of Theorem 8 which ties the approximation error of optimization to that of extrapolation. Lemma C.9 below shows that if the parameters are within close proximity to an extrapolating solution, the generalization error (i.e. loss values) are small.

Lemma C.9. *Assume $\|\boldsymbol{X} - \boldsymbol{X}^*\|_\infty < \delta$ for $\boldsymbol{X} \in \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}\}$, and $\Theta^* = (\boldsymbol{A}^*, \boldsymbol{B}^*, \boldsymbol{C}^*)$ is an extrapolating solution (with $\mathcal{L}(\Theta^*) = 0$). Then, there exists a constant depending on k and Θ^*, $c(k, \Theta^*)$, such that $\mathcal{L}(\Theta) \leqslant c\delta^2$.*

Proof of Lemma C.9. First we note that Θ^* attains zero loss and is an extrapolating with respect to the teacher $\hat{\Theta}$, specifically,

$$\boldsymbol{C}^*(\boldsymbol{A}^*)^i\boldsymbol{B} = \hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^i\hat{\boldsymbol{B}}, \ \forall i = 0, 1, \dots, \tag{C.28}$$

From the above, the loss can be written with respect to Θ^* instead of $\hat{\Theta}$, i.e.

$$\mathcal{L}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = \sum_{j=0}^{k-1} \left(\boldsymbol{C}\boldsymbol{A}^j\boldsymbol{B} - \boldsymbol{C}^*(\boldsymbol{A}^*)^j\boldsymbol{B}^*\right)^2 . \tag{C.29}$$

Denote by $\boldsymbol{l} \in \mathbb{R}^k$ the residuals vector with respect to Θ^*, i.e. $l_i \equiv \boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B} - \boldsymbol{B} - \boldsymbol{C}^*(\boldsymbol{A}^*)^i\boldsymbol{B}^*$. With these notations, the loss amounts to $\mathcal{L}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = \|\boldsymbol{l}\|^2$.

From equivalence of norms, we have that for $\boldsymbol{v} \in \mathbb{R}^n$, $\|\boldsymbol{v}\|_\infty \leqslant \|\boldsymbol{v}\|_2 \leqslant \sqrt{n}\|\boldsymbol{v}\|_\infty$. The latter means that bounding $\|\boldsymbol{v}\|_\infty$ will provide the desired bound on the loss with a factor of $\sqrt{k}$.

Recall that Θ^* is fixed, specifically, $\|\boldsymbol{X}^*\|$ is bounded by some constant $m_{\boldsymbol{X}^*}$ for $\boldsymbol{X} \in \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}\}$. Denote $m \equiv \max\{m_{\boldsymbol{A}^*}, m_{\boldsymbol{B}^*}, m_{\boldsymbol{C}^*}, 1\}$, the ith entry of $\boldsymbol{l}$ can be bounded by

$$\begin{aligned}
\left|\boldsymbol{C}\boldsymbol{A}^{i-1}\boldsymbol{B} - \boldsymbol{C}^*(\boldsymbol{A}^*)^{i-1}\boldsymbol{B}^*\right| &= \left|(\boldsymbol{C} - \boldsymbol{C}^* + \boldsymbol{C}^*)\boldsymbol{A}^{i-1}\boldsymbol{B} - \boldsymbol{C}^*(\boldsymbol{A}^*)^{i-1}\boldsymbol{B}^*\right| && \text{(C.30)}\\
&\leqslant \left|\boldsymbol{C}^*\left(\boldsymbol{A}^{i-1}\boldsymbol{B} - (\boldsymbol{A}^*)^{i-1}\boldsymbol{B}^*\right)\right| + \left|(\boldsymbol{C} - \boldsymbol{C}^*)\boldsymbol{A}^{i-1}\boldsymbol{B}\right| && \text{(C.31)}\\
&\leqslant m\left\|\boldsymbol{A}^{i-1}\boldsymbol{B} - (\boldsymbol{A}^*)^{i-1}\boldsymbol{B}^*\right\| + \delta\left\|\boldsymbol{A}^{i-1}\boldsymbol{B}\right\| && \text{(C.32)}
\end{aligned}$$

[6] Equality of the 0th moment ensures the student induces a valid probability, i.e. $\sum_i C_i B_i = \sum_i \hat{C}_i \hat{B}_i = 1$.

From Cauchy-Schwarz inequality, the RHS satisfies $\delta\|\boldsymbol{A}^{i-1}\boldsymbol{B}\| \leqslant \delta\|\boldsymbol{A}\|^{i-1}\|\boldsymbol{B}\| \leqslant \delta(m+\delta)^i$. The last transition uses,

$$\|\boldsymbol{X}\| = \|\boldsymbol{X} - \boldsymbol{X} + \boldsymbol{X}^*\| \leqslant \|\boldsymbol{X} - \boldsymbol{X}^*\| + \|\boldsymbol{X}^*\| \leqslant \delta + m \tag{C.33}$$

for $\boldsymbol{X} \in \{\boldsymbol{A}, \boldsymbol{B}\}$.

In a similar manner, the LHS can be bounded by

$$\left\|\boldsymbol{A}^{i-1}\boldsymbol{B} - (\boldsymbol{A}^*)^{i-1}\boldsymbol{B}^*\right\| = \left\|\boldsymbol{A}^{i-1}\boldsymbol{B} - \boldsymbol{A}^{i-1}\boldsymbol{B}^* + \boldsymbol{A}^{i-1}\boldsymbol{B}^* - (\boldsymbol{A}^*)^{i-1}\boldsymbol{B}^*\right\| \tag{C.34}$$

$$\leqslant \underbrace{\|\boldsymbol{A}\|^{i-1}\|\boldsymbol{B} - \boldsymbol{B}^*\|}_{\delta(m+\delta)^{i-1}} + \|\boldsymbol{A}^{i-1} - (\boldsymbol{A}^*)^{i-1}\|\|\boldsymbol{B}^*\| \tag{C.35}$$

again, we have used the triangle inequality and Cauchy-Schwarz inequality. In order to deal with $\|\boldsymbol{A}^{i-1} - (\boldsymbol{A}^*)^{i-1}\|$, we make use of the equality $x^n - y^n = (x-y)(x^{n-1} + x^{n-2}y + x^{n-3}y^2 + \cdots + xy^{n-2}y^{n-1}$.

Assigning $\boldsymbol{A}$ and $\boldsymbol{A}^*$, we have

$$\|\boldsymbol{A}^{i-1} - (\boldsymbol{A}^*)^{i-1}\| = \left\|(\boldsymbol{A} - \boldsymbol{A}^*)\sum_{j=0}^{i-2}\boldsymbol{A}^{i-2-j}(\boldsymbol{A}^*)^j\right\| \tag{C.36}$$

$$\leqslant \underbrace{\|\boldsymbol{A} - \boldsymbol{A}^*\|}_{\leqslant\delta}\left\|(\sum_{j=0}^{i-2}\boldsymbol{A}^{i-2-j}(\boldsymbol{A}^*)^j\right\| \leqslant \delta(i-1)(m+\delta)^{i-2} \tag{C.37}$$

where the last inequality uses,

$$\left\|\sum_{j=0}^{i-2}\boldsymbol{A}^{i-2-j}(\boldsymbol{A}^*)^j\right\| \leqslant \sum_{j=0}^{i-2}\|\boldsymbol{A}\|^{i-2-j}\|\boldsymbol{A}^*\|^j \tag{C.38}$$

$$\leqslant \sum_{j=0}^{i-2}(m+\delta)^{i-2-j}m^j \tag{C.39}$$

$$\leqslant \sum_{j=0}^{i-2}(m+\delta)^{i-2} = (i-1)(m+\delta)^{i-2}. \tag{C.40}$$

Plugging all terms back together,

$$|l_i| \leqslant m\left(\delta(m+\delta)^{i-1} + \delta(i-1)(m+\delta)^{i-2}m\right) + \delta(m+\delta)^i = O(\delta m^i). \tag{C.41}$$

The above is maximized for $i = k$, to result with $\|\boldsymbol{l}\|_\infty = O(\delta m^k)$.

Plugging back to the bound on the loss, we have

$$\mathcal{L}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = \|\boldsymbol{l}\|^2 \leqslant k\|\boldsymbol{l}\|_\infty^2 = O(\delta^2 k m^{2k}) \tag{C.42}$$

When k is small which is the case of interest, $\mathcal{L}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = O(\delta^2)$.

□

Theorem C.10. ***[Theorem 8 in main paper]*** *Assume the conditions of Theorem 5, and that the teacher parameters $\hat{\Theta}$ are stable, i.e. the singular values of $\hat{\boldsymbol{A}}$ are in $[-1, 1]$. Assume also that $\hat{\Theta}$ are non-degenerate, in the sense that the input-output mapping they realize is not identically zero. Then, for any $\epsilon > 0$ and $q \in \mathbb{N}$, there exists $\delta(\epsilon, q) > 0$ which admits the following: whenever GF leads the student parameters Θ to within (Frobenius) distance $\delta(\epsilon, q)$ from a balanced point Θ^* satisfying $\mathcal{L}(\Theta^*) = 0$, the student ϵ-extrapolates with horizon q.*

Proof of Theorem C.10. Given a desired approximation error $\epsilon > 0$, and horizon $q \in \{k+1, k+2, \dots\}$, our goal is to find $\delta(\epsilon, q)$ such that if $\|\boldsymbol{X} - \boldsymbol{X}^*\|_F < \delta$ for $\boldsymbol{X} \in \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}\}$, then $|\boldsymbol{C}\boldsymbol{A}^j\boldsymbol{B} - \hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^j\hat{\boldsymbol{B}}| < \epsilon$.

First we note that by Theorem 5, Θ^* is an extrapolating solution with respect to $\hat{\Theta}$ (i.e. Θ^* and $\hat{\Theta}$ generate identical impulse responses), therefore we can measure extrapolation with respect to Θ^*. In a similar manner to Lemma C.9, we can define the residual vector $\boldsymbol{l}$ such that its ith entry is defined as

$$l_i = \boldsymbol{C}\boldsymbol{A}^{i-1}\boldsymbol{B} - \boldsymbol{C}^*(\boldsymbol{A}^*)^{i-1}\boldsymbol{B}^*. \tag{C.43}$$

Note that Θ is initialized balanced and by Lemma 6, this property is preserved throughout GF. By Lemma C.3 both Θ and Θ^* have equivalent balanced systems $\bar{\Theta}$ and $\bar{\Theta}^*$ for which the transition matrices are diagonal. For the sake of readability, we overload notations and carry on under the assumption that $\boldsymbol{A}$ and $\boldsymbol{A}^*$ are diagonal with entries $\{A_1, \ldots, A_d\}$ and $\{A_1^*, \ldots, A_d^*\}$ on their respective diagonals.

Another observation is that a balanced RNN must satisfy that its first element in the impulse response moment is non-negative ($\boldsymbol{C}\boldsymbol{B} = \sum c_i b_i = \sum b_i^2 \geqslant 0$). Without loss of generality, we can assume the teacher RNN $\hat{\Theta}$ satisfies $\hat{\boldsymbol{C}}\hat{\boldsymbol{B}} = 1$. Suppose this is not the case, i.e. $\hat{\boldsymbol{C}}\hat{\boldsymbol{B}} = \eta > 0$ (recall we assume the teacher is non-degenerate). One can re-scale $\hat{\boldsymbol{C}}$ and $\hat{\boldsymbol{B}}$ by a factor of $\eta^{-0.5}$ and result with $\frac{\hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^{i-1}\hat{\boldsymbol{B}}}{\eta}$ on the ith entry of its impulse response. The re-scaled system satisfies $\frac{\hat{\boldsymbol{C}}\hat{\boldsymbol{B}}}{\eta} = 1$ and provides no extra information.

With the claims above, we may interpret $\hat{\Theta}$ as a discrete random variable, i.e. $\hat{\boldsymbol{p}} = (\hat{C}_1\hat{B}_1, \ldots, \hat{C}_{\hat{d}}\hat{B}_{\hat{d}})$ defines a probability vector,[7] the random variable is assigned the value $\hat{A}_i$ with probability $\hat{p}_i$. Denote the random variable associated with $\hat{\Theta}$ by $\hat{\mathrm{z}}$. Recall that Θ^* is balanced and attains zero loss, hence it also admits the random variable interpretation (denoted as z^*).

Under the probabilistic perspective, the impulse response of the systems correspond to the moment of the random variables, i.e.

$$\mathbb{E}\left[\hat{\mathrm{z}}^n\right] = \sum_{i=1}^{\hat{d}} \hat{p}_i \hat{A}_i^n = \sum_{i=1}^{\hat{d}} \hat{C}_i \hat{B}_i \hat{A}_i^n = \hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^n\hat{\boldsymbol{B}} \tag{C.44}$$

We can write the loss in terms of the defined residual vector, $\mathcal{L}(\Theta) = \|\boldsymbol{l}\|^2$. From Lemma C.9, we have $\mathcal{L}(\Theta) = O(\delta^2)$. Specifically, $\|\boldsymbol{l}\| \leqslant c\delta^2$ for some constant $c \in \mathbb{R}$. Recall that for $\boldsymbol{v} \in \mathbb{R}^n$, $\|\boldsymbol{v}\|_\infty \leqslant \|\boldsymbol{v}\|_2$, applying to our case, we have $\|\boldsymbol{l}\|_\infty \leqslant \|\boldsymbol{l}\| \leqslant \sqrt{c}\delta$.

Specifically, $|l_0| \leqslant \|\boldsymbol{l}\|_\infty \leqslant \sqrt{c}\delta$ and since we assume $\boldsymbol{C}^*\boldsymbol{B}^* = 1$, we have

$$|l_0| = |\boldsymbol{C}\boldsymbol{B} - \boldsymbol{C}^*\boldsymbol{B}^*| = |\boldsymbol{C}\boldsymbol{B} - 1| \leqslant \sqrt{c}\delta \tag{C.45}$$

which leads to $1 - \sqrt{c}\delta \leqslant \boldsymbol{C}\boldsymbol{B} \leqslant 1 + \sqrt{c}\delta$. Denote $\boldsymbol{C}\boldsymbol{B} = 1 + \beta$ where $\beta \in [-\sqrt{c}\delta, \sqrt{c}\delta]$. A re-scaled student introduces a factor δ into the approximation error, that is

$$\left|\frac{\boldsymbol{C}\boldsymbol{A}^{i-1}\boldsymbol{B}}{1+\beta} - \boldsymbol{C}^*(\boldsymbol{A}^*)^{i-1}\boldsymbol{B}^*\right| = \left|\frac{\boldsymbol{C}\boldsymbol{A}^{i-1}\boldsymbol{B}}{1+\beta} - (1+\beta)\frac{\boldsymbol{C}^*(\boldsymbol{A}^*)^{i-1}\boldsymbol{B}^*}{1+\beta}\right| \tag{C.46}$$

$$= \frac{1}{1+\beta}\left|\boldsymbol{C}\boldsymbol{A}^{i-1}\boldsymbol{B} - (1+\beta)\boldsymbol{C}^*(\boldsymbol{A}^*)^{i-1}\boldsymbol{B}^*\right| \tag{C.47}$$

$$\leqslant \frac{1}{1+\beta}\left|\boldsymbol{C}\boldsymbol{A}^{i-1}\boldsymbol{B} - \boldsymbol{C}^*(\boldsymbol{A}^*)^{i-1}\boldsymbol{B}^*\right| + \frac{1}{1+\beta}|\beta\boldsymbol{C}^*(\boldsymbol{A}^*)^{i-1}\boldsymbol{B}^*| \tag{C.48}$$

Recall that Θ^* is fixed, specifically, $\|\boldsymbol{X}^*\|$ is bounded by some constant $m_{\boldsymbol{X}^*}$ for $\boldsymbol{X} \in \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}\}$. Denote $m \equiv \max\{m_{\boldsymbol{A}^*}, m_{\boldsymbol{B}^*}, m_{\boldsymbol{C}^*}, 1\}$. Combining the latter with Cauchy-Schwarz inequality,

$$\left|\frac{\boldsymbol{C}\boldsymbol{A}^{i-1}\boldsymbol{B}}{1+\beta} - \boldsymbol{C}^*(\boldsymbol{A}^*)^{i-1}\boldsymbol{B}^*\right| \leqslant \frac{1}{1+\beta}|l_i| + \frac{\beta m^{i+1}}{1+\beta} \leqslant 2|l_i| + 2\beta m^{i+1} = O(\delta) \tag{C.49}$$

where the last inequality uses $0.5 < 1 + \beta$ for $\sqrt{c}\delta < 0.5$. Per our claim above, we can assume with out loss of generality that $\boldsymbol{C}\boldsymbol{B} = 1$. A balanced student can thus be interpreted as a discrete random variable with probability vector $\boldsymbol{p}$ taking on values $\{A_1, \ldots, A_d\}$ with probability p_i.

[7] $\sum \hat{p}_i = 1$ and $\forall i,\ \hat{p}_i \geqslant 0$

Recall that under the assumption of a stable teacher, we have that for all $i = 1, \dots, \hat{d},\ |\hat{A}_i| \leqslant 1$. Denote the joint support of $\boldsymbol{A}$ and $\hat{\boldsymbol{A}}$ by $\mathbb{S} \equiv \{A_i\}_{i=1}^{d} \bigcup \{\hat{A}_j\}_{j=1}^{\hat{d}}$, and define $s = \max\{\mathbb{S}\} - \min\{\mathbb{S}\}$. By our assumption on convergence and stability of the teacher, we have $s \leqslant 2(1+\delta)$.

From Proposition 2 in Wu & Yang (2020), we have that if $|\boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B} - \hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^i\hat{\boldsymbol{B}}| \leqslant \delta$ for $i \in \{1, \dots, 2\hat{d}\}$, then

$$\mathcal{W}_1(\boldsymbol{p}, \hat{\boldsymbol{p}}) = O\left(\delta^{\frac{1}{4\hat{d}}}\right). \tag{C.50}$$

By (Panaretos & Zemel, 2019, Section 2.3), we can tie the Wasserstein distance to higher order q^{th} Wasserstien distance. For the s defined above, we have that for any $q \in \mathbb{N}$:

$$\mathcal{W}_q^q(\boldsymbol{p}, \hat{\boldsymbol{p}}) \leqslant \mathcal{W}_1(\boldsymbol{p}, \hat{\boldsymbol{p}}) s^{q-1} < (2+2\delta)^{q-1} O(\delta^{\frac{1}{2\hat{d}}}). \tag{C.51}$$

Denote by $m_q(\boldsymbol{p})$ the qth moment of $\boldsymbol{p}$, from (Biswas & Mackey, 2021, Equation 1), we have:

$$\left|m_q^q(\boldsymbol{p}) - m_q^q(\hat{\boldsymbol{p}})\right| \leqslant \mathcal{W}_q^q(\boldsymbol{p}, \hat{\boldsymbol{p}}) \leqslant (2+2\delta)^{q-1} O(\delta^{\frac{1}{2\hat{d}^*}}). \tag{C.52}$$

Recall that the difference in the ith moments of the student and teacher correspond to the difference in the impulse responses. So we have that if $\|\boldsymbol{X} - \boldsymbol{X}^*\|_F < \delta$ for $\boldsymbol{X} \in \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}\}$ then $\forall q \in \mathbb{N}$,

$$|\boldsymbol{l}_q| \leqslant (2+2\delta)^{q-1} O(\delta^{\frac{1}{2\hat{d}}}). \tag{C.53}$$

□

C.4 Proposition 9 (Implicit Bias for Balancedness)

The proof of Proposition 9 consists of several steps. First, we bound with high probability the norms of $\boldsymbol{B}$ and $\boldsymbol{C}$ at initialization (Lemma C.11). We then derive bounds on the differential equations of $\frac{d}{dt}(\boldsymbol{B}(t) + \boldsymbol{C}^\top(t))$ and $\frac{d}{dt}(\boldsymbol{B}(t) - \boldsymbol{C}^\top(t))$ (Lemma C.15). We show that when the initialization scale tends to zero, the ratio between the differential equations tends to zero. (Lemma C.14).

Before we turn to prove Proposition 9, we first need to bound the initial values for a vector $\boldsymbol{v} \in \mathbb{R}^n$ initialized with $\mathcal{N}(0, \frac{\epsilon^2}{n})$.

Lemma C.11. *Assume a vector $\boldsymbol{v} \in \mathbb{R}^n$ with $\mathcal{N}(0, \frac{\epsilon^2}{n})$ per coordinate. Then:*

$$Pr\left(\frac{\epsilon}{2} < \|\mathbf{v}\|_2 < \frac{3\epsilon}{2}\right) \geqslant 1 - 2\exp(-9d/64). \tag{C.54}$$

Proof of Lemma C.11. The proof of C.11 uses known results on the Chi-square distribution Laurent & Massart (2000), applied to our specific setting to achieve the desired bounds. We will begin by changing variables, $\tilde{v}_i = v_i \cdot \frac{\sqrt{n}}{\epsilon}$. The entries $\tilde{v}_i$, are standard Gaussian variables. The squared norm of $\tilde{\boldsymbol{v}}$ distributes according to the χ-squared distribution.

By (Laurent & Massart, 2000, Lemma 1), in our case (plugging in $x = 9n/64$), the following inequalities hold:

$$Pr\left(\|\tilde{\boldsymbol{v}}\|^2 \geqslant (1.75 + 9/32)n\right) \leqslant \exp(-9n/64), \tag{C.55}$$

$$Pr\left(\|\tilde{\boldsymbol{v}}\|^2 \leqslant d/4\right) \leqslant \exp(-9n/64). \tag{C.56}$$

We simplify these inequalities. From the first bound:

$$Pr\left(\|\tilde{\boldsymbol{v}}\|^2 \geqslant 2.25n\right) \leqslant \exp(-9n/64) \Longrightarrow Pr\left(\|\tilde{\boldsymbol{v}}\| \geqslant 1.5\sqrt{n}\right) \leqslant \exp(-9n/64) \tag{C.57}$$

Changing variables back to $\boldsymbol{v}$:

$$Pr\left(\|\boldsymbol{v}\| \geqslant 1.5\epsilon\right) \leqslant \exp(-9n/64). \tag{C.58}$$

Similarly, for the second bound:

$$Pr\left(\|\tilde{\boldsymbol{v}}\| \leqslant \sqrt{n}/2\right) \leqslant \exp(-9n/64) \Longrightarrow Pr\left(\|\boldsymbol{v}\| \leqslant \epsilon/2\right) \leqslant \exp(-9n/64). \tag{C.59}$$

Taking the complementary probability, we have the desired result of

$$Pr\left(\frac{\epsilon}{2} < \|\boldsymbol{v}\| < \frac{3\epsilon}{2}\right) \leqslant 1 - 2\exp(-9n/64). \tag{C.60}$$

□

Proposition C.12. ***[Proposition 9 in main paper]*** *Suppose that:* (i) $d > 4$*;* (ii) *the teacher parameters* $\hat{\Theta}$ *are balanced and are non-degenerate, in the sense that the input-output mapping they realize is not identically zero; and* (iii) *the student parameters are learned by applying GF to the loss* $\mathcal{L}(\cdot)$*. Let* $\tilde{\Theta}$ *be a random point in parameter space, with entries drawn independently from the standard normal distribution. For* $\epsilon > 0$*, consider the case where GF emanates from the initialization* $\epsilon\tilde{\Theta}$*, and denote the resulting curve by* $\Theta_\epsilon(\tau) = (\boldsymbol{A}_\epsilon(\tau), \boldsymbol{B}_\epsilon(\tau), \boldsymbol{C}_\epsilon(\tau))$*, with* $\tau \geqslant 0$*. Then, w.p. at least* 0.75*, for every* $\epsilon > 0$ *there exists* $\tau_\epsilon \geqslant 0$ *such that:*

$$\lim_{\epsilon \to 0^+} \frac{||\boldsymbol{B}_\epsilon(\tau_\epsilon) - \boldsymbol{C}_\epsilon^\top(\tau_\epsilon)||_F}{||\boldsymbol{B}_\epsilon(\tau_\epsilon) + \boldsymbol{C}_\epsilon^\top(\tau_\epsilon)||_F} = 0\,. \tag{C.61}$$

The consequence of Proposition 9 is that as ϵ converges to zero, $\boldsymbol{B}$ and $\boldsymbol{C}$ converge towards each other.

For convenience, we refer to the mentioned initialization scheme (where every coordinate in a vector is initialized as $\mathcal{N}\left(0, \frac{\epsilon^2}{d}\right)$) as ***$\epsilon$-normal initialization***. In order to prove the proposition we define a few relevant terms,

$$Y = \boldsymbol{B} - \boldsymbol{C}^\top, \qquad W = \boldsymbol{B} + \boldsymbol{C}^\top, \qquad w_0 = \hat{\boldsymbol{C}}\hat{\boldsymbol{B}}. \tag{C.62}$$

We will in fact prove the stronger, following lemma, for any matrix $\boldsymbol{A}$, not necessarily symmetric.

Lemma C.13. *Assume* $w_0 > 0$*, and* $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ *are* ϵ*-normally initialized. Then* $\exists t$ *such that*

$$\lim_{\epsilon \to 0} \frac{\|Y(t)\|^2}{\|W(t)\|^2} = 0, \qquad \lim_{\epsilon \to 0} \frac{\|\boldsymbol{A}(t)\|_F^2}{\|W(t)\|^2} = 0. \tag{C.63}$$

The proof of Lemma C.13 follows three steps. (I) We show that there is a time in the optimization in which the norms of all parameters are bounded (Lemma C.14); (II) we derive upper (and lower) bounds for the differential equations describing the evolvement of Y, W and $\boldsymbol{A}$. Our approximations are limited to the initial phase of training. Concretely, we show that for $0 \leqslant t \leqslant \frac{1}{2w_0}\ln\left(\frac{1}{\epsilon^{0.5}}\right)$, all norms are bounded. Thus, it is possible to obtain meaningful bounds on the ODEs of Y and W while $\boldsymbol{A}$ remains in the magnitude of initialization. (III) Using the relevant bounds, we show that

Lemma C.14. *Assume* ϵ*-initialization and bounded teacher. Then, for all* $0 \leqslant t \leqslant \frac{1}{2w_0}\ln\left(\frac{1}{\epsilon^{0.5}}\right)$*, there exist* M_1, M_2 *such that:*

$$\|\boldsymbol{C}(t)\|, \|\boldsymbol{B}(t)\| < M_1\epsilon^{0.75} \tag{C.64}$$

and

$$\|\boldsymbol{A}(t)\|_F < M_2\epsilon \tag{C.65}$$

To prove this, we note that at initialization, $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{C}$ satisfy these bounds. From continuity, there exists a maximal time for which they are satisfied. We bound the rate of their growth, and thus show that for all t as described, we are within this region.

Lemma C.15. *Assume* $w_0 > 0$ *and assume* $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ *are* ϵ*-normally initialized, we have the following bounds for all* $0 \leqslant t \leqslant \frac{1}{2w_0}\ln\left(\frac{1}{\epsilon^{0.5}}\right)$

$$Y(t)^\top Y(t) \leqslant c_1\epsilon^2 e^{-2w_0 t} + c_2\epsilon^{2.5}, \tag{C.66}$$

and

$$W(t)^\top W(t) \geqslant c_3\epsilon^2 e^{2w_0 t} - c_4\epsilon^{2.5}. \tag{C.67}$$

Lemma C.15 shows that the growth rate of $W(t)^\top W(t)$ and the decay rate of $Y(t)^\top Y(t)$ both depend on the sign of w_0. In our analysis we assume $w_0 > 0$ for simplicity of presentation, the same analysis applies for $w_0 < 0$ with opposite roles for Y and W. The proof of Lemma C.15 follows from writing the leading terms of the ODE and bounding the remaining terms by their upper bounds in the time considered. Using these lemmas, we proceed to prove Lemma C.13.

Proof of Lemma C.13. Consider the dynamics at time $0 \leqslant t \leqslant \frac{1}{2w_0} \ln\left(\frac{1}{\epsilon^{0.5}}\right)$. By Lemma C.15, we have

$$Y(t)^\top Y(t) \leqslant c_1 \epsilon^2 e^{-\ln \frac{1}{\epsilon^{0.5}}} + c_2 \epsilon^{2.5} = (c_1 + c_2)\epsilon^{2.5} \tag{C.68}$$

and

$$W(t)^\top W(t) \geqslant c_3 \epsilon^2 e^{\ln \frac{1}{\epsilon^{0.5}}} - c_4 \epsilon^{2.5} = c_3 \epsilon^{1.5} - c_4 \epsilon^{2.5} \tag{C.69}$$

We can calculate the limit

$$\lim_{\epsilon \to 0} \frac{\|Y(t)\|^2}{\|W(t)\|^2} \leqslant \lim_{\epsilon \to 0} \frac{(c_1 + c_2)\epsilon^{2.5}}{c_3 \epsilon^{1.5} - c_4 \epsilon^{2.5}} = 0 \tag{C.70}$$

From Lemma C.14, $\|\boldsymbol{A}(t)\|_F \leqslant M_2 \epsilon$, so we can calculate the limit,

$$\lim_{\epsilon \to 0} \frac{\|\boldsymbol{A}(t)\|_F^2}{\|W(t)\|^2} \leqslant \lim_{\epsilon \to 0} \frac{M_2 \epsilon^2}{c_3 \epsilon^{1.5} - c_4 \epsilon^{2.5}} = 0 \tag{C.71}$$

which concludes the proof. □

Proof of Lemma C.14. At $t = 0$, by Lemma C.11, $\|\boldsymbol{A}\|_F < 2\epsilon < M_2 \epsilon, \|\boldsymbol{B}\|, \|\boldsymbol{C}\| < 2\epsilon < M_1 \epsilon^{0.75}$, where we bound $M_1 > 2\epsilon^{0.25}$.

If there exists t s.t. $\|\boldsymbol{A}(t)\|_F > M_2 \epsilon$ (otherwise, we get trivially the desired bound for $\|\boldsymbol{A}(t)\|$), from continuity and the intermediate value theorem, there exists a maximal t_A s.t. for $0 \leqslant t < t_1, \|\boldsymbol{A}(t)\|_F < \epsilon, \|\boldsymbol{A}(t_1)\|_F = M_2 \epsilon$. By similar argument for B, C, we can define t_B, t_C. By taking the minimal value between them, we get T^* s.t. $\forall 0 \leqslant t < T^*, \|\boldsymbol{A}(t)\|_F < M_2 \epsilon, \|\boldsymbol{B}\|, \|\boldsymbol{C}\| < M_1 \epsilon^{0.75}$.

We start by proving the lemma under the assumption that $0 \leqslant t < T^*$. Later, we will show that $T^* > T$. Let us recall the derivative of $\boldsymbol{B}$:

$$\dot{\boldsymbol{B}} = -\sum_{i=0}^{k-1} \nabla \ell_i (\boldsymbol{A}^\top)^i \boldsymbol{C}^\top. \tag{C.72}$$

Bounding it by the Cauchy-Schwartz inequality, we have

$$\|\dot{\boldsymbol{B}}\| = \left\| -\sum_{i=0}^{k-1} \nabla \ell_i (\boldsymbol{A}^\top)^i \boldsymbol{C}^\top \right\| \leqslant \sum_{i=0}^{k-1} |\nabla \ell_i| \left\|(\boldsymbol{A}^\top)^i\right\|_F \left\|\boldsymbol{C}^\top\right\|. \tag{C.73}$$

Denote $M = \max_i(|w_i|) + M_1^2 \epsilon^{1.5}$, then we have

$$M = \max_i(|w_i|) + M_1^2 \epsilon^{1.5} \geqslant \max_i(|w_i|) + \|\boldsymbol{C}\|\|\boldsymbol{B}\| \geqslant \max_i(|w_i|) + |\boldsymbol{C}\boldsymbol{B}|, \tag{C.74}$$

$$\geqslant \max_i(|w_i|) + \max_i(|\boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B}|) \geqslant \max_i(|w_i| + |\boldsymbol{C}\boldsymbol{A}^i\boldsymbol{B}|) \geqslant \max_i(|\nabla \ell_i|). \tag{C.75}$$

Recall the conservation law from Lemma C.7 for the norms of $\boldsymbol{B}$ and $\boldsymbol{C}$:

$$\forall t,\ \frac{d}{dt}\left(\|\boldsymbol{B}(t)\| - \|\boldsymbol{C}(t)\|\right) = 0. \tag{C.76}$$

By Lemma C.11, we have

$$\|\boldsymbol{B}(0)\|, \|\boldsymbol{C}(0)\| < 2\epsilon \Rightarrow (|\|\boldsymbol{B}(0)\| - \|\boldsymbol{C}(0)\||) < 4\epsilon. \tag{C.77}$$

Therefore, we get

$$\forall t,\ \|\boldsymbol{C}(t)\| < \|\boldsymbol{B}(t)\| + 4\epsilon. \tag{C.78}$$

Note also that by assuming $M_2\epsilon < 1$, we have

$$\sum_{i=0}^{k-1}(M_2\epsilon)^i < k. \tag{C.79}$$

Plugging the various values into equation C.73, we have:

$$\|\dot{\boldsymbol{B}}\| \leqslant \sum_{i=0}^{k-1}|\nabla\ell_i|\|(\boldsymbol{A}^\top)\|_F^i\|\boldsymbol{C}^\top\| < M(\|\boldsymbol{B}\|+4\epsilon)\sum_{i=0}^{k-1}\|(\boldsymbol{A}^\top)\|_F^i \tag{C.80}$$

$$< Mk(\|\boldsymbol{B}\|+4\epsilon). \tag{C.81}$$

Denoting $R = \|\boldsymbol{B}\|^2 = \boldsymbol{B}^\top\boldsymbol{B}$, then

$$\dot{R} = \dot{\boldsymbol{B}}^\top\boldsymbol{B} + \boldsymbol{B}^\top\dot{\boldsymbol{B}} = 2\boldsymbol{B}^\top\dot{\boldsymbol{B}} \tag{C.82}$$

Taking absolute value and then plugging equation C.80 results with

$$|\dot{R}| = |2\boldsymbol{B}^\top\dot{\boldsymbol{B}}| \leqslant 2\|\boldsymbol{B}^\top\|\|\dot{\boldsymbol{B}}\| < 2k\|\boldsymbol{B}^\top\|M(\|\boldsymbol{B}\|+4\epsilon). \tag{C.83}$$

Using the definition of R, we get that

$$|\dot{R}| < 2kM(|R| + 4\epsilon\sqrt{|R|}) \tag{C.84}$$

Notice that for all t s.t. $\|\boldsymbol{B}\| = \sqrt{R} < 4\epsilon < M_1\epsilon^{0.75}$, we are done. For any t_2 s.t. $\sqrt{R(t_2)} > 4\epsilon$, from continuity and the intermediate value theorem, there exist a maximal t_1 s.t. $0 < t_1 < t_2, \sqrt{R(t_1)} = 4\epsilon$. For all $t_1 < t < t_2, \sqrt{R(t)} > 4\epsilon$. Thus,

$$\forall t_1 < t \leqslant t_2,\ |\dot{R}| < 4kM|R| \tag{C.85}$$

Integrating by t and substituting, we get that for all $t_1 < t_{int} < t_2$:

$$\int_{t_1}^{t_{int}}\frac{1}{|R|}|\dot{R}|dt < \int_{t_1}^{t_{int}}4kMdt \tag{C.86}$$

$$\Rightarrow \int_{R(t_1)}^{R(t_{int})}\frac{1}{|R|}dR < 8M[t_{int}-t_1] < 4kM[T-0] \tag{C.87}$$

$$\Rightarrow \ln\left(\frac{|R(t_{int})|}{|R(t_1)|}\right) < 4kMT \tag{C.88}$$

$$\Rightarrow |R(t_{int})| < \exp(4kMT)|R(t_1)| = (4\epsilon)^2\exp\left(4kM\cdot\ln\left(\frac{1}{\epsilon^{0.5}}\right)\right) \tag{C.89}$$

$$\Rightarrow |R(t_{int})| < (4\epsilon)^2\exp(4kM)\frac{1}{\epsilon^{0.5}}. \tag{C.90}$$

Denoting $M_1^2 = 16\cdot\exp(4kM)$, we have:

$$\|\boldsymbol{B}(t_{int})\| < M_1\epsilon^{0.75} \tag{C.91}$$

To summarize, we have shown that for all $0 < t < T$, there exists M_1 s.t $\|\boldsymbol{B}(t)\| < M_1\epsilon^{0.75}$.

Using an identical proof with the roles of $\boldsymbol{C}^\top, \boldsymbol{B}$ reversed, we have $|\boldsymbol{C}(t)| < M_1\epsilon^{0.75}$. As we have defined t_B, t_C as the maximal points s.t. this property occurs for $T < t_B, t_C$.

Next, we will bound $\|\boldsymbol{A}\|_F$. Recall that $\|\boldsymbol{A}\|_F^2 = Tr(\boldsymbol{A}^\top\boldsymbol{A})$, and

$$\frac{d}{dt}Tr(\boldsymbol{A}^\top\boldsymbol{A}) = Tr\left(\frac{d}{dt}(\boldsymbol{A}^\top\boldsymbol{A})\right) \tag{C.92}$$

$$= Tr\left(\dot{\boldsymbol{A}}^\top\boldsymbol{A}\right) + Tr\left(\boldsymbol{A}^\top\dot{\boldsymbol{A}}\right) \tag{C.93}$$

$$= 2Tr\left(\boldsymbol{A}^\top\dot{\boldsymbol{A}}\right), \tag{C.94}$$

where we have used the linearity of trace and its invariance to transpose. Recall that

$$\dot{\boldsymbol{A}} = -\sum_{i=1}^{k-1} \nabla \ell_i \sum_{r=0}^{i-1} (\boldsymbol{A}^\top)^r \boldsymbol{C}^\top \boldsymbol{B}^\top (\boldsymbol{A}^\top)^{i-r-1}. \tag{C.95}$$

Multiplying it from the left by $\boldsymbol{A}^\top$ and then taking trace provides us with

$$Tr(\boldsymbol{A}^\top \dot{\boldsymbol{A}}) = -\sum_{i=1}^{k-1} \nabla \ell_i \sum_{r=0}^{i-1} Tr\left((\boldsymbol{A}^\top)^{r+1} \boldsymbol{C}^\top \boldsymbol{B}^\top (\boldsymbol{A}^\top)^{i-r-1}\right). \tag{C.96}$$

Taking a transpose and then using the cyclic property of trace and, we have for the terms in the sum:

$$Tr\left((\boldsymbol{A}^\top)^{r+1} \boldsymbol{C}^\top \boldsymbol{B}^\top (\boldsymbol{A}^\top)^{i-r-1}\right) = Tr\left(\boldsymbol{B}^\top (\boldsymbol{A}^\top)^i \boldsymbol{C}^\top\right) \tag{C.97}$$

$$= Tr\left(\boldsymbol{C}\boldsymbol{A}^i \boldsymbol{B}\right) \tag{C.98}$$

$$= \boldsymbol{C}\boldsymbol{A}^i \boldsymbol{B} \tag{C.99}$$

Thus, we end up with

$$Tr(\boldsymbol{A}^\top \dot{\boldsymbol{A}}) = -\sum_{i=1}^{k-1} \nabla \ell_i \sum_{r=0}^{i-1} \boldsymbol{C}\boldsymbol{A}^i \boldsymbol{B} = -\sum_{i=1}^{k-1} \nabla \ell_i \cdot i \cdot \boldsymbol{C}\boldsymbol{A}^i \boldsymbol{B}. \tag{C.100}$$

We use it to bound $Tr(\boldsymbol{A}^\top \dot{\boldsymbol{A}})$:

$$Tr(\boldsymbol{A}^\top \dot{\boldsymbol{A}}) \leqslant \sum_{i=1}^{k-1} |\nabla \ell_i| \cdot i \cdot |\boldsymbol{C}\boldsymbol{A}^i \boldsymbol{B}| \tag{C.101}$$

Using the Cauchy-Schwartz inequality and then plugging $M > |\nabla \ell_i|$ and the bounds found for $\|\boldsymbol{C}\|, \|\boldsymbol{B}\| < M_1 \epsilon^{0.75}$ leads to

$$|\nabla \ell_i| \cdot i \cdot |\boldsymbol{C}\boldsymbol{A}^i \boldsymbol{B}| \leqslant M \cdot i \cdot \|\boldsymbol{C}\| \|\boldsymbol{A}\|_F^i \|\boldsymbol{B}\| \leqslant M \cdot i \cdot M_1^2 \epsilon^{1.5} \|\boldsymbol{A}\|_F^i \tag{C.102}$$

Putting all together provides us with

$$Tr(\boldsymbol{A}^\top \dot{\boldsymbol{A}}) \leqslant \sum_{i=1}^{k-1} M \cdot i \cdot M_1^2 \epsilon^{1.5} \|\boldsymbol{A}\|_F^i = M \cdot M_1^2 \epsilon^{1.5} \sum_{i=1}^{k-1} i \cdot \|\boldsymbol{A}\|_F^i \tag{C.103}$$

Noting that $\|\boldsymbol{A}\|_F < 1$ and denoting $\tilde{M}_2 = \frac{k^2}{2} M \cdot M_1^2$ leads to

$$\sum_{i=1}^{k-1} i \|\boldsymbol{A}\|_F^i < \|\boldsymbol{A}\|_F \frac{k(k-1)}{2} < \frac{k^2}{2} \|\boldsymbol{A}|\|_F \tag{C.104}$$

Therefore, we get that

$$Tr(\boldsymbol{A}^\top \dot{\boldsymbol{A}}) < \tilde{M}_2 \epsilon^{1.5} \|\boldsymbol{A}\|_F. \tag{C.105}$$

Notice that

$$\|\boldsymbol{A}\|_F \cdot \frac{d}{dt}(\|\boldsymbol{A}\|_F) = 0.5 \frac{d}{dt}(\|\boldsymbol{A}\|_F^2) = Tr(\boldsymbol{A}^\top \dot{\boldsymbol{A}}) \tag{C.106}$$

$$\Rightarrow \|\boldsymbol{A}\|_F \cdot \frac{d}{dt}(\|\boldsymbol{A}\|_F) < \tilde{M}_2 \epsilon^{1.5} \|\boldsymbol{A}\|_F. \tag{C.107}$$

$$\Rightarrow \frac{d}{dt}(\|\boldsymbol{A}\|_F) < \tilde{M}_2 \epsilon^{1.5} \tag{C.108}$$

Therefore, for any $0 \leqslant t_3 \leqslant T$:

$$\|\boldsymbol{A}(t_3)\|_F - \|\boldsymbol{A}(0)\|_F = \int_0^{t_3} \frac{d}{dt}(\|\boldsymbol{A}\|_F) dt < \tilde{M}_2 \epsilon^{1.5} t_3 \tag{C.109}$$

$$\Rightarrow \|\boldsymbol{A}(t_3)\|_F < \tilde{M}_2 \epsilon^{1.5} T + \|\boldsymbol{A}(0)\|_F. \tag{C.110}$$

Recall that by Lemma C.11, we have for any $x > 0, \ln(x) < x$ that

$$\|\boldsymbol{A}(0)\|_F < 4d\epsilon. \tag{C.111}$$

Using the fact that

$$T = \frac{1}{2w_0} \ln\left(\frac{1}{\epsilon^{0.5}}\right) < \frac{1}{2w_0}\epsilon^{-0.5} \tag{C.112}$$

we get that

$$\Rightarrow \|\boldsymbol{A}(t_3)\|_F < \frac{\tilde{M}_2}{2w_0}\epsilon + 4d\epsilon. \tag{C.113}$$

Taking $M_2 = \frac{\tilde{M}_2}{2w_0} + 4\epsilon$, we have for all $0 \leqslant t \leqslant T$ that

$$\Rightarrow \|\boldsymbol{A}(t)\|_F < M_2\epsilon \tag{C.114}$$

Once again, recall that we defined t_A as the maximal t with this property, thus $T < t_A$. Since we already have also that $T < t_B, t_C$, we get $T < \min(T_A, t_B, t_C) = T^*$. □

C.4.1 Bounding $W^\top W$ and $Y^\top Y$

Proof of Lemma C.15. Denote

$$Y \equiv \boldsymbol{C}^\top - \boldsymbol{B}, \qquad W \equiv \boldsymbol{C}^\top + \boldsymbol{B}. \tag{C.115}$$

Recall that

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{B}} = \sum_{i=0}^{k-1} \nabla \ell_i (\boldsymbol{A}^\top)^i C^\top, \qquad \frac{\partial \mathcal{L}}{\partial \boldsymbol{C}} = B^\top \sum_{i=0}^{k-1} \nabla \ell_i (\boldsymbol{A}^\top)^i \tag{C.116}$$

We can write the change in Y,

$$\begin{aligned} \dot{Y} = \dot{\boldsymbol{C}}^\top - \dot{\boldsymbol{B}} &= -\sum_{i=0}^{k-1} \nabla \ell_i \boldsymbol{A}^i \boldsymbol{B} + \sum_{i=0}^{k-1} \nabla \ell_i (\boldsymbol{A}^\top)^i \boldsymbol{C}^\top \\ &= \sum_{i=0}^{k-1} \nabla \ell_i \left((\boldsymbol{A}^\top)^i \boldsymbol{C}^\top - \boldsymbol{A}^i \boldsymbol{B}\right) \end{aligned} \tag{C.117}$$

Denote $(\boldsymbol{A}^i)_S = \frac{\boldsymbol{A}^i + (\boldsymbol{A}^\top)^i}{2}$ and $(\boldsymbol{A}^i)_{\bar{S}} = \frac{\boldsymbol{A}^i - (\boldsymbol{A}^\top)^i}{2}$ the symmetric and anti-symmetric parts of $\boldsymbol{A}^i$. We can now write

$$\begin{aligned} (\boldsymbol{A}^\top)^i \boldsymbol{C}^\top - \boldsymbol{A}^i \boldsymbol{B} &= \left[(\boldsymbol{A}^i)_S - (\boldsymbol{A}^i)_{\bar{S}}\right] \boldsymbol{C}^\top - \left[(\boldsymbol{A}^i)_S + (\boldsymbol{A}^i)_{\bar{S}}\right] \boldsymbol{B} \\ &= (\boldsymbol{A}^i)_S (\boldsymbol{C}^\top - \boldsymbol{B}) - (\boldsymbol{A}^i)_{\bar{S}} (\boldsymbol{C}^\top + \boldsymbol{B}) \\ &= (\boldsymbol{A}^i)_S Y - (\boldsymbol{A}^i)_{\bar{S}} W \end{aligned} \tag{C.118}$$

Note also that for $i = 0$, we have $\boldsymbol{A}^0 = I$ and its anti-symmetric part is the zero matrix, writing $i = 0$ separately and assigning equation C.118 into equation C.117,

$$\dot{Y} = \nabla \ell_0 Y + \sum_{i=1}^{k-1} \nabla \ell_i \left((\boldsymbol{A}^i)_S Y - (\boldsymbol{A}^i)_{\bar{S}} W\right) \tag{C.119}$$

Let us look at $\frac{d}{dt}(Y^\top Y) = \dot{Y}^\top Y + Y^\top \dot{Y} = 2Y^\top \dot{Y}$.

Multiplying equation C.119 from the left with $Y^\top$ evaluates to

$$Y^\top \dot{Y} = \nabla \ell_0 Y^\top Y + \sum_{i=1}^{k-1} \nabla \ell_i \left(Y^\top (\boldsymbol{A}^i)_S Y - Y^\top (\boldsymbol{A}^i)_{\bar{S}} W\right) \tag{C.120}$$

We now turn to bound the terms in the sum.

$$Y^\top \dot{Y} \leqslant \nabla \ell_0 Y^\top Y + \sum_{i=1}^{k-1} |\nabla \ell_i| \left(|Y^\top (\boldsymbol{A}^i)_S Y| + |Y^\top (\boldsymbol{A}^i)_{\bar{S}} W|\right) \tag{C.121}$$

We can bound each term using Cauchy–Schwarz. We first need to bound $\|Y\|$ and $\|W\|$, which are trivially bounded by

$$\|Y\|, \|W\| \leqslant \|\boldsymbol{C}\| + \|\boldsymbol{B}\| \leqslant 2M_1\epsilon^{0.75} \tag{C.122}$$

As for the symmetric and anti-symmetric parts of $\boldsymbol{A}$,

$$\|(\boldsymbol{A}^i)_S\|_F = \left\| \frac{\boldsymbol{A}^i + (\boldsymbol{A}^i)^\top}{2} \right\|_F \leqslant \frac{1}{2}\left(\|\boldsymbol{A}^i\|_F + \|(\boldsymbol{A}^i)^\top\|_F\right) = \|\boldsymbol{A}^i\|_F \leqslant \|\boldsymbol{A}\|_F^i, \tag{C.123}$$

where the last inequality follows again from Cauchy–Schwarz (the same considerations apply for $(\boldsymbol{A}^i)_{\bar{S}}$).

From Cauchy–Schwarz we can bound $|Y^\top(\boldsymbol{A}^i)_{\bar{S}}W| \leqslant \|Y^\top\|\|(\boldsymbol{A}^i)_{\bar{S}}\|_F\|W\| \leqslant \|Y\|\|\boldsymbol{A}\|_F^i\|W\|$, we denote $M_3 = max(M_1, M_2)$ and derive:

$$|Y^\top(\boldsymbol{A}^i)_{\bar{S}}W| \leqslant 2M_1\epsilon^{0.75}(M_2\epsilon)^i 2M_1\epsilon^{0.75} < 4M_3^3\epsilon^{1.5+i}, \tag{C.124}$$

which is maximized when $i = 1$. We again bound: $M = \max_i(|w_i|) + M_1^2\epsilon^{1.5} > |\nabla\ell_i|$. We can bound the terms in equation C.121 by

$$|\nabla\ell_i|\left(|Y^\top(\boldsymbol{A}^i)_S Y| + |Y^\top(\boldsymbol{A}^i)_{\bar{S}}W|\right) \leqslant 8 \cdot M \cdot M_3^3\epsilon^{2.5} \tag{C.125}$$

Plugging back into equation C.121:

$$Y^\top\dot{Y} \leqslant \nabla\ell_0 Y^\top Y + \sum_{i=1}^{k-1} 8 \cdot M \cdot M_3^3\epsilon^{2.5} \tag{C.126}$$

We can also bound $\nabla\ell_0 = (\boldsymbol{CB} - w_0) \leqslant -w_0 + |\boldsymbol{CB}| \leqslant -w_0 + \|\boldsymbol{C}\|\|\boldsymbol{B}\| \leqslant -w_0 + M_1^2\epsilon^{1.5}$. Note also that we multiply by $Y^\top Y$ so we can bound

$$\nabla\ell_0 Y^\top Y \leqslant -w_0 Y^\top Y + \underbrace{M_1^2\epsilon^{1.5}(M_2 1\epsilon^{0.75})^2}_{=4M_1^4\epsilon^3} \tag{C.127}$$

putting back together, we get

$$Y^\top\dot{Y} \leqslant -w_0 Y^\top Y + 4M_1^4\epsilon^3 + (k-1)(M_1+1)8 \cdot M \cdot M_3^3\epsilon^{2.5} \tag{C.128}$$

In particular, there exists M_4 such that

$$Y^\top\dot{Y} \leqslant -w_0 Y^\top Y + \frac{M_4}{2}\epsilon^{2.5} \tag{C.129}$$

Recall that we were interested in bounding $\frac{d}{dt}(Y^\top Y) = \dot{Y}^\top Y + Y^\top\dot{Y} = 2Y^\top\dot{Y}$,

$$\frac{d}{dt}(Y^\top Y) \leqslant -2w_0 Y^\top Y + M_4\epsilon^{2.5} \tag{C.130}$$

Denoting $z(t) \equiv Y(t)^\top Y(t)$ and $x(t) \equiv W(t)^\top W(t)$, and using Lemma C.16, we have the desired bounds

$$z(t) \leqslant \frac{1}{2w_0}\left[(33w_0 d\epsilon^2)e^{-2w_0 t} + M_4\epsilon^{2.5}\right] \tag{C.131}$$

In particular, we can write

$$z(t) \leqslant c_1\epsilon^2 e^{-2w_0 t} + c_2\epsilon^{2.5} \tag{C.132}$$

and

$$x(t) \geqslant c_3\epsilon^2 e^{2w_0 t} - c_4\epsilon^{2.5} \tag{C.133}$$

where c_i's are positive constants.

Note that the derivation of $x(t)$ is exactly the same as $z(t)$ with opposite signs and bounding from below instead.

□

C.4.2 INTEGRAL BOUND OF $\|Y\|^2$ AND $\|W\|^2$

Lemma C.16. *Assume $\dot{z} < -2w_0 z + M_4\epsilon^{2.5} < 0, \dot{x} > 2w_0 z - M_4\epsilon^{2.5} > 0$, where $w_0 > 0$. Then, under the assumptions of Lemma C.11:*

$$z(t_1) < \frac{1}{2w_0}\left(\exp(-2w_0 t_1)\cdot(75w_0 d\epsilon^2) + M_4\epsilon^{2.5}\right) \tag{C.134}$$

$$x(t_2) > \frac{1}{2w_0}\left(\exp(2w_0 t_2)\cdot(\frac{w_0\epsilon^2}{25d}) + M_4\epsilon^{2.5}\right) \tag{C.135}$$

Proof of Lemma C.16. Assume $\dot{z} < -2w_0 z + M_4\epsilon^{2.5}$, where $w_0 > 0, 2w_0 z > M_4\epsilon^{2.5}$. Similarly, assume $\dot{x} > 2w_0 x - M_4\epsilon^{2.5}$.

Then:

$$\frac{\dot{z}}{2w_0 z - M_4\epsilon^{2.5}} < -1 \tag{C.136}$$

$$\frac{\dot{x}}{2w_0 x - M_4\epsilon^{2.5}} > 1 \tag{C.137}$$

Integrating both sides by dt, and using integration by substitution, we get:

$$-t_1 = \int_0^{t_1} -1 > \int_0^{t_1} \frac{1}{2w_0 z - M_4\epsilon^{2.5}}\frac{dz}{dt}dt = \int_{z(0)}^{z(t_1)} \frac{1}{2w_0 z - M_4\epsilon^{2.5}}dz \tag{C.138}$$

$$t_2 = \int_0^{t_2} 1 < \int_0^{t_2} \frac{1}{2w_0 x - M_4\epsilon^{2.5}}\frac{dx}{dt}dt = \int_{x(0)}^{x(t_2)} \frac{1}{2w_0 x - M_4\epsilon^{2.5}}dx \tag{C.139}$$

We note:

$$\int_{z(0)}^{z(t_1)} \frac{1}{2w_0 z - M_4\epsilon^{2.5}}dz = \frac{1}{2w_0}[\ln(2w_0 z(t_1) - M_4\epsilon^{2.5}) - \ln(2w_0 z(0) - M_4\epsilon^{2.5})] \tag{C.140}$$

$$\Rightarrow \int_{z(0)}^{z(t_1)} \frac{1}{2w_0 z - M_4\epsilon^{2.5}}dz = \frac{1}{2w_0}\left[\ln\left(\frac{2w_0 z(t_1) - M_4\epsilon^{2.5}}{2w_0 z(0) - M_4\epsilon^{2.5}}\right)\right] \tag{C.141}$$

$$\int_{x(0)}^{x(t_2)} \frac{1}{2w_0 x - M_4\epsilon^{2.5}}dx = \frac{1}{2w_0}\left[\ln\left(\frac{2w_0 x(t_2) - M_4\epsilon^{2.5}}{2w_0 x(0) - M_4\epsilon^{2.5}}\right)\right] \tag{C.142}$$

Combining equations, we have:

$$\frac{1}{2w_0}\left[\ln\left(\frac{2w_0 z(t_1) - M_4\epsilon^{2.5}}{2w_0 z(0) - M_4\epsilon^{2.5}}\right)\right] < -t_1 \tag{C.143}$$

$$\Rightarrow \ln\left(\frac{2w_0 z(t_1) - M_4\epsilon^{2.5}}{2w_0 z(0) - M_4\epsilon^{2.5}}\right) < -2w_0 t_1 \tag{C.144}$$

$$\Rightarrow \frac{2w_0 z(t_1) - M_4\epsilon^{2.5}}{2w_0 z(0) - M_4\epsilon^{2.5}} < \exp(-2w_0 t_1) \tag{C.145}$$

$$\Rightarrow 2w_0 z(t_1) - M_4\epsilon^{2.5} < \exp(-2w_0 t_1)\cdot\left(2w_0 z(0) - M_4\epsilon^{2.5}\right) \tag{C.146}$$

$$\Rightarrow z(t_1) < \frac{1}{2w_0}\left[\exp(-2w_0 t_1)\cdot(2w_0 z(0) - M_4\epsilon^{2.5}) + M_4\epsilon^{2.5}\right] \tag{C.147}$$

$$x(t_2) > \frac{1}{2w_0}\left[\exp(2w_0 t_2)\cdot(2w_0 x(0) - M_4\epsilon^{2.5}) + M_4\epsilon^{2.5}\right] \tag{C.148}$$

We note, $z(0) = Y^T(0)Y(0)$. By linearity of sum of variances, $Y(0)$'s entries are distributed according to $\mathcal{N}(\sqrt{2}\cdot\epsilon)$, by Lemma C.11:

$$\frac{\sqrt{2}}{2}\epsilon < Y(0) < \frac{3\sqrt{2}}{2}\epsilon \tag{C.149}$$

From Cauchy-Schwartz, $z(0) < 3\epsilon^2$ with high probability. W is distributed as Y, therefore, similarly, $x(0) > \frac{1}{2}\epsilon^2$. Assuming $M_4\epsilon^{2.5} < \frac{w_0\epsilon^2}{2}$, we have:

$$z(t_1) < \frac{1}{2w_0}\left(\exp(-2w_0t_1)\cdot(6w_0\epsilon^2) + M_4\epsilon^{2.5}\right) \tag{C.150}$$

$$x(t_2) > \frac{1}{2w_0}\left(\exp(2w_0t_2)\cdot(w_0\epsilon^2) + M_4\epsilon^{2.5}\right) \tag{C.151}$$

Concluding the proof. □