A. Background on Convergence of Vector Sequences and Random Variables

In this section we review some the background on convergence of random variables, definitions of convergence of matrix sequences and some of their properties that we use throughout this paper.

Pseudo-Lipschitz continuity For a given $p \geq 1$, a function $f : \mathbb{R}^d \to \mathbb{R}^m$ is called pseudo-Lipschitz of order p, denoted by PL(p), if

$$\|f(\mathbf{x}_1) - f(\mathbf{x}_2)\| \leq C\|\mathbf{x}_1 - \mathbf{x}_2\| \left(1 + \|\mathbf{x}_1\|^{p-1} + \|\mathbf{x}_2\|^{p-1}\right) \tag{26}$$

for some constant $C > 0$.

This is a generalization of the standard definition of Lipshitiz continuity. A PL(1) function is Lipschitz with constant $3C$.

Empirical convergence of a sequence Consider a sequence of vectors $\mathbf{x}(N) = \{\mathbf{x}_n(N)\}_{n=1}^N$ with $\mathbf{x}_n(N) \in \mathbb{R}^d$. So, each $\mathbf{x}(N)$ is a block vector with a total of Nd components. For a finite $p \geq 1$, we say that the vector sequence $\mathbf{x}(N)$ converges empirically with p-th order moments if there exists a random variable $X \in \mathbb{R}^d$ such that

(i) $\mathbb{E}\|X\|_p^p < \infty$; and

(ii) for any $f : \mathbb{R}^d \to \mathbb{R}$ that is pseudo-Lipschitz continuous of order p,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(\mathbf{x}_n(N)) = \mathbb{E}\left[f(X)\right]. \tag{27}$$

In this case, with some abuse of notation, we will write

$$\lim_{n\to\infty} \mathbf{x}_n \overset{PL(p)}{=} X, \tag{28}$$

where we have omitted the dependence on N in $\mathbf{x}_n(N)$. We note that the sequence $\{\mathbf{x}(N)\}$ can be random or deterministic. If it is random, we will require that for every pseudo-Lipschitz function $f(\cdot)$, the limit (27) holds almost surely. In particular, if $\mathbf{x}_n \sim X$ are i.i.d. and $\mathbb{E}\|X\|_p^p < \infty$, then $\mathbf{x}$ empirically converges to X with p^{th} order moments.

Weak convergence (or convergence in distribution) of random variables is equivalent to

$$\lim_{n\to\infty} \mathbb{E}f(X_n) = \mathbb{E}f(X), \quad \text{for all bounded functions } f. \tag{29}$$

It is shown in (Bayati & Montanari, 2011) that PL(p) convergence is equivalent to weak convergence plus convergence in p moment.

Wasserstein-2 distance Let ν and μ be two distributions on some Euclidean space $\mathcal{X}$. The Wasserstein-2 distance between ν and μ is defined as

$$W_2(\nu, \mu) = \left(\inf_{\gamma\in\Gamma} \mathbb{E}\|X - X'\|_2^2\right)^{\frac{1}{2}}, \tag{30}$$

where Γ is the set of all distributions with marginals consistent with ν and μ.

A sequence x_n converges PL(2) to X if and only if the empirical measure $\hat{\mathbb{P}}_N = \frac{1}{N}\sum_{n=1}^N \delta(x - x_n)$ (where $\delta(\cdot)$ is the Dirac measure,) converges in Wasserstein-2 distance to distribution of X (Villani, 2008), i.e.

$$x_n \overset{PL(2)}{=} X \quad \Longleftrightarrow \quad \lim_{n\to\infty} W_2(\hat{\mathbb{P}}_N, \mathbb{P}_X) = 0. \tag{31}$$

For two zero mean Gaussian measure $\nu = \mathcal{N}(0, \Sigma_1), \mu = \mathcal{N}(0, \Sigma_2)$ the Wasserstein-2 distance is given by (Givens et al., 1984)

$$W_2^2(\nu, \mu) = \text{tr}(\Sigma_1 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2} + \Sigma_2). \tag{32}$$

Therefore, for zero mean Gaussian measures, convergence in covariance, implies convergence in Wasserstein-2 distance, and hence if the empirical covariance of a zero mean Gaussian sequence x_n converges to some covaraince matrix Σ, then using (31) $x_n \overset{PL(2)}{=} X$ where $X \sim \mathcal{N}(0, \Sigma)$.

B. Proofs

B.1. Proof of Proposition 2.1

Suppose we are given a convolutional model (5) with impulse response coefficients L_t, $t = 0, \ldots, T-1$. It is well-known from linear systems theory (Kailath, 1980) that linear time-invariant systems are input-output equivalent if and only if they have the same impulse response coefficients. So, we simply need to find matrices (W, F, C) satisfying (9). First consider the single input single output (SISO) case where $n_x = n_y = 1$. Take any set of real non-zero scalars λ_i, $i = 0, \ldots, T-1$, that are distinct and set

$$W = \text{diag}(\lambda_0, \ldots, \lambda_{T-1}), \quad F = 1_T, \tag{33}$$

so there are $n = T$ hidden states. Then, for any t,

$$(CW^tF) = \sum_{k=0}^{T-1} C_k \lambda_k^t. \tag{34}$$

Equivalently, the impulse response coefficients in (9) are given by,

$$[L_0, \cdots, L_{T-1}] = CV, \tag{35}$$

where V is the Vandermode matrix $V_{jt} = \lambda_j^t$. Since the values λ_j are distinct, V is invertible and we can find a vector C matching arbitrary impulse response coefficients. Thus, when $n_x = n_y = 1$, we can find a linear RNN with at most $n = T$ hidden states that match the first T impulse response coefficients. To extend to the case of arbitrary n_x and n_y, we simply create $n_x n_y$ systems, one for each input-output component pair. Since each system will have T hidden states, the total number of states would be $n = Tn_x n_y$.

B.2. Proof of Theorem 3.2

Given $y_t = \sum_{j=0}^{t} \sqrt{\rho_j}\theta_j x_{t-j}$ and $\theta = (\theta_0, \ldots, \theta_{T-1})$, we consider a perturbation in θ, namely Δ_θ. Therefore,

$$\widetilde{y}_t = \sum_{j=0}^{t} \sqrt{\rho_j}\Delta_{\theta j} x_{t-j} \tag{36}$$

and the NTK for this model is given by

$$K_{t,s}(x, x') = \sum_{\Delta_\theta \in T_\theta} \widetilde{y}_t(\Delta_\theta)\widetilde{y}'_s(\Delta_\theta)^{\mathsf{T}}. \tag{37}$$

where T_θ is the standard basis for the parameter space. The following lemma shows this sum can be calculated as an expectation over a Gaussian random variable.

Lemma B.1. *Let V be a finite dimensional Hilbert space and $W = \mathbb{R}^m$ with the standard inner product and let $T, T' : V \to W$ be linear transformations. Let $\{v_i\}_{i=1}^n$ be an ordered orthonormal basis for V. Then we have*

$$\sum_{i=1}^{n} T(v_i)(T'(v_i))^{\mathsf{T}} = \mathbb{E}_{\alpha \sim \mathcal{N}(0, I_n)} \left[T\Big(\sum_{i=1}^{n} \alpha_i v_i\Big) \Big(T'\Big(\sum_{i=1}^{n} \alpha_i v_i\Big) \Big)^{\mathsf{T}} \right]. \tag{38}$$

Proof.

$$\begin{aligned}\mathbb{E}_{\alpha \sim \mathcal{N}(0, I_n)} \left[T\Big(\sum_{i=1}^{n} \alpha_i v_i\Big) \Big(T'\Big(\sum_{j=1}^{n} \alpha_j v_j\Big) \Big)^{\mathsf{T}} \right] &= \mathbb{E}_{\alpha \sim \mathcal{N}(0, I_n)} \left[\sum_{i,j=1}^{n} \alpha_i \alpha_j T\big(v_i\big) T'\big(v_j\big)^{\mathsf{T}} \right] \\ &= \sum_{i=1}^{n} T(v_i)(T'(v_i))^{\mathsf{T}}. \end{aligned} \tag{39}$$

□

Since $\widetilde{y}_t(\Delta_\theta)$ is a linear operator, by applying Lemma B.1 we have,

$$\begin{aligned}
K_{t,s}(x,x') &= \mathbb{E}_{\Delta_\theta\sim\mathcal{N}(0,1),\text{i.i.d.}}\,[\widetilde{y}_t(\Delta_\theta)\widetilde{y}'_s(\Delta_\theta)^\mathsf{T}] \\
&= \mathbb{E}_{\Delta_\theta\sim\mathcal{N}(0,1),\text{i.i.d.}}\,\left[(\sum_{j=0}^{t}\sqrt{\rho_j}\Delta_{\theta j}x_{t-j})(\sum_{k=0}^{t}\sqrt{\rho_k}\Delta_{\theta k}x'_{s-k})^\mathsf{T}\right] \\
&= \mathbb{E}_{\Delta_\theta\sim\mathcal{N}(0,1),\text{i.i.d.}}\,\left[(\sum_{j=0}^{t}\rho_j\;\Delta_{\theta j}x_{t-j}{x'_{s-j}}^\mathsf{T}\Delta_{\theta j}^\mathsf{T})\right]
\end{aligned}$$

Therefore,

$$\begin{aligned}
\Big(K_{t,s}(x,x')\Big)_{m,m'} &= \mathbb{E}_{\Delta_\theta\sim\mathcal{N}(0,1),\text{i.i.d.}}\,\left[\sum_{j=0}^{t}\rho_j\;\sum_{k,k'}(\Delta_{\theta j})_{m,k}x_{t-j,k}x'_{s-j,k'}(\Delta_{\theta j})_{k',m'}\right] \\
&= (\sum_{j=0}^{t}\rho_j\;x_{t-j}^\mathsf{T}x'_{s-j})\delta_{m,m'}
\end{aligned}$$

Thus, $K_{t,s}(x,x') = (\sum_{j=0}^{t}\rho_j\;x_{t-j}^\mathsf{T}x'_{s-j})I_{n_y}$ and we can write the full kernel as

$$K(x,x') = \mathcal{T}(x)^\mathsf{T}D(\rho)\mathcal{T}(x')\otimes I_{n_y}, \tag{40}$$

where $\mathcal{T}(x)$ and $D(\rho)$ are defined in (18) and (19) respectively.

B.3. Proof of Theorem 3.3

Part (a) is a special case of a more general lemma, Lemma C.1 which we present in Appendix C. Let

$$q_{t+1} = \frac{1}{\sqrt{n}}Wq_t, \quad q_0 = F, \tag{41}$$

so that q_t represents the impulse response from x_t to h_t. That is,

$$h_t = \sum_{j=0}^{t}q_{t-j}x_j, \tag{42}$$

which is the convolution of q_t and h_t. The system (41) is a special case of (66) with $L = 1$, no input u_t and

$$A = W, \quad G(q) = q.$$

Since there is only $L = 1$ transform, we have dropped the dependence on the index ℓ. Lemma C.1 shows that $(q_0,\ldots,q_t)$ converges $PL(2)$ to a Gaussian vector $(Q_0,\ldots,Q_t)$ with zero mean. We claim that the Q_i's are independent. We prove this with induction. Suppose $(Q_0,\ldots,Q_t)$ are independent. We need to show $(Q_0,\ldots,Q_{t+1})$ are independent by using the SE equations (70). Specifically, from (70b), $Z_i = Q_i$ for all i. Also, since each Q_i is zero mean, $\mu_i = 0$ and $\widetilde{Z}_i = Z_i = Q_i$. Since the $\widetilde{Z}_i$ are independent, the linear predictor coefficients in (70d) are zero: $F_{ti} = 0$. Therefore, $\widetilde{R}_t = R_t \sim \mathcal{N}(0,\nu_W P_t)$ is independent of $(R_0,\ldots,R_{t-1})$. From (70h), $Q_{t+1} = R_t$. So, we have that $(Q_0,\ldots,Q_{t-1})$ is an independent Gaussian vector. Finally, to compute the variance of the Q_{t+1}, observe

$$\begin{aligned}
\text{cov}(Q_{t+1}) &\overset{(a)}{=} \text{cov}(R_t) \overset{(b)}{=} \nu_W P_t \\
&\overset{(c)}{=} \nu_W\text{cov}(Z_t) \overset{(d)}{=} \nu_W\text{cov}(Q_t),
\end{aligned} \tag{43}$$

where (a) follows from (70h); (b) follows from (70f) and the fact that $F_{ti} = 0$ for all i; (c) follows from (70e); and (d) follows from the fact that $Z_t = Q_t$. Also, since $q_0 = F$, it follows that $Q_0 \sim \mathcal{N}(0,\nu_F I)$. We conclude that $\text{cov}(Q_t) = \nu_F\nu_W^t I_{n_x}$. This proves part (a).

For part (b), we consider perturbations Δ_W, Δ_F, and Δ_C of the parameters W, F, and C. We have that,

$$\widetilde{h}_t = \frac{1}{\sqrt{n}} W \widetilde{h}_{t-1} + \frac{1}{\sqrt{n}} \Delta_W h_t + \Delta_F x_t, \qquad \widetilde{y}_t = \frac{1}{\sqrt{n}} C \widetilde{h}_t + \Delta_C h_t \tag{44}$$

Combining this equation with (1), we see that the mapping from x_t to $[h_t \ \widetilde{h}_t]$ is a linear time-invariant system. Let $q_t \in \mathbb{R}^{n \times 2n_x}$ be its impulse response. The impulse response coefficients satisfy the recursive equations,

$$q_{t+1} = \left[\frac{1}{\sqrt{n}} W q_{t,1}, \frac{1}{\sqrt{n}} (W q_{t,2} + \Delta_W q_{t,1}) \right], \quad q_0 = [F, \Delta_F].$$

We can analyze these coefficients in the LSL using Lemma C.1. Specifically, let $L = 2$ and set

$$A_1 = W, \quad A_2 = \Delta_W.$$

Also, let

$$z_{t1} = \overline{G}_1(q_t) := q_t, \tag{45a}$$

$$z_{t2} = \overline{G}_2(q_t) := [0 \ \ q_{t1}]. \tag{45b}$$

Then, we have the updates,

$$q_{t+1} = \frac{1}{\sqrt{n}} W z_{t1} + \frac{1}{\sqrt{n}} \Delta_W z_{t2}, \quad q_0 = [F, \Delta_F].$$

It follows from Lemma C.1 that $(q_0, \ldots, q_{T-1})$ converges $PL(2)$ to zero mean Gaussian random variables $(Q_0, \ldots, Q_{T-1})$. Note that $Q_t = [Q_{t1}, Q_{t2}]$ where each Q_{t1} and Q_{t2} are random vectors $\in \mathbb{R}^{1 \times n_x}$.

Similar to the proof of the previous theorem, we use induction to show that $(Q_0, \ldots, Q_t)$ are independent. Suppose that the claim is true for t. Then, Z_{i1} and Z_{i2} are functions of Q_i. So, for $\ell = 1, 2$, $Z_{t\ell}$ is independent of $Z_{i\ell}$ for $i < t$. Thus, the prediction coefficients $F_{ti\ell} = 0$ and, as before, $R_{t\ell} \sim \mathcal{N}(0, P_{t\ell})$ independent of $R_{i\ell}$, $i < t$. Thus, $Q_{t+1} = R_{t1} + R_{t2}$ is independent of $(Q_0, \ldots, Q_t)$.

We conclude by computing the $\mathrm{cov}(Q_t)$. We claim that, for all t, the variance of Q_t is of the form,

$$\mathrm{cov}(Q_t) = \begin{bmatrix} \tau_{t1} I_{n_x} & 0 \\ 0 & \tau_{t2} I_{n_x} \end{bmatrix} \tag{46}$$

for scalar τ_{t1}, τ_{t2}. Since $q_0 = [F, \Delta_F]$, we have

$$\tau_{t1} = \nu_F, \quad \tau_{t2} = 1.$$

Now suppose that (46) is true for some t. From (70b),

$$Z_{t1} = Q_t, \quad Z_{t1} = (0, Q_{t1}),$$

from which we obtain that

$$\mathrm{cov}(Z_{t1}) = \begin{bmatrix} \tau_{t1} I_{n_x} & 0 \\ 0 & \tau_{t2} I_{n_x} \end{bmatrix}, \quad \mathrm{cov}(Z_{t2}) = \begin{bmatrix} 0 & 0 \\ 0 & \tau_{t1} I_{n_x} \end{bmatrix}. \tag{47}$$

Therefore, we have

$$\begin{aligned} \mathrm{cov}(Q_{t+1}) &\overset{(a)}{=} \mathrm{cov}(R_{t,1}) + \mathrm{cov}(R_{t,2}) \\ &\overset{(b)}{=} \nu_W P_{t1} + P_{t2} \\ &\overset{(c)}{=} \nu_W \mathrm{cov}(Z_{t1}) + \mathrm{cov}(Z_{t2}) \overset{(d)}{=} \begin{bmatrix} \nu_W \tau_{t1} I_{n_x} & 0 \\ 0 & (\tau_{t1} + \nu_W \tau_{t2}) I_{n_x} \end{bmatrix}, \end{aligned} \tag{48}$$

where (a) follows from (70h); (b) follows from (70f) and the fact that $F_{ti\ell} = 0$ for all i; (c) follows from (70e); and (d) follows from (47). It follows that

$$\tau_{t+1,1} = \nu_W \tau_{t1}, \quad \tau_{t+1,2} = \nu_W \tau_{t2} + \tau_{t1}.$$

These recursions have the solution,

$$\tau_{t1} = \nu_W^t \nu_F, \quad \tau_{t2} = t\nu_F \nu_W^{t-1} + \nu_W^t. \tag{49}$$

Since $[h_t, \widetilde{h}_t] = \sum_{j=0}^t q_j \begin{bmatrix} x_{t-j} \\ x_{t-j} \end{bmatrix}$ and we know each q_t converges $PL(2)$ to random Q_t with covariances calculated in (46) and (49), we have

$$[h_t, \widetilde{h}_t] \overset{PL(2)}{=} [H_t, \widetilde{H}_t] = \sum_{j=0}^t Q_j \begin{bmatrix} x_{t-j} \\ x_{t-j} \end{bmatrix} \tag{50}$$

where $H_t, \widetilde{H}_t$ are scalar random variables. For each t, s, we can now calculate the auto-correlation function for H as follows

$$\begin{aligned} \mathbb{E}[H_t H_s] &= \mathbb{E}[\sum_{j=0}^t \sum_{k=0}^t x_{t-j}^\mathsf{T} Q_{j,1}^\mathsf{T} Q_{k,1} x_{s-k}] \\ &= \sum_{j=0}^t x_{t-j}^\mathsf{T} \mathbb{E}[Q_{j,1}^\mathsf{T} Q_{j,1}] x_{s-j} \\ &= \sum_{j=0}^t \nu_W^j \nu_F \; x_{t-j}^\mathsf{T} x_{s-j}. \end{aligned} \tag{51}$$

Similarly for $\widetilde{H}$ we have

$$\mathbb{E}[\widetilde{H}_t \widetilde{H}_s] = \sum_{j=0}^t (j\nu_F \nu_W^{j-1} + \nu_W^j) \; x_{t-j}^\mathsf{T} x_{s-j}. \tag{52}$$

Thus, the impulse response of the system $L_j = C q_{j,1}$ converge empirically to $\mathcal{N}(0, \Lambda)$ where,

$$\Lambda = \nu_C \lim_{n \to \infty} \frac{1}{n} q_{j,1}^\mathsf{T} q_{j,1} = \nu_C \mathbb{E}[Q_{j,1}^\mathsf{T} Q_{j,1}] = \nu_C \nu_F \nu_W^j I_{n_x}. \tag{53}$$

This proves part (a).

Note that $\mathbb{E}[\widetilde{H}_t \widetilde{H}'_s]$ and $\mathbb{E}[H_t H'_s]$ can be calculated similarly by substituting x_{s-j} with x'_{s-j} in (51) and (52).

Next, we calculate the NTK in this case

$$\begin{aligned} K_{t,s}(x, x') &= \sum_{\Delta_\Theta \in T_\Theta} \widetilde{y}_t(\Delta_\Theta) \widetilde{y}'_s(\Delta_\Theta)^\mathsf{T} \\ &\overset{(a)}{=} \mathbb{E}_{\Delta_\Theta \sim \mathcal{N}(0,1), \text{i.i.d.}} \; [\widetilde{y}_t(\Delta_\Theta) \widetilde{y}'_s(\Delta_\Theta)^\mathsf{T}] \end{aligned} \tag{54}$$

where (a) follows from Lemma B.1. Combining with (44) we have

$$\begin{aligned} K_{t,s}(x, x') &= \mathbb{E}_{C, \Delta_C \sim \mathcal{N}(0,1), \text{i.i.d.}} \; \left[(C\widetilde{h}_t + \Delta_C h_t)(C\widetilde{h}'_s + \Delta_C h'_s)^\mathsf{T}\right] \\ &= \left(\nu_C \mathbb{E}[\widetilde{H}_t \widetilde{H}'_s] + \mathbb{E}[H_t H'_s]\right) I_{n_y} \end{aligned} \tag{55}$$

Therefore,

$$K(x, x') = \mathcal{T}(x)^\mathsf{T} D(\rho) \mathcal{T}(x') \otimes I_{n_y}, \tag{56}$$

where $\mathcal{T}(x)$ and $D(\rho)$ are given in (18) and (19) and

$$\rho_i = \nu_C (i\nu_F \nu_W^{i-1} + \nu_W^i) + \nu_W^i \nu_F. \tag{57}$$

This proves part (b).

B.4. Proof of Theorem 4.1

Bounding the Initial Impulse Response from Theorem 3.3, each coefficient of $L^0_{\mathrm{RNN},j}$ has mean zero and variance $\nu_C\nu_F\nu_W^j$. There are n_xn_y such components. This proves (24).

Convolutional Equivalent Linear Model The key for the remainder of the proof is to use Theorems 3.2 and 3.3 to construct a scaled convolutional model that has the same NTK and intial conditions as the RNN. Then, we analyze the convolutional model to obtain the desired bound. To this end, let $\rho = [\rho_0, \ldots, \rho_{T-1}]$ be the scaling factors given in Theorem 3.3. For each initial condition $\theta^0_{\mathrm{RNN}} = (W^0, F^0, C^0)$ of the RNN, suppose that we initialize the scaled convolutional model with

$$\theta^0_{\mathrm{conv},j} = \frac{1}{\sqrt{\rho_j}n^{(j+1)/2}}C^0(W^0)^jF^0.$$

The initial impulse response of the scaled convolutional model will then be

$$L^0_{\mathrm{conv},j} = \sqrt{\rho_j}\theta_{\mathrm{conv},j} = \frac{1}{n^{(j+1)/2}}C^0(W^0)^jF^0 = L^0_{\mathrm{RNN},j}. \tag{58}$$

Hence, the scaled convolutional model and the RNN have the same initial impulse response coefficients. We then train the scaled convolutional model on the training data using gradient descent with the same learning rate η used in the training of the RNN. Let $L^\ell_{\mathrm{conv},j}$ denote the impulse response of the scaled convolutional model after ℓ steps of gradient descent.

Gradient Descent Analysis of the Convolutional Model Next, we look at how the impulse response of the scaled convolutional model evolves over the gradient descent steps. It is convenient to do this analysis using some matrix notation. For each parameter, $\theta = [\theta_0, \ldots, \theta_{T-1}]$, the convolutional filter parameters are $L_j = \sqrt{\rho_j}\theta_j$. Thus, we can write

$$\mathbf{L} = \mathbf{D}^{1/2}\theta,$$

where $\mathbf{D}$ is a block diagonal operator with values ρ_j. Also, let $\widehat{\mathbf{y}} = [\widehat{\mathbf{y}}^1, \ldots, \widehat{\mathbf{y}}^N]$ be the set of predictions on the N training samples. Since the convolutional model is linear, we can write $\hat{\mathbf{y}} = \mathbf{A}\mathbf{L}$ for some linear operator $\mathbf{A}$. The operator $\mathbf{A}$ would be a block Toeplitz with the input data $\mathbf{x} = [\mathbf{x}^1, \ldots, \mathbf{x}^N]$. Also, if we let $\mathbf{y} = [\mathbf{y}^1, \ldots, \mathbf{y}^N]$ be the N training samples, the least squares cost is

$$\|\mathbf{y} - \mathbf{A}\mathbf{D}^{1/2}\theta\|_F^2.$$

Minimizing this loss function will result in GD steps,

$$\theta^{\ell+1} = \theta^\ell + \eta\mathbf{D}^{1/2}\mathbf{A}^\mathsf{T}(\mathbf{y} - \mathbf{A}\mathbf{D}^{1/2}\theta^\ell).$$

Now let $\mathbf{u}^\ell = \mathbf{D}^{-1/2}(\theta^\ell - \theta^0)$ and $\mathbf{b} := \mathbf{y} - \mathbf{A}\mathbf{D}^{1/2}\theta^0$. Then,

$$\mathbf{u}^{\ell+1} = \mathbf{u}^\ell + \eta\mathbf{A}^\mathsf{T}(\mathbf{b} - \mathbf{A}\mathbf{D})\mathbf{u}^\ell = (\mathbf{I} - \eta\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{D})\mathbf{u}^\ell + \eta\mathbf{A}^\mathsf{T}\mathbf{b} \tag{59}$$

For $0 < \nu_W < 1$, we have that ρ_j satisfies the bound (22). Since $\mathbf{D}$ is a block diagonal matrix with entries ρ_j, $\|\mathbf{D}\| \leq \rho_{\max}$. Now select

$$B_1 := \frac{1}{\rho_{\max}\|\mathbf{A}\|^2}, \quad B_2 := \|\mathbf{A}^\mathsf{T}\mathbf{b}\|. \tag{60}$$

If we take $\eta < B_1$ then

$$\eta\mathbf{D}^{1/2}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{D}^{1/2} \leq \eta\rho_{\max}\|\mathbf{A}\|^2 \leq \mathbf{I} \Rightarrow \|\mathbf{I} - \eta\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{D}\| \leq 1.$$

Hence, (59) shows that

$$\|\mathbf{u}^{\ell+1}\|_F \leq \|\mathbf{u}^\ell\|_F + \eta B_2 \Rightarrow \|\mathbf{u}^\ell\|_F \leq \eta\ell B_2, \tag{61}$$

where we have used the fact that $\mathbf{u}^0 = \mathbf{0}$. Now, since $\mathbf{u}^\ell = \mathbf{D}^{-1/2}(\theta^\ell - \theta^0)$, the j-th component of θ^ℓ is

$$\theta^\ell_j = \theta^0_j + \sqrt{\rho_j}\mathbf{u}^\ell_j.$$

Hence,

$$L^\ell_{\mathrm{conv},j} = \sqrt{\rho_j}\theta^\ell_j = L^0_{\mathrm{conv},j} + \rho_j\mathbf{u}^\ell_j.$$

Applying (61) we obtain the bound on the convolutional model

$$\|L^\ell_{\mathrm{conv},j} - L^0_{\mathrm{conv},j}\|_F \leq \rho_j\eta\ell B_2. \tag{62}$$

Bounding the RNN Impulse Response From Theorems 3.2 and 3.3, the scaled convolutional model and linear RNN have the same NTK. Due to (58), they have the same input-output mapping at the initial conditions. Since the scaled convolutional model is linear in its parameters it follows that it is linear NTK model for the RNN. Therefore, using the NTK results such as Proposition 3.1, we have that for all input sequences $\mathbf{x}$ and GD time steps ℓ,

$$\lim_{n\to\infty} \left\| f_{\mathrm{RNN}}(\mathbf{x}, \theta^{\ell}_{\mathrm{RNN}}) - f_{\mathrm{conv}}(\mathbf{x}, \theta^{\ell}_{\mathrm{conv}}) \right\| = 0, \tag{63}$$

where the convergence is in probability. Thus, if we fix an input $\mathbf{x}$ and iteration ℓ and define

$$\mathbf{y}_{\mathrm{RNN}} = f_{\mathrm{RNN}}(\mathbf{x}, \theta^{\ell}_{\mathrm{RNN}}), \quad \mathbf{y}_{\mathrm{conv}} = f_{\mathrm{conv}}(\mathbf{x}, \theta^{\ell}_{\mathrm{conv}}),$$

the limit (63) can be re-written as

$$\lim_{n\to\infty} \| y_{\mathrm{RNN},j} - y_{\mathrm{conv},j} \| = 0, \tag{64}$$

for all j. Again, the convergence is in probability. Now consider the case where the input sequence $\mathbf{x} = (x_0, \ldots, x_{T-1})$ is a sequence with $x_j = 0$ for all $j > 0$ That is, it is only non-zero at the initial time step. Then for all time steps $y_{\mathrm{RNN},j} = L^{\ell}_{\mathrm{RNN},j} x_0$ and $y_{\mathrm{conv},j} = L^{\ell}_{\mathrm{conv},j} x_0$. Since this is true for all x_0, (64) shows that for all time steps $j = 0, \ldots, T-1$,

$$\lim_{n\to\infty} \left\| L_{\mathrm{RNN},j}(x, \theta^{\ell}_{\mathrm{RNN}}) - L_{\mathrm{conv},j}(x, \theta^{\ell}_{\mathrm{conv}}) \right\|_F = 0 \tag{65}$$

where the convergence is in probability. Combining (65) with (62) proves (25).

C. Recursions with Random Gaussians

We consider a recursion of the form,

$$q_{t+1} = \sum_{\ell=1}^{L} \frac{1}{\sqrt{n}} A_\ell \overline{G}_\ell(q_t, u_t), \tag{66}$$

where $q_t \in \mathbb{R}^{n \times d_q}$, $u_t \in \mathbb{R}^{n \times d_u}$, and $\overline{G}_\ell(q_t, u_t)$ acts row-wise, meaning

$$\overline{G}_\ell(q_t, u_t)_{i,:} = G_\ell(q_{t,i,:}, u_{t,i,:}), \tag{67}$$

for some Lipschitz functions $G_\ell : \mathbb{R}^{d_q} \times \mathbb{R}^{d_u} \times \to \mathbb{R}^{d_q}$. That is, the outptut of row i of $\overline{G}_\ell(\cdot)$ depends only the i-th rows of its inputs. We will analyze this system for a fixed horizon, $t = 0, \ldots, T-1$. Assume that

$$(q_0, u_0, \ldots, u_{T-1}) \overset{PL(2)}{\to} (Q_0, U_0, \ldots, U_{T-1}), \tag{68}$$

to random variables $(Q_0, U_0, \ldots, U_{T-1})$ where Q_0 is independent of $(U_0, \ldots, U_{T-1})$, and $Q_0 \sim \mathcal{N}(0, P_0)$ for some covariance matrix $P_0 \in \mathbb{R}^{d_q \times d_q}$. Assume the matrices $A_\ell \in \mathbb{R}^{n \times n}$ are independent with i.i.d. components, $(A_\ell)_{i,j} \sim \mathcal{N}(0, \nu_\ell)$.

Lemma C.1. *Under the above assumptions,*

$$(q_0, q_1, \ldots, q_{T-1}, u_0, u_1, \ldots, u_{T-1}) \overset{PL(2)}{\to} (Q_0, Q_1, \ldots, Q_{T-1}, U_0, \ldots U_{T-1}) \tag{69}$$

where each $Q_i \in \mathbb{R}^{d_q}$ and $(Q_0, Q_1, \ldots, Q_{T-1})$ are zero mean Gaussian processes independent of $(U_0, \ldots U_{T-1})$, generated recursively through SE equations given by

$$D_\ell = \mathcal{N}(0, \nu_\ell) \tag{70a}$$

$$Z_{t\ell} = G_\ell(Q_t, U_t) \tag{70b}$$

$$\mu_{t\ell} = \mathbb{E}(Z_{t\ell}), \quad \tilde{Z}_{t\ell} = Z_{t\ell} - \mu_{t\ell} \tag{70c}$$

$$F_{t,:,\ell} = \min_{F_1, \ldots, F_{t-1}} \mathbb{E} \left\| \tilde{Z}_{t\ell} - \sum_{j=1}^{t} \tilde{Z}_{t-j,\ell} F_j \right\|^2 \tag{70d}$$

$$P_{t\ell} = \mathbb{E}(\tilde{Z}_{t\ell} - \sum_{j=1}^{t} \tilde{Z}_{t-j,\ell} F_{tj\ell})^{\mathsf{T}} (\tilde{Z}_{t\ell} - \sum_{j=1}^{t} \tilde{Z}_{t-j,\ell} F_{tj\ell}) \tag{70e}$$

$$\tilde{R}_{t\ell} = \sum_{j=1}^{t} \tilde{R}_{t-j,\ell} F_{tj\ell} + \mathcal{N}(0, \nu_W P_{t\ell}), \tag{70f}$$

$$R_{t\ell} = \tilde{R}_{t\ell} + D_\ell \mu_{t\ell} \tag{70g}$$

$$Q_{t+1} = \sum_{\ell=1}^{L} R_{t\ell} \tag{70h}$$

Proof: We prove this by induction. Let $\mathcal{M}_t$ be the hypothesis that this result is true up to iteration t. We show that $\mathcal{M}_0$ is true and that $\mathcal{M}_t$ implies $\mathcal{M}_{t+1}$.

Base case ($\mathcal{M}_0$): Define $z_{0\ell} = \overline{G}_\ell(q_0, u_0)$. We have that rows of $z_{0\ell}$ converge $PL(2)$ to $Z_{0\ell} = \overline{G}_\ell(Q_0, U_0)$.

Now, let $\mu_{0\ell} = \mathbb{E}(Z_{0\ell})$ and define the following:

$$d_\ell = \frac{1}{\sqrt{n}} A_\ell \mathbf{1}, \quad \widetilde{z}_{0\ell} = z_{0\ell} - \mathbf{1}\mu_{0\ell} \tag{71}$$

$$\widetilde{r}_{1\ell} = \frac{1}{\sqrt{n}} A_\ell \widetilde{z}_{0\ell}, \quad r_{1\ell} = \widetilde{r}_{1\ell} + d_\ell \mu_{0\ell}, \quad q_1 = \sum_{\ell=1}^{L} r_{1\ell}. \tag{72}$$

We know that

$$d_\ell \overset{PL(2)}{=} D_\ell \sim \mathcal{N}(0, \nu_\ell), \qquad \widetilde{z}_{0\ell} \overset{PL(2)}{=} \widetilde{Z}_{0\ell} = Z_{0\ell} - \mu_{0\ell}. \tag{73}$$

Note that $\widetilde{Z}_{0\ell}$ are zero mean. Now since A_ℓ are i.i.d Gaussian matrices, rows of $\widetilde{r}_{1\ell}$ converge PL(2) to random variable

$$\widetilde{R}_{1\ell} \sim \mathcal{N}(0, P_{1\ell}) \qquad \text{where,} \quad P_{1\ell} = \lim_{n\to\infty} \frac{1}{n} \widetilde{z}_{0\ell}^\intercal \widetilde{z}_{0\ell} \overset{a.s.}{=} \mathbb{E}(\widetilde{Z}_{0\ell}^\intercal \widetilde{Z}_{0\ell}) \tag{74}$$

Furthermore, one can show that $\mathbb{E}(\widetilde{R}_{1\ell_1} \widetilde{R}_{1\ell_2}) = \mathbb{E}(\widetilde{Z}_{0\ell_1}^\intercal \widetilde{Z}_{0\ell_2}) = 0$ $\widetilde{R}_{1\ell_1}$ and $\widetilde{R}_{1\ell_2}$ are independent Therefore,

$$q_1 \overset{PL(2)}{=} Q_1 = \sum_{\ell=1}^{L} \left[\widetilde{R}_{1\ell} + D_\ell \mu_{0\ell}\right] \tag{75}$$

This proves $\mathcal{M}_0$ holds true.

Induction recursion: We next assume that the SE system is true up to iteration t. We write the recursions as

$$d_\ell = \frac{1}{\sqrt{n}} A_\ell \mathbf{1} \in \mathbb{R}^n \tag{76a}$$

$$z_{t\ell} = \overline{G}_\ell(q_t, u_t), \quad \widetilde{z}_{t\ell} = z_{t\ell} - \mathbf{1}\mu_{t\ell} \tag{76b}$$

$$\widetilde{r}_{t+1,\ell} = \frac{1}{\sqrt{n}} A_\ell \widetilde{z}_{t\ell}, \quad r_{t+1,\ell} = \widetilde{r}_{t+1,\ell} + d_\ell \mu_{t\ell}, \quad q_{t+1} = \sum_{\ell=1}^{L} r_{t+1,\ell}. \tag{76c}$$

The main issue in dealing with a recursion of the form Equation (76) is that for $t \geq 1$, matrices $\{A_\ell\}_{\ell=1}^L$ and $\{\widetilde{r}_{t\ell}\}_{\ell=1}^L$ are no longer independent. The key idea is to use a conditioning technique (Bolthausen conditioning) as in (Bayati & Montanari, 2011) to deal with this dependence. Instead of conditioning $\widetilde{r}_{t\ell}$ on A_ℓ, we condition A_ℓ on the event

$$\mathcal{E}_{t,\ell} = \{\widetilde{r}_{t'+1,\ell} = \frac{1}{\sqrt{n}} A_\ell \widetilde{z}_{t'\ell}, t' = 0, \ldots, t-1\}. \tag{77}$$

Note that this event is a set of linear constraints, and i.i.d. Gaussian random variables conditioned on linear constraints have Gaussian densities that we can track.

Let $\widetilde{\mathcal{H}}_{t\ell}$ be the linear operator

$$\widetilde{\mathcal{H}}_{t\ell} : A_\ell \mapsto (\widetilde{r}_{1\ell}, \ldots, \widetilde{r}_{t\ell}). \tag{78}$$

With these definitions, we have

$$A_\ell|_{\mathcal{E}_{t,\ell}} \overset{d}{=} \widetilde{\mathcal{H}}_{t\ell}^\dagger(\widetilde{r}_{1\ell}, \ldots, \widetilde{r}_{t\ell}) + \widetilde{\mathcal{H}}_{t\ell}^\perp(\widetilde{A}_\ell), \tag{79}$$

where $\widetilde{\mathcal{H}}_{t,\ell}^\dagger$ is the Moore-Penrose pseudo-inverse operator of $\widetilde{\mathcal{H}}_{t,\ell}$, $\widetilde{\mathcal{H}}_{t,\ell}^\perp$ is the orthogonal projection operator onto the subspace orthogonal to the kernel of $\widetilde{\mathcal{H}}_t$, and $\widetilde{A}_\ell$ is an independent copy of A_ℓ. Therefore, we can write $\widetilde{r}_{t+1,\ell}$ as sum of two terms

$$\widetilde{r}_{t+1,\ell} = \widetilde{r}_{t+1,\ell}^{\mathrm{det}} + \widetilde{r}_{t+1,\ell}^{\mathrm{ran}}, \tag{80}$$

where $\widetilde{r}_{t+1,\ell}^{\mathrm{det}}$ is what we call the deterministic part:

$$\widetilde{r}_{t+1,\ell}^{\mathrm{det}} = \frac{1}{\sqrt{n}} \widetilde{\mathcal{H}}_{t\ell}^\dagger(\widetilde{r}_1, \ldots, \widetilde{r}_t)\, \widetilde{z}_{t\ell} \tag{81}$$

and $\widetilde{r}_{t+1,\ell}^{\mathrm{ran}}$ is the random part:

$$\widetilde{r}_{t+1,\ell}^{\mathrm{ran}} = \frac{1}{\sqrt{n}} \widetilde{\mathcal{H}}_{t\ell}^\perp(\widetilde{A}_\ell)\, \widetilde{z}_{t\ell}. \tag{82}$$

It is helpful to write the linear operators defined in this section in matrix form for derivations that follow.

$$\widetilde{\mathcal{H}}_{t\ell}(A_\ell) = \frac{1}{\sqrt{n}} [A_\ell] \begin{bmatrix} \widetilde{z}_{0\ell} & \ldots & \widetilde{z}_{t-1,\ell} \end{bmatrix}. \tag{83}$$

Deterministic part: We first characterizes the limiting behavior of $\widetilde{r}^{\text{det}}_{t+1,\ell}$.

It is easy to show that if the functions $\overline{G}_\ell$ are non-constant, then the operator $\widetilde{\mathcal{H}}_{t\ell}\widetilde{\mathcal{H}}_{t\ell}^{\mathsf{T}}$ where $\widetilde{\mathcal{H}}_{t\ell}^{\mathsf{T}}$ is the adjoint of $\widetilde{\mathcal{H}}_{t\ell}$, is full-rank almost surely for any finite t. Thus, we have

$$\widetilde{\mathcal{H}}_{t\ell}^{\dagger} = \widetilde{\mathcal{H}}_{t\ell}^{\mathsf{T}}(\widetilde{\mathcal{H}}_{t\ell}\widetilde{\mathcal{H}}_{t\ell}^{\mathsf{T}})^{-1} \tag{84}$$

Form equation (78) we have

$$\widetilde{\mathcal{H}}_{t\ell}^{\mathsf{T}}(\widetilde{r}_{1\ell},\dots,\widetilde{r}_{t\ell}) = \frac{1}{\sqrt{n}}\sum_{t'=1}^{t}\widetilde{r}_{t'\ell}(\widetilde{z}_{t'-1,\ell})^{\mathsf{T}}. \tag{85}$$

Combining (85) and (78) we get

$$\left(\widetilde{\mathcal{H}}_{t\ell}(\widetilde{\mathcal{H}}_{t\ell}^{\mathsf{T}})(\widetilde{r}_{1\ell},\dots,\widetilde{r}_{t\ell})\right)_s = \frac{1}{n}\sum_{t'=1}^{t}\widetilde{r}_{t'\ell}\,(\widetilde{z}_{t'-1,\ell})^{\mathsf{T}}\widetilde{z}_{s-1,\ell} \tag{86}$$

Now, under the induction hypothesis, using the definition of PL(2) convergence we have

$$R_{\widetilde{Z}\ell}(t',s) := \lim_{n\to\infty}\frac{1}{n}(\widetilde{z}_{t'-1,\ell})^{\mathsf{T}}\widetilde{z}_{s-1,\ell} \overset{a.s.}{=} \mathbb{E}\left((\widetilde{Z}_{t'-1,\ell})^{\mathsf{T}}\widetilde{Z}_{s-1,\ell}\right) \tag{87}$$

Therefore we have,

$$\widetilde{\mathcal{H}}_{t\ell}\widetilde{\mathcal{H}}_{t\ell}^{\mathsf{T}}(\widetilde{r}_{1\ell},\dots,\widetilde{r}_{t\ell}) = [\widetilde{r}_{1\ell} \quad \dots \quad \widetilde{r}_{t\ell}]\underbrace{\begin{bmatrix} R_{\widetilde{Z}\ell}(0,0) & R_{\widetilde{Z}\ell}(0,1) & \dots & R_{\widetilde{Z}\ell}(0,t-1) \\ R_{\widetilde{Z}\ell}(1,0) & R_{\widetilde{Z}\ell}(1,1) & \dots & R_{\widetilde{Z}\ell}(1,t-1) \\ \vdots & \vdots & \ddots & \vdots \\ R_{\widetilde{Z}\ell}(t-1,0) & R_{\widetilde{Z}\ell}(t-1,1) & \dots & R_{\widetilde{Z}\ell}(t-1,t-1) \end{bmatrix}}_{\mathcal{R}_{\widetilde{Z}\ell}} \tag{88}$$

Let $\mathcal{R}^{-1}_{\widetilde{Z}\ell}$ denote the inverse of $\mathcal{R}_{\widetilde{Z}\ell}$ and index its blocks similarly to $\mathcal{R}_{\widetilde{Z}\ell}$. Then, the pseudo-inverse is

$$\widetilde{\mathcal{H}}_{t\ell}^{\dagger}(\widetilde{r}_{1\ell},\dots,\widetilde{r}_{t\ell}) = \frac{1}{\sqrt{n}}\sum_{t'=1}^{t}\sum_{t''=1}^{t}\widetilde{r}_{t''\ell}\mathcal{R}^{-1}_{\widetilde{Z}\ell}(t''-1,t'-1)(\widetilde{z}_{t'-1,\ell})^{\mathsf{T}} + o(\frac{1}{n}). \tag{89}$$

Define $\widetilde{P}_{t\ell} := \widetilde{Z}_{t\ell} - \sum_{j=1}^{t}\widetilde{Z}_{t-j,\ell}F_{tj\ell},$, where $F_{t,:,\ell}$ are defined in (70d). Using equation (81) we get:

$$\widetilde{r}^{\text{det}}_{t+1,\ell} = \frac{1}{n}\sum_{t''=1}^{t}\widetilde{r}_{t''\ell}\sum_{t'=1}^{t}\mathcal{R}^{-1}_{\widetilde{Z}\ell}(t''-1,t'-1)(\widetilde{z}_{t'-1,\ell})^{\mathsf{T}}\widetilde{z}_{t,\ell} + o(\frac{1}{n}) \tag{90a}$$

$$\overset{a.s.}{=}\sum_{t''=1}^{t}\widetilde{r}_{t''\ell}\sum_{t'=1}^{t}\mathcal{R}^{-1}_{\widetilde{Z}\ell}(t''-1,t'-1)\,\mathbb{E}\left((\widetilde{Z}_{t'-1,\ell})^{\mathsf{T}}\widetilde{Z}_{t,\ell}\right) + o(\frac{1}{n}) \tag{90b}$$

$$= \sum_{t''=1}^{t}\widetilde{r}_{t''\ell}\sum_{t'=1}^{t}\mathcal{R}^{-1}_{\widetilde{Z}\ell}(t''-1,t'-1)\,\mathbb{E}\left((\widetilde{Z}_{t'-1,\ell})^{\mathsf{T}}(\widetilde{P}_{t\ell} + \sum_{j=1}^{t}\widetilde{Z}_{t-j,\ell}F_{t,j,\ell})\right) + o(\frac{1}{n}) \tag{90c}$$

$$= \sum_{t''=1}^{t}\widetilde{r}_{t''\ell}\sum_{j=1}^{t}\underbrace{\sum_{t'=1}^{t}\mathcal{R}^{-1}_{\widetilde{Z}\ell}(t''-1,t'-1)\,\mathcal{R}_{\widetilde{Z}\ell}(t'-1,t-j)}_{I\delta(t''=t-j+1)}F_{t,j,\ell} + o(\frac{1}{n}) \tag{90d}$$

$$= \sum_{j=1}^{t}\widetilde{r}_{t-j+1,\ell}F_{t,j,\ell} + o(\frac{1}{n}), \tag{90e}$$

where (a) follows from the fact that $\mathbb{E}(\widetilde{Z}^{\mathsf{T}}_{t'\ell}\widetilde{P}_{t\ell}) = 0$ for $t' = 0,\dots,t-1$. Now by induction hypothesis we know that $\widetilde{r}_{t-j+1,\ell} \overset{PL(2)}{=} \widetilde{R}_{t-j+1,\ell}$, therefore,

$$\widetilde{r}^{\text{det}}_{t+1,\ell} \overset{PL(2)}{=} \widetilde{R}^{\text{det}}_{t+1,\ell} = \sum_{j=1}^{t}\widetilde{R}_{t-j+1,\ell}F_{t,j,\ell} \tag{91}$$

Random part We next consider the random part:

$$\widetilde{r}^{\mathrm{ran}}_{t+1,\ell} = \frac{1}{\sqrt{n}} \widetilde{\mathcal{H}}^{\perp}_{t\ell}(\widetilde{A}_\ell)\widetilde{z}_{t\ell} \tag{92}$$

$$= \frac{1}{\sqrt{n}}(\widetilde{A}_\ell \widetilde{z}_{t\ell} - \widetilde{\mathcal{H}}^{\dagger}_{t\ell}\widetilde{\mathcal{H}}_{t\ell}(\widetilde{A}_\ell)\widetilde{z}_{t\ell}). \tag{93}$$

We know that,

$$\widetilde{\mathcal{H}}^{\dagger}_{t}\widetilde{\mathcal{H}}_{t}(\widetilde{A}_\ell) = \frac{1}{n}\sum_{t'=1}^{t}\sum_{t''=1}^{t} \widetilde{A}_\ell \widetilde{z}_{t''-1,\ell}\mathcal{R}^{-1}_{\widetilde{Z}\ell}(t''-1,t'-1)(\widetilde{z}_{t'-1,\ell})^{\intercal} + o(\frac{1}{n}). \tag{94}$$

Then, we have

$$\widetilde{r}^{\mathrm{ran}}_{t+1,\ell} = \frac{1}{\sqrt{n}}\widetilde{A}_\ell \widetilde{z}_{t\ell} - \frac{1}{\sqrt{n}}\sum_{t'=1}^{t}\sum_{t''=1}^{t} \widetilde{A}_\ell \widetilde{z}_{t''-1,\ell}\, \mathcal{R}^{-1}_{\widetilde{Z}\ell}(t''-1,t'-1)\left(\frac{1}{n}(\widetilde{z}_{t'-1,\ell})^{\intercal}\widetilde{z}_{t\ell}\right) + o(\frac{1}{n}) \tag{95}$$

$$= \frac{1}{\sqrt{n}}\widetilde{A}_\ell \widetilde{z}_{t\ell} - \frac{1}{\sqrt{n}}\sum_{t''=1}^{t} \widetilde{A}_\ell \widetilde{z}_{t''-1,\ell}\ \sum_{j=1}^{t}\sum_{t'=1}^{t}\mathcal{R}^{-1}_{\widetilde{Z}\ell}(t''-1,t'-1)\,\mathcal{R}_{\widetilde{Z}\ell}(t'-1,t-j)F_{t,j,\ell} + o(\frac{1}{n}) \tag{96}$$

$$= \frac{1}{\sqrt{n}}\widetilde{A}_\ell(\widetilde{z}_{t\ell} - \sum_{j=1}^{t}\widetilde{z}_{t-j,\ell}F_{t,j,\ell}) + o(\frac{1}{n}) \tag{97}$$

Therefore, since $\widetilde{A}_\ell$ are i.i.d. Gaussian matrices, $\widetilde{r}^{\mathrm{ran}}_{t+1,\ell}$ converges PL(2) to a Gaussian random variable $\widetilde{R}^{\mathrm{ran}}_{t+1,\ell} \sim \mathcal{N}(0, P_{t+1,\ell})$ such that,

$$P_{t+1,\ell} = \mathbb{E}(\widetilde{Z}_{t\ell} - \sum_{j=1}^{t}\widetilde{Z}_{t-j,\ell}F_{tj\ell})^{\intercal}(\widetilde{Z}_{t\ell} - \sum_{j=1}^{t}\widetilde{Z}_{t-j,\ell}F_{tj\ell}) \tag{98}$$

We can now write $\widetilde{R}_{t+1,\ell}$ as,

$$\widetilde{R}_{t+1,\ell} = \widetilde{R}^{\mathrm{det}}_{t+1,\ell} + \widetilde{R}^{\mathrm{ran}}_{t+1,\ell} \tag{99}$$

$$= \sum_{j=1}^{t}\widetilde{R}_{t-j+1,\ell}F_{t,j,\ell} + \mathcal{N}(0, P_{t+1,\ell}), \tag{100}$$

and by equation (76) we have

$$Q_{t+1} = \sum_{\ell=1}^{L} R_{t+1,\ell}, \qquad R_{t+1,\ell} = \widetilde{R}_{t+1,\ell} + D_\ell \mu_{t\ell}.$$

This proves $\mathcal{M}_t$ implies $\mathcal{M}_{t+1}$.