STOCHASTIC GRADIENT DESCENT LEARNS STATE EQUATIONS WITH NONLINEAR ACTIVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study discrete time dynamical systems governed by the state equation $h_{t+1} = \phi(Ah_t + Bu_t)$. Here $A, B$ are weight matrices, $\phi$ is an activation function, and $u_t$ is the input data. This relation is the backbone of recurrent neural networks (e.g. LSTMs) which have broad applications in sequential learning tasks. We utilize stochastic gradient descent to learn the weight matrices from a finite input/state trajectory $\{u_t, h_t\}_{t=0}^{N}$. We prove that SGD estimate linearly converges to the ground truth weights while using near-optimal sample size. Our results apply to increasing activations whose derivatives are bounded away from zero. The analysis is based on i) an SGD convergence result with nonlinear activations and ii) careful statistical characterization of the state vector. Numerical experiments verify the fast convergence of SGD on ReLU and leaky ReLU in consistence with our theory.

1 INTRODUCTION

A wide range of problems involve sequential data with a natural temporal ordering. Examples include natural language processing, time series prediction, system identification, and control design, among others. State-of-the-art algorithms for sequential problems often stem from dynamical systems theory and are tailored to learn from temporally dependent data. Linear models and algorithms; such as Kalman filter, PID controller, and linear dynamical systems, have a long history and are utilized in control theory since 1960’s with great success [Brown et al. (1992); Ho & Kalman (1966); Åström & Hägglund (1995)]. More recently, nonlinear models such as recurrent neural networks (RNN) found applications in complex tasks such as machine translation and speech recognition [Bahdanau et al. (2014); Graves et al. (2013); Hochreiter & Schmidhuber (1997)]. Unlike feedforward neural networks, RNNs are dynamical systems that use their internal state to process inputs. The goal of this work is to shed light on the inner workings of RNNs from a theoretical point of view. In particular, we focus on the RNN state equation which is characterized by a nonlinear activation function $\phi$, state weight matrix $A$, and input weight matrix $B$ as follows

$$h_{t+1} = \phi(Ah_t + Bu_t),$$

(1.1)

Here $h_t$ is the state vector and $u_t$ is the input data at timestamp $t$. This equation is the source of dynamic behavior of RNNs and distinguishes RNN from feedforward networks. The weight matrices $A$ and $B$ govern the dynamics of the state equation and are inferred from data. We will explore the statistical and computational efficiency of stochastic gradient descent (SGD) for learning these weight matrices.

Contributions: Suppose we are given a finite trajectory of input/state pairs $\{(u_t, h_t)\}_{t=0}^{N}$ generated from the state equation (1.1). We consider a least-squares regression obtained from $N$ equations; with inputs $\{(u_t, h_t)\}_{t=1}^{N}$ and outputs $\{(h_{t+1})\}_{t=0}^{N}$. For a class of activation functions including leaky ReLU and for stable systems, we show that SGD linearly converges to the ground truth weight matrices while requiring near-optimal trajectory length $N$. In particular, the required sample size is $O(n + p)$ where $n$ and $p$ are the dimensions of the state and input vectors respectively. Our results are extended to unstable systems when the samples are collected from multiple independent RNN trajectories rather than a single trajectory. Our results apply to increasing activation functions whose derivatives are bounded away from zero; which includes leaky ReLU. Numerical experiments on

1Throughout this work, a system is called stable if the spectral norm of the state matrix $A$ is less than 1.
ReLU and leaky ReLU corroborate our theory and demonstrate that SGD converges faster as the activation slope increases. To obtain our results, we i) characterize the statistical properties of the state vector (e.g. well-conditioned covariance) and ii) derive a novel SGD convergence result with nonlinear activations; which may be of independent interest. As a whole, this paper provides a step towards foundational understanding of RNN training via SGD.

1.1 Related Work

Our work is related to the recent optimization and statistics literature on linear dynamical systems (LDS) and neural networks.

**Linear dynamical systems:** The state-equation (1.1) reduces to a LDS when $\phi$ is the linear activation ($\phi(x) = x$). Identifying the weight matrices is a core problem in linear system identification and is related to the optimal control problem (e.g. linear quadratic regulator) with unknown system dynamics. While these problems are studied since 1950’s (Ljung (1998; 1987); Åström & Eykhoff (1971)), our work is closer to the recent literature that provides data dependent bounds and characterize the non-asymptotic learning performance. Recht and coauthors have a series of papers exploring optimal control problem (Simchowitz et al. (2018); Tu et al. (2018; 2017)). In particular, Hardt et al. (2016) shows gradient descent learns single-input-single-output (SISO) LDS with polynomial guarantees. Oymak & Ozay (2018) and Faradonbeh et al. (2018) provide sample complexity bounds for learning LDS. Sanandaji et al. (2011b;a) study the identification of sparse systems.

**Neural networks:** There is a growing literature on the theoretical aspects of deep learning and provable algorithms for training neural networks. Most of the existing results are concerned with feedforward networks. Ge et al. (2018); Li & Yuan (2017); Mei et al. (2018); Soltanolkotabi (2017); Janzamin et al. (2015); Soltanolkotabi et al. (2017); Zhong et al. (2017b) consider learning fully-connected shallow networks with gradient descent. Brutzkus & Globerson (2017); Zhong et al. (2017a); Du et al. (2017) address convolutional neural networks; which is an efficient weight-sharing architecture. Brutzkus et al. (2017); Wang et al. (2018) studies over-parameterized networks when data is linearly separable. Janzamin et al. (2015); Oymak & Soltanolkotabi (2018) utilize tensor decomposition techniques for learning feedforward neural nets. For recurrent networks, Sedghi & Anandkumar (2016) proposed tensor algorithms with polynomial guarantees and Khrulkov et al. (2017) studied their expressive power. More recently, Miller & Hardt (2018) showed that stable RNNs can be approximated by feed-forward networks.

2 Problem Setup

We first introduce the notation. $\|\cdot\|$ returns the spectral norm of a matrix and $s_{\text{min}}(\cdot)$ returns the minimum singular value. The activation $\phi : \mathbb{R} \rightarrow \mathbb{R}$ applies entry-wise if its input is a vector. Throughout, $\phi$ is assumed to be a 1-Lipschitz function. With proper scaling of its parameters, the system (1.1) with a Lipschitz activation can be transformed into a system with 1-Lipschitz activation. The functions $\Sigma[\cdot]$ and $\text{var}[\cdot]$ return the covariance of a random vector and variance of a random variable respectively. $I_n$ is the identity matrix of size $n \times n$. Normal distribution with mean $\mu$ and covariance $\Sigma$ is denoted by $\mathcal{N}(\mu, \Sigma)$. Throughout, $c, C, c_0, c_1, \ldots$ denote positive absolute constants.

**Setup:** We consider the dynamical system parametrized by an activation function $\phi(\cdot)$ and weight matrices $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times p}$ as described in (1.1). Here, $h_t$ is the $n$ dimensional state-vector and $u_t$ is the $p$ dimensional input to the system at time $t$. As mentioned previously, (1.1) corresponds to the state equation of a recurrent neural network. For most RNNs of interest, the state $h_t$ is hidden and we only get to interact with $h_t$ via an additional output equation. For Elman networks (Elman 1990), this equation is characterized by some output activation $\phi_y$ and output weights $C, D$ as follows

$$y_t = \phi_y(CH_t + Du_t).$$

(2.1)

In this work, our attention is restricted to the state equation (1.1); which corresponds to setting $y_t = h_t$ in the output equation. To analyze (1.1) in a non-asymptotic data-dependent setup, we assume a finite input/state trajectory of $\{u_t, h_t\}_{t=0}^T$ generated by some ground truth weight matrices $(A, B)$. Our goal is learning the unknown weights $A$ and $B$ in a data and computationally efficient
With this reparameterization, we find the input/output identity

We formulate a regression problem by defining the output vector $y_t$, input vector $x_t$, and the target parameter $C$ as follows

$$y_t = h_{t+1}, \quad x_t = \begin{bmatrix} \mu h_t \\ u_t \end{bmatrix} \in \mathbb{R}^{n+p}, \quad C = [\mu^{-1} A B] \in \mathbb{R}^{n \times (n+p)}. \tag{2.2}$$

With this reparameterization, we find the input/output identity $y_t = \phi(C x_t)$. We will consider the least-squares regression given by

$$L(\Theta) = \frac{1}{N} \sum_{t=1}^{N} L_t(\Theta) \quad \text{where} \quad L_t(\Theta) = \frac{1}{2} \|y_t - \phi(\Theta x_t)\|^2_{L_2}. \tag{2.3}$$

For learning the ground truth parameter $C$, we utilize SGD on the loss function (2.3) with a constant learning rate $\eta$. Starting from an initial point $\Theta_0$, after END SGD iterations, Algorithm 1 returns an estimate $C = \Theta_{\text{END}}$. Estimates of $A$ and $B$ are decoded from the left and right submatrices of $C$ respectively.

### 3 Main Results

#### 3.1 Preliminaries

The analysis of the state equation naturally depends on the choice of the activation function; which is the source of nonlinearity. We first define a class of Lipschitz and increasing activation functions.

**Definition 3.1 ($\beta$-increasing activation).** Given $1 \geq \beta \geq 0$, the activation function $\phi$ satisfies $\phi(0) = 0$ and $1 \geq \phi'(x) \geq \beta$ for all $x \in \mathbb{R}$.

Our results will apply to strictly increasing activations where $\phi$ is $\beta$-increasing for some $\beta > 0$. Observe that, this excludes ReLU activation which has zero derivative for negative values. However, it includes Leaky ReLU which is a generalization of ReLU. Parameterized by $1 \geq \beta \geq 0$, Leaky ReLU is a $\beta$-increasing function given by

$$\text{LReLU}(x) = \max(\beta x, x). \tag{3.1}$$

In general, given an increasing and 1-Lipschitz activation $\phi$, a $\beta$-increasing function $\phi_\beta$ can be obtained by blending $\phi$ with the linear activation, i.e. $\phi_\beta(x) = (1 - \beta) \phi(x) + \beta x$.

A critical property that enables SGD is that the state-vector covariance $\Sigma[h_t]$ is well-conditioned under proper assumptions. The lemma below provides upper and lower bounds on this covariance matrix in terms of problem variables.

**Lemma 3.2 (State vector covariance).** Consider the state equation (1.1) where $h_0 = 0$ and $u_t \sim N(0, I_p)$. Define the upper bound term $B_t$ as

\begin{algorithm}[H]
\caption{Learning state equations with nonlinear activations}
1: **Inputs:** $(y_t, h_t, u_t)_{t=1}^{N}$ sampled from a trajectory. Scaling $\mu$, learning rate $\eta$. Initialization $A_0, B_0$.
2: **Outputs:** Estimates $\hat{A}, \hat{B}$ of the weight matrices $A, B$.
3: $x_t \leftarrow [\mu h_t^T u_t^T]^T$ for $1 \leq t \leq N$.
4: $\Theta_0 \leftarrow [\mu^{-1} A_0 B_0]$.
5: for $r$ from 1 to END do
6: \hspace{1em} Pick $\gamma_r$ from $\{1, 2, \ldots, N\}$ uniformly at random.
7: \hspace{1em} $\Theta_r \leftarrow \Theta_{r-1} - \eta \nabla \mathcal{L} \left( \Theta_{r-1} \right)$
8: end for
9: return $[\hat{A} \hat{B}] \leftarrow \Theta_{\text{END}} \left[ \begin{array}{cc} \mu I_n & 0 \\ 0 & I_p \end{array} \right]$.
\end{algorithm}
Additionally, assume that $\phi$ is a $\beta$-increasing function and $p \geq n$. Then, $\Sigma[h_t] \geq \beta^2 s_{\min}(B)^2 I_n$.

As a natural extension from linear dynamical systems, we will say the system is stable if $\|A\| < 1$ and unstable otherwise. For activations we consider, stability implies that if the input is set to 0, state vector $h_t$ will exponentially converge to 0 i.e. the system forgets the past states quickly. This is also the reason $(B_t)_{t \geq 0}$ sequence converges for stable systems and diverges otherwise. The condition number of the covariance will play a critical role in our analysis. Using Lemma 3.2, this number can be upper bounded by $\rho$ defined as

$$\rho = \left( \frac{B_\infty}{\beta s_{\min}(B)} \right)^2 = \left( \frac{\|B\|}{s_{\min}(B)} \right)^2 \frac{1}{\beta^2(1 - \|A\|^2)}.$$

Observe that, the condition number of $B$ appears inside the $\rho$ term.

### 3.2 Learning from Single Trajectory

Our main result applies to stable systems ($\|A\| < 1$) and provides a non-asymptotic convergence guarantee for SGD in terms of the upper bound on the state vector covariance. This result characterizes the sample complexity and the rate of convergence of SGD; and also provides insights into the role of activation function and the spectral norm of $A$.

**Theorem 3.3 (Main result).** Let $\{u_t, h_{t+1}\}_{t=1}^N$ be a finite trajectory generated from the state equation (1.1). Suppose $\|A\| < 1$, $\phi$ is $\beta$-increasing, $h_0 = 0$, $p \geq n$, and $u_t \sim N(0, I_p)$. Let $\rho$ be same as (3.3) and $c, C, c_0$ be properly chosen absolute constants. Pick the trajectory length $N$ to satisfy

$$N \geq CL\rho^2(n + p),$$

where $L = 1 - \log(c_0)/\log\|A\|$. Pick scaling $\mu = 1/B_\infty$, learning rate $\eta = c_0 \beta^2/\log(n + p)$, and consider the loss function (2.3). With probability $1 - 4N \exp(-100n) - 8L \exp(-O(L^2))$, starting from an initial point $\Theta_0$, for all $\tau \geq 0$, the SGD iterations described in Algorithm 1 satisfies

$$\mathbb{E}[\|\Theta_\tau - C\|^2_F] \leq (1 - c_0 \beta^4/2\rho^2(n + p))^{\tau} \|\Theta_0 - C\|^2_F.$$

The expectation is over the randomness of the SGD updates.

**Sample complexity:** Theorem 3.3 essentially requires $N \gtrsim (n + p)/\beta^4$ samples for learning. This can be seen by unpacking (3.3) and ignoring the logarithmic $L$ term and the condition number of $B$. Observe that $O(n + p)$ growth achieves near-optimal sample size for our problem. Each state equation (1.1) consists of $n$ sub-equations (one for each entry of $h_{t+1}$). We collect $N$ state equations to obtain a system of $Nn$ equations. On the other hand, the total number of unknown parameters in $A$ and $B$ are $(n + p)$. This implies Theorem 3.3 is applicable as soon as the problem is mildly overdetermined i.e. $Nn \gtrsim (n + p)$.

**Computational complexity:** Theorem 3.3 requires $O(n(n + p) \log \rho)$ iterations to reach $\varepsilon$-neighborhood of the ground truth. Our analysis reveals that, this rate can be accelerated if the state vector is zero-mean. This happens for odd activation functions satisfying $\phi(-x) = -\phi(x)$ (e.g. linear activation). The result below is a corollary and requires $\times n$ less iterations.

**Theorem 3.4 (Faster learning for odd activations).** Consider the same setup provided in Theorem 3.3. Additionally, assume that $\phi$ is an odd function. Pick scaling $\mu = 1/B_\infty$, learning rate $\eta = c_0 \beta^2/\log(n + p)$, and consider the loss function (2.3). With probability $1 - 4N \exp(-100n) - 8L \exp(-O(L^2))$, starting from an initial point $\Theta_0$, for all $\tau \geq 0$, the SGD iterations described in Algorithm 1 satisfies

$$\mathbb{E}[\|\Theta_\tau - C\|^2_F] \leq (1 - c_0 \beta^4/2\rho^2(n + p))^{\tau} \|\Theta_0 - C\|^2_F,$$

where the expectation is over the randomness of the SGD updates.
Another aspect of the convergence rate is the dependence on $\beta$. In terms of $\beta$, the SGD error (3.4) decays as $(1 - O(\beta^8))^{\tau}$. While it is not clear how optimal is the exponent 8, numerical experiments in Section 6 demonstrate that larger $\beta$ indeed results in drastically faster convergence.

4 MAIN IDEAS AND PROOF STRATEGY

To prove the results of the previous section, we derive a deterministic result that establishes the linear convergence of SGD for $\beta$-increasing functions. For linear convergence proofs, a typical strategy is showing the strong convexity of the loss function i.e. showing that, for some $\alpha > 0$ and all points $v, u$, the gradient satisfies

$$\langle \nabla \mathcal{L}(v) - \nabla \mathcal{L}(u), v - u \rangle \geq \alpha \|v - u\|_2^2.$$ 

The core idea of our convergence result is that the strong convexity parameter of the loss function with $\beta$-increasing activations can be connected to the loss function with linear activations. In particular, recalling (2.3), set $y_t^{lin} = Cx_t$ and define the linear loss to be

$$\mathcal{L}^{lin}(\Theta) = \frac{1}{2N} \sum_{i=1}^{N} \|y_t^{lin} - \Theta x_i\|_2^2.$$ 

Denoting the strong convexity parameter of the original loss by $\alpha_\phi$ and that of linear loss by $\alpha_{lin}$, we argue that $\alpha_\phi \geq \beta^2 \alpha_{lin}$; which allows us to establish a convergence result as soon as $\alpha_{lin}$ is strictly positive. Next result is our SGD convergence theorem which follows from this discussion.

**Theorem 4.1** (Deterministic convergence). Suppose a data set $\{x_i, y_i\}_{i=1}^{N}$ is given; where output $y_i$ is related to input $x_i$ via $y_i = \phi((x_i, \theta))$ for some $\theta \in \mathbb{R}^n$. Suppose $\beta > 0$ and $\phi$ is a $\beta$-increasing. Let $\gamma_+ \geq \gamma_- > 0$ be scalars. Assume that input samples satisfy the bounds

$$\gamma_+ I_n \geq \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T \geq \gamma_- I_n \quad \text{and} \quad \|x_i\|_2^2 \leq B < \infty \quad \text{for all } i.$$ 

Let $\{r_\tau\}_{\tau=0}^{\infty}$ be a sequence of i.i.d. integers uniformly distributed between 1 to $N$. Then, starting from an arbitrary point $\theta_0$, setting learning rate $\eta = \frac{\beta^2 \gamma_+}{\gamma_+ B}$, for all $\tau \geq 0$, the SGD iterations for quadratic loss

$$\theta_{\tau+1} = \theta_\tau - \eta(\phi(x_{r_\tau}^T, \theta_\tau) - y_{r_\tau})\phi'(x_{r_\tau}^T, \theta_\tau)x_{r_\tau}, \quad (4.1)$$

satisfies the error bound

$$\mathbb{E}[\|\theta_\tau - \theta\|_2^2] \leq \|\theta_0 - \theta\|_2^2 (1 - \frac{\beta^4 \gamma_+^2}{\gamma_+ B})^{\tau}, \quad (4.2)$$

where the expectation is over the random selection of the SGD iterations $\{r_\tau\}_{\tau=0}^{\infty}$. 

This theorem provides a clean convergence rate for SGD for $\beta$-increasing activations and naturally generalizes standard results on linear regression which corresponds to $\beta = 1$. Its extension to proximal gradient methods might be beneficial for high-dimensional nonlinear problems (e.g. sparse/low-rank approximation and generalized linear models Cai et al. (2010), Beck & Teboulle (2009), Oymak et al. (2018), Agarwal et al. (2010)) and is left as a future work.

To derive the results from Section 3 we need to determine the conditions under which Theorem 4.1 is applicable to the data obtained from RNN state equation with high probability. Below we provide desirable characteristics of the state vector; which enables our statistical results.

**Assumption 1** (Well-behaved state vector). Let $L > 1$ be an integer. There exists positive scalars $\gamma_+, \gamma_- \cdot \theta$ and an absolute constant $C > 0$ such that $\theta \leq \gamma_+ n$ and the following holds

- **Lower bound**: $\Sigma[h_{L-1}] \geq \gamma_- I_n$.
- **Upper bound**: for all $t$, the state vector satisfies

$$\Sigma[h_t] \leq \gamma_+ I_n \quad \text{and} \quad \|h_t - \mathbb{E}[h_t]\|_{\psi_2} \leq C \sqrt{\gamma_+} \quad \text{and} \quad \|\mathbb{E}[h_t]\|_{\psi_2} \leq \theta \sqrt{\gamma_+}. \quad (4.3)$$

Here $\|\cdot\|_{\psi_2}$ returns the subgaussian norm of a vector (see Def. 5.22 of Vershynin (2010)).

Under review as a conference paper at ICLR 2019
Assumption\(^1\) ensures that covariance is well-conditioned, state vector is well-concentrated, and it has a reasonably small expectation. Our next theorem establishes statistical guarantees for learning the RNN state equation based on this assumption.

**Theorem 4.2** (General result). Let \( \{u_i, h_{i+1}\}_{i=1}^N \) be a length \( N \) trajectory of the state equation \((1.1)\). Suppose \( \|A\| < 1, \) is \( \beta \)-increasing, \( h_0 = 0, \) and \( u_i \overset{i.i.d.}{\sim} N(0, I_p). \) Given scalars \( \gamma_+ \geq \gamma_- > 0, \) set the condition number as \( \rho = \gamma_+/\gamma_- \). For absolute constants \( C, c, c_0 > 0, \) choose trajectory length \( N \) to satisfy

\[
N \geq CL\rho^2(n + p) \quad \text{where} \quad L = \left[ 1 + \frac{\log(c_0\rho)}{\log \|A\|} \right].
\]

Suppose Assumption\(^1\) holds with \( L, \gamma_+, \gamma_-, \theta. \) Pick scaling to be \( \mu = 1 / \sqrt{\gamma} + \) and learning rate to be \( \eta = c_0 \frac{\beta^4}{2\rho^2(\theta + \sqrt{2})^2(n + p)} \). With probability \( 1 - 4N \exp(-100\eta) - 8L \exp(-O(\frac{N}{\lambda^2})) \), starting from \( \Theta_0, \) for all \( \tau \geq 0, \) the SGD iterations on loss \((2.3)\) as described in Algorithm\(^1\) satisfies

\[
E[\|\Theta_\tau - C\|_F^2] \leq (1 - c_0 \frac{\beta^4}{2\rho^2(\theta + \sqrt{2})^2(n + p)})^\tau \|\Theta_0 - C\|_F^2,
\]

where the expectation is over the randomness of SGD updates.

The advantage of this theorem is that, it isolates the optimization problem from the statistical properties of state vector. If one can prove tighter bounds on achievable \((\gamma_+, \gamma_-, \theta), \) it will immediately imply improved performance for SGD. In particular, Theorems 3.3 and 4.2 are simple corollaries of Theorem 4.2 with proper choices.

- Theorem 3.3 follows by setting \( \gamma_+ = B_{\infty}^2, \gamma_- = \beta^2 s_{\min}(B)^2, \) and \( \theta = \sqrt{n}. \)
- Theorem 4.2 follows by setting \( \gamma_+ = B_{\infty}^2, \gamma_- = \beta^2 s_{\min}(B)^2, \) and \( \theta = 0. \)

5 Learning Unstable Systems

So far, we considered learning from a single RNN trajectory for stable systems (\( \|A\| < 1 \)). For such systems, as the time goes on, the impact of the earlier states disappear. In our analysis, this allows us to split a single trajectory into multiple nearly-independent trajectories. This approach will not work for unstable systems (\( A \) is arbitrary) where the impact of older states may be amplified over time. To address this, we consider a model where the data is sampled from multiple independent trajectories.

Suppose \( N \) independent trajectories of the state-equation \((1.1)\) are available. Pick some integer \( T_0 \geq 1. \) Denoting the \( i \)th trajectory by the triple \((h_{t_0}^{(i)}, h_{t_1}^{(i)}, u_{t_0}^{(i)}); 0 \leq t \leq T_0 \) we collect a single sample from each trajectory at time \( T_0 \) to obtain the triple \((h_{T_0}^{(i)}, h_{T_0 + 1}^{(i)}, u_{T_0}^{(i)}). \) To utilize the existing optimization framework \((2.3)\), for \( 1 \leq i \leq N, \) we set

\[
(y_i, h_i, u_i) = (h_{t_0}^{(i)}, h_{t_0 + 1}^{(i)}, u_{t_0}^{(i)}).
\]

With this setup, we can again use the SGD Algorithm\(^1\) to learn the weights \( A \) and \( B. \) The crucial difference compared to Section 3 is that, the samples \((y_i, h_i, u_i); 1 \leq i \leq N, \) are now independent of each other; hence, the analysis is simplified. As previously, having an upper bound on the condition number of the state-vector covariance is critical. This upper bound can be shown to be \( \rho \) defined as

\[
\rho = \begin{cases} 
\tilde{\rho} & \text{if } n > 1 \\
\frac{\rho_{\min}^2}{\rho_{\min}^2 - \|A\|^2} & \text{if } n = 1
\end{cases}
\]

where \( \tilde{\rho} = \frac{B_{T_0}^2}{\beta^2 s_{\min}(B)^2}. \)

The \( \tilde{\rho} \) term is similar to the earlier definition \((5.3)\); however it involves \( B_{T_0} \) rather than \( B_{\infty}. \) This modification is indeed necessary since \( B_{\infty} = \infty \) when \( \|A\| > 1. \) On the other hand, note that, \( B_{T_0}^2 \) grows proportional to \( \|A\|^{2T_0} \); which results in exponentially bad condition number in \( T_0. \) Our \( \rho \) definition remedies this issue for single-output systems; where \( n = 1 \) and \( A \) is a scalar. In particular, when \( 1 < \beta \) (e.g. \( \phi \) is linear) \( \rho \) becomes equal to the correct value \((5.3)\). The next theorem provides our result on unstable systems in terms of this condition number and other model parameters.

\(^2\)Clearly, any nonzero \( 1 \times 1 \) covariance matrix has condition number \( 1. \) However, due to subleties in the proof strategy, we don’t use \( \rho = 1 \) for \( \beta < 1. \) Obtaining tighter bounds on the subgaussian norm of the state-vector would help resolve this issue.
We conducted experiments on ReLU and Leaky ReLU activations. Let us first describe the experimental setup. We pick the state dimension $n = 50$ and the input dimension $p = 100$. We choose the ground truth matrix $A$ to be a scaled random unitary matrix; which ensures that all singular values of $A$ are equal. $B$ is generated with i.i.d. $N(0, 1)$ entries. Instead of using the theoretical scaling choice, we determine the scaling $\mu$ from empirical covariance matrices outlined in Algorithm 2. Similar to our proof strategy, this algorithm equalizes the spectral norms of the input and state covariances to speed up convergence. We also empirically determined the learning rate and used $\eta = 1/100$ in all experiments.

Evaluation: We consider two performance measures in the experiments. Let $\hat{C}$ be an estimate of the ground truth parameter $C = [\mu^{-1} A \ B]$. The first measure is the normalized error defined as $\|\hat{C} - C\|_F^2 / \|C\|_F^2$. The second measure is the normalized loss defined as

$$\frac{\sum_{i=1}^N \|y_i - \phi(\hat{C} x_i)\|_2^2}{\sum_{i=1}^N \|y_i\|_2^2}.$$  

In all experiments, we run Algorithm 1 for 50000 SGD iterations and plot these measures as a function of $\tau$; by using the estimate available at the end of the $\tau$th SGD iteration for $0 \leq \tau \leq 50000$. Each curve is obtained by averaging the outcomes of 20 independent realizations.

Algorithm 2 Empirical hyperparameter selection.

1: Inputs: $(h_i, u_i)_{i=1}^N$ sampled from a trajectory.
2: Outputs: Scaling $\mu$.
3: Form the empirical covariance matrix $\Sigma_h$ from $(h_i)_{i=1}^N$.
4: Form the empirical covariance matrix $\Sigma_u$ from $(u_i)_{i=1}^N$.
5: return $\sqrt{\|\Sigma_u\|/\|\Sigma_h\|}$.  

Figure 1: SGD convergence behavior for Leaky ReLUs with varying minimum slope $\beta$. Figures a) and b) repeat the same experiments. The difference is the spectral norm of the ground truth state matrix $A$.

**Theorem 5.1** (Unstable systems). Suppose we are given $N$ independent trajectories $(h_i^{(i)}, u_i^{(i)})_{i \geq 0}$ for $1 \leq i \leq N$. Each trajectory is sampled at time $T_0$ to obtain $N$ samples $(y_i, h_i, u_i)_{i=1}^N$ where the $i$th sample is given by (5.1). Suppose the sample size satisfies

$$N \geq C \rho^2 (n + p),$$

where $\rho$ is given by (5.2). Assume the initial states are 0, $\phi$ is $\beta$-increasing, $p \geq n$, and $u_i \sim \mathcal{N}(0, I_p)$. Set scaling $\mu = 1 / \sqrt{B_{T_0}}$, learning rate $\eta = c_0 \rho^2 (n + p)$, and run SGD over the equations described in (2.2) and (2.3). Starting from $\Theta_0$, with probability $1 - 2N \exp(-100(n + p)) - 4 \exp(-\mathcal{O}(\frac{n^2}{\sqrt{p}}))$, all SGD iterations satisfy

$$\mathbb{E}[\|\Theta_T - C\|_F^2] \leq (1 - c_0 \beta^4 \frac{\rho^2 n (n + p)}{2})^\gamma \|\Theta_0 - C\|_F^2,$$

where the expectation is over the randomness of the SGD updates.

6 Numerical Experiments

We conducted experiments on ReLU and Leaky ReLU activations. Let us first describe the experimental setup. We pick the state dimension $n = 50$ and the input dimension $p = 100$. We choose the ground truth matrix $A$ to be a scaled random unitary matrix; which ensures that all singular values of $A$ are equal. $B$ is generated with i.i.d. $N(0, 1)$ entries. Instead of using the theoretical scaling choice, we determine the scaling $\mu$ from empirical covariance matrices outlined in Algorithm 2. Similar to our proof strategy, this algorithm equalizes the spectral norms of the input and state covariances to speed up convergence. We also empirically determined the learning rate and used $\eta = 1/100$ in all experiments.

Evaluation: We consider two performance measures in the experiments. Let $\hat{C}$ be an estimate of the ground truth parameter $C = [\mu^{-1} A \ B]$. The first measure is the normalized error defined as $\|\hat{C} - C\|_F^2 / \|C\|_F^2$. The second measure is the normalized loss defined as

$$\frac{\sum_{i=1}^N \|y_i - \phi(\hat{C} x_i)\|_2^2}{\sum_{i=1}^N \|y_i\|_2^2}.$$  

In all experiments, we run Algorithm 1 for 50000 SGD iterations and plot these measures as a function of $\tau$; by using the estimate available at the end of the $\tau$th SGD iteration for $0 \leq \tau \leq 50000$. Each curve is obtained by averaging the outcomes of 20 independent realizations.
weight matrices and an activation function. This problem is of interest for recurrent neural networks
(2016). What can be said when we add the nonlinear activations?

This work showed that SGD can learn the nonlinear dynamical system (1.1); which is characterized by
linear convergence and as $\beta$ increases, the rate of convergence drastically improves. The improvement
is more visible for less stable systems driven by $\mathbf{A}$ with a larger spectral norm. In particular, while
ReLU converges for small $\|\mathbf{A}\|$, SGD gets stuck before reaching the ground truth when $\|\mathbf{A}\| = 0.8$.

To understand, how well SGD fits the training data, in Figure 2a we plotted the normalized loss
for ReLU activation. For more unstable system ($\|\mathbf{A}\| = 0.9$), training loss stagnates in a similar
fashion to the parameter error. We also verified that the norm of the overall gradient $\|\nabla \mathcal{L}(\Theta_r)\|_F$
continues to decay (where $\Theta_r$ is the $r$th SGD iterate); which implies that SGD converges before
reaching a global minima. As $\mathbf{A}$ becomes more stable, rate of convergence improves and linear rate
is visible. Finally, to better understand the population landscape of the quadratic loss with ReLU
activations, Figure 2b repeats the same ReLU experiments while increasing the sample size five
times to $N = 2500$. For this more overdetermined problem, SGD converges even for $\|\mathbf{A}\| = 0.9$;
indicating that
- population landscape of loss with ReLU activation is well-behaved,
- however ReLU problem requires more data compared to the Leaky ReLU for finding global
  minima.

Overall, as predicted by our theory, experiments verify that SGD indeed quickly finds the optimal
weight matrices of the state equation (1.1) and as the activation slope $\beta$ increases, the convergence
rate improves.

7 CONCLUSIONS

This work showed that SGD can learn the nonlinear dynamical system (1.1); which is characterized by
weight matrices and an activation function. This problem is of interest for recurrent neural networks
as well as nonlinear system identification. We showed that efficient learning is possible with optimal
sample complexity and good computational performance. Our results apply to strictly increasing
activations such as Leaky ReLU. We empirically showed that Leaky ReLU converges faster than
ReLU and requires less samples; in consistence with our theory. We list a few unanswered problems
that would provide further insights into recurrent neural networks.

- **Covariance of the state-vector:** Our results depend on the covariance of the state-vector and
  requires it to be positive definite. One might be able to improve the current bounds on the condition
  number and relax the assumptions on the activation function. Deriving similar performance bounds
  for ReLU is particularly interesting.

- **Hidden state:** For RNNs, the state vector is hidden and is observed through an additional equation
  (2.1); which further complicates the optimization landscape. Even for linear dynamical systems,
  learning the $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ system ((1.1), (2.1)) is a non-trivial task [20]
  what can be said when we add the nonlinear activations?

- **Classification task:** In this work, we used normally distributed input and least-squares regression
  for our theoretical guarantees. More realistic input distributions might provide better insight into
  contemporary problems, such as natural language processing; where the goal is closer to classification
  (e.g. finding the best translation from another language).
REFERENCES


A Deterministic Convergence Result for SGD

Proof of Theorem 4.1. Given two distinct scalars $a, b$; define $\phi'(a, b) = \frac{\phi(a) - \phi(b)}{a - b}$. Observe $\phi'(a, b) \geq \beta$ since $\phi$ is $\beta$-increasing. Define $w_r$ to be the residual $w_r = \theta_r - \theta$. Observing $w_r = \phi'(x_r^T \theta_r, x_r^T \theta) x_r^T w_r$, the SGD recursion obeys

$$||w_{r+1}||^2_2 = ||w_r - \eta w_r + \tau_r||^2_2 = \frac{\tau_r + 2}{2} \cdot ||w_r||^2_2 + \frac{\tau_r - 2}{2} \cdot E[||w_r + \eta \mathbf{G}_r w_r||^2_2]$$

where $\mathbf{G}_r = \phi'(x_r^T \theta_r, x_r^T \theta) x_r^T$. Using the bound on the rows we find that

$$E[||w_r + \eta \mathbf{G}_r w_r||^2_2] \leq \frac{\beta^2 \gamma}{\gamma_t + B} ||w_r||^2_2$$

Consequently, we find the following bounds in expectation

$$\gamma_t I_n \geq E[\mathbf{G}_r] \geq \beta \gamma I_n,$$

and $\gamma_t \gamma_n I_n \geq E[\mathbf{G}_r^T \mathbf{G}_r] \leq B \gamma_t I_n$.

Observe that (A.1) essentially lower bounds the strong convexity parameter of the problem with $\beta^2 \gamma$; which is the strong convexity of the identical problem with the linear activation (i.e. $\beta = 1$). However, we only consider strong convexity around the ground truth parameter $\theta$ i.e. we restricted our attention to $(\theta, \theta_r)$ pairs. With this, $w_{r+1}$ can be controlled as,$\frac{\tau_r + 2}{2} E[||w_r||^2_2] + \frac{\tau_r - 2}{2} E[||w_r + \mathbf{G}_r w_r||^2_2] \leq \frac{\beta^2 \gamma}{\gamma_t + B} ||w_r||^2_2 (1 - \frac{\beta^2 \gamma}{\gamma_t + B})$.

Lemma A.1 (Merging L splits). Assume matrices $X^{(i)} \in \mathbb{R}^{N_i \times q}$ are given for $1 \leq i \leq L$. Suppose for all $1 \leq i \leq L$, rows of $X^{(i)}$ has $\ell_2$ norm at most $\sqrt{B}$ and each $X^{(i)}$ satisfies

$$\gamma_t I_n \geq \frac{X^{(i)^T} X^{(i)}}{N_i} \geq \gamma I_n.$$ 

Set $N = \sum_{i=1}^L N_i$ and form the concatenated matrix $X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \\ \vdots \\ X^{(L)} \end{bmatrix}$. Denote $i$th row of $X$ by $x_i$. Then, for each $i$, $||x_i||^2_2 \leq B$ and

$$\gamma_t I_n \geq \frac{X^T X}{N} \geq \frac{1}{N} \sum_{i=1}^N x_i x_i^T \geq \gamma I_n.$$ 

Proof. The bound on the rows $||x_i||^2_2$ directly follows by assumption. For the remaining result, first observe that $X^T X = \sum_{i=1}^L X^{(i)^T} X^{(i)}$. Next, we have $N \gamma_t I_n = \sum_{i=1}^L N_i \gamma_t I_n \geq \sum_{i=1}^L X^{(i)^T} X^{(i)} \geq \sum_{i=1}^L N_i \gamma I_n = N \gamma I_n$.

Combining these two yields the desired upper/lower bounds on $X^T X / N$. \hfill \square
B PROPERTIES OF THE NONLINEAR STATE EQUATIONS

This section characterizes the properties of the state vector $h_t$ when input sequence is normally distributed. These bounds will be crucial for obtaining upper and lower bounds for the singular values of the data matrix $X = [x_1 \ldots x_N]^T$ described in (2.2). For probabilistic arguments, we will use the properties of subgaussian random variables. Orlicz norm provides a general framework that subsumes subgaussianity.

**Definition B.1 (Orlicz norms).** For a scalar random variable Orlicz-a norm is defined as

$$\|X\|_\psi_a = \sup_{k \geq 1} k^{-1/a}(\mathbb{E}[|X|^k])^{1/k}$$

Orlicz-a norm of a vector $x \in \mathbb{R}^p$ is defined as $\|x\|_\psi_a = \sup_{v \in \mathbb{R}^p} |v^T x|_\psi_a$ where $B^p$ is the unit $\ell_2$ ball. The subexponential norm is the Orlicz-1 norm $\|\cdot\|_{\psi_1}$ and the subgaussian norm is the Orlicz-2 norm $\|\cdot\|_{\psi_2}$.

**Lemma B.2 (Lipschitz properties of the state vector).** Consider the state equation governed by equation (1.1). Suppose activation $\phi$ is 1-Lipschitz. Observe that $h_{t+1}$ is a deterministic function of the input sequence $\{u_{\tau}\}_{\tau=0}^t$. Fixing all vectors $\{u_i\}_{i=\tau}^{t+1}$ (i.e. all except $u_{\tau}$), $h_{t+1}$ is $|A|^{t-\tau}|B|$ Lipschitz function of $u_{\tau}$ for $0 \leq \tau \leq t$.

**Proof.** Fixing $\{u_i\}_{i=\tau}^{t+1}$, denote $h_{t+1}$ as a function of $u_{\tau}$ by $h_{t+1}(u_{\tau})$. Given a pair of vectors $u_{\tau}, u_{\tau}'$ using 1-Lipschitzness of $\phi$, for any $t > \tau$, we have

$$\|h_{t+1}(u_{\tau}) - h_{t+1}(u_{\tau}')\|_{\ell_2} \leq \|\phi(\mathbf{A}h_t(u_{\tau}) + B u_{\tau}) - \phi(\mathbf{A}h_t(u_{\tau}') + B u_{\tau}')\|_{\ell_2}$$

$$\leq \|\mathbf{A}(h_t(u_{\tau}) - h_t(u_{\tau}'))\|_{\ell_2}$$

$$\leq \|\mathbf{A}\| \|h_t(u_{\tau}) - h_t(u_{\tau}')\|_{\ell_2}.$$  

Proceeding with this recursion until $t = \tau$, we find

$$\|h_{t+1}(u_{\tau}) - h_{t+1}(u_{\tau}')\|_{\ell_2} \leq \|\mathbf{A}\|^{t-\tau} \|h_{t+1}(u_{\tau}) - h_{t+1}(u_{\tau}')\|_{\ell_2}$$

$$\leq \|\mathbf{A}\|^{t-\tau} \|\mathbf{A}^T\| \|\mathbf{A} h_t - \mathbf{A} h_t\|_{\ell_2}$$

$$\leq \|\mathbf{A}\|^{t-\tau} \|\mathbf{A}\| \|\mathbf{A} h_t - \mathbf{A} h_t\|_{\ell_2}.$$  

This bound implies $h_{t+1}(u_{\tau})$ is $|A|^{t-\tau}|B|$ Lipschitz function of $u_{\tau}$.  

**Lemma B.3 (Upper bound).** Consider the state equation governed by equation (1.1). Suppose $\{u_t\}_{t=0}^{T} \sim \mathcal{N}(0, I_p)$, $\phi$ is 1-Lipschitz, $\phi(0) = 0$ and $h_0 = 0$. Recall the definition of $B_i$. We have the following properties

- $h_t$ is a $B_t$-Lipschitz function of the vector $q_t = [u_0^T \ldots u_{t-1}^T]^T \in \mathbb{R}^{tp}$.
- There exists an absolute constant $c > 0$ such that $\|h_t - \mathbb{E}[h_t]\|_{\psi_2} \leq cB_t$ and $\Sigma[h_t] \preceq B_t^2 I_n$.
- $h_t$ satisfies

$$\mathbb{E}[\|h_t\|^2_{\ell_2}] \leq \text{tr}(BB_t^T) \frac{1 - \|A\|_{\ell_2}^2}{1 - \|A\|_{\ell_2}^2} \leq \min\{n, p\} B_t^2.$$  

Also, there exists an absolute constant $c > 0$ such that for any $m \geq n$, with probability $1 - 2 \exp(-100m)$, $\|h_t\|_{\ell_2} \leq c\sqrt{\ln B_t}$.

**Proof.** i) Bounding Lipschitz constant: Observe that $h_t$ is a deterministic function of $q_t$: i.e. $h_t = f(q_t)$ for some function $f$. To bound Lipschitz constant of $f$, for all (deterministic) vector pairs $q_t$ and $\hat{q}_t$, we find a scalar $L_f$ satisfying

$$\|f(q_t) - f(\hat{q}_t)\|_{\ell_2} \leq L_f \|q_t - \hat{q}_t\|_{\ell_2}. \quad (B.1)$$

Define the vectors, $\{a_i\}_{i=0}^t$, as follows

$$a_t = [u_0^T \ldots u_{t-1}^T u_t^T \ldots u_{t-1}^T]^T.$$  

Observing that $a_0 = q_t$, $a_t = \hat{q}_t$, we write the telescopic sum

$$\|f(q_t) - f(\hat{q}_t)\|_{\ell_2} \leq \sum_{i=0}^{t-1} \|f(a_{i+1}) - f(a_i)\|_{\ell_2}.$$  

Focusing on the individual terms $f(a_{i+1}) - f(a_i)$, observe that the only difference is the $u_i, \hat{u}_i$ terms. Viewing $h_t$ as a function of $u_t$ and applying Lemma B.2,

$$\|f(a_{i+1}) - f(a_i)\|_{\ell_2} \leq \|A\|^{t-1-i} |B| \|u_i - \hat{u}_i\|_{\ell_2}. \quad (B.2)$$
To bound the sum, we apply the Cauchy-Schwarz inequality, which yields

\[
|f(q_t) - f(\hat{q}_t)| \leq \sum_{i=0}^{t-1} \|A\|^{1-\ell} \|B\| \|u_i - \hat{u}_i\|_{\ell_2}
\]

\[
\leq \left( \sum_{i=0}^{t-1} \|A\|^{2(1-\ell)} \|B\|^2 \right)^{1/2} \left( \sum_{i=0}^{t-1} \|u_i - \hat{u}_i\|_{\ell_2}^2 \right)^{1/2}
\]

\[
\leq \sqrt{\|B\|^2 (1 - \|A\|^2)} \|q_t - \hat{q}_t\|_{\ell_2}
\]

\[
= B_t \|q_t - \hat{q}_t\|_{\ell_2}.
\]

(B.2)

The final line achieves the inequality (B.1) with \(L_f = B_t\) hence \(h_t\) is \(B_t\) Lipschitz function of \(q_t\).

ii) Bounding subgaussian norm: When \(u_t \sim \mathcal{N}(0, I_p)\), the vector \(q_t\) is distributed as \(\mathcal{N}(0, I_p)\). Since \(h_t\) is a \(B_t\) Lipschitz function of \(q_t\), for any fixed unit length vector \(v\), \(\alpha_v := v^T h_t = v^T f(q_t)\) is still \(B_t\)-Lipschitz function of \(q_t\). Hence, using Gaussian concentration of Lipschitz functions, \(\alpha_v\) satisfies

\[
P(|\alpha_v - E[\alpha_v]| \geq t) \leq 2 \exp(-\frac{t^2}{2B_t^2}).
\]

This implies that for any \(v\), \(\alpha_v - E[\alpha_v] = \mathcal{O}(B_t)\) subgaussian [Vershynin (2010)]. This is true for all unit \(v\), hence using Definition B.1 the vector \(h_t\) satisfies \(\|h_t - E[h_t]\|_{\ell_2} \leq \mathcal{O}(B_t)\) as well. Secondly, \(B_t\)-Lipschitz function of a Gaussian vector obeys the variance inequality \(\text{var}[\alpha_v] \leq B_t^2\) (page 49 of Ledoux [2001]), which implies the covariance bound

\[
\Sigma[h_t] \leq B_t^2 I_n.
\]

iii) Bounding \(\ell_2\)-norm: To obtain this result, we first bound \(E[\|h_t\|_{\ell_2}^2]\). Since \(\phi\) is 1-Lipschitz and \(\phi(0) = 0\), we have the deterministic relation

\[
\|h_t+1\|_{\ell_2} \leq \|Ah_t + Bu_t\|_{\ell_2}.
\]

Taking squares of both sides, expanding the right hand side, and using the independence of \(h_t, u_t\) and the covariance information of \(u_t\), we obtain

\[
E[\|h_{t+1}\|_{\ell_2}^2] \leq E[\|Ah_t + Bu_t\|_{\ell_2}^2] = E[\|Ah_t\|_{\ell_2}^2] + E[\|Bu_t\|_{\ell_2}^2]
\]

(B.3)

\[
\leq \|A\|^2 E[\|h_t\|_{\ell_2}^2] + \text{tr}(BB^T).
\]

(B.4)

Now that the recursion is established, expanding \(h_t\) on the right hand side until \(h_1 = 0\), we obtain

\[
E[\|h_{t+1}\|_{\ell_2}^2] \leq \sum_{i=0}^t \|A\|^2 \text{tr}(BB^T) \leq \text{tr}(BB^T) \frac{1 - \|A\|^{2(t+1)}}{1 - \|A\|^2}.
\]

Now using the fact that \(\text{tr}(BB^T) \leq \text{rank}(B)\|B\|^2 \leq \min\{n, p\}\|B\|^2\), we find

\[
E[\|h_{t+1}\|_{\ell_2}^2] \leq \text{tr}(BB^T) \leq \min\{n, p\}B_{t+1}^2.
\]

Finally, using the fact that \(h_t\) is \(B_t\)-Lipschitz function and utilizing Gaussian concentration of \(q_t \sim \mathcal{N}(0, I_p)\), we find

\[
P(\|h_{t+1}\|_{\ell_2}^2 - E[\|h_{t+1}\|_{\ell_2}^2] \geq t) \leq \exp(-\frac{t^2}{2B_t^2}).
\]

Setting \(t = (c - 1)\sqrt{m}B_t\) for sufficiently large \(c > 0\), we find \(P(\|h_{t}\|_{\ell_2} \geq \sqrt{m}B_t + (c - 1)\sqrt{m}B_t) \leq \exp(-100m)\).

Lemma B.4 (Odd activations). Suppose \(\phi\) is strictly increasing and obeys \(\phi(x) = -\phi(-x)\) for all \(x\) and \(h_0 = 0\). Consider the state equation (1) driven \(u_t \sim \mathcal{N}(0, I_p)\). We have that \(E[h_t] = 0\).

Proof. We will inductively show that \(\{h_t\}_{t \geq 0}\) has a symmetric distribution around 0. Suppose the vector \(h_t\) satisfies this assumption. Let \(S \subset \mathbb{R}^n\) be a set. We will argue that \(P(h_{t+1} \subset S) = P(h_{t+1} \subset -S)\). Since \(\phi\) is strictly increasing, it is bijective on vectors, and we can define the unique inverse set \(S' = \phi^{-1}(S)\). Also since \(\phi\) is odd, \(\phi(-S') = -S\). Since \(h_t, u_t\) are independent and symmetric, we reach the desired conclusion as follows

\[
P(h_{t+1} \subset S) = P(Ah_t + Bu_t \subset S') = P(A(-h_t) + B(-u_t) \subset S')
\]

(B.5)

\[
= P(Ah_t + Bu_t \subset -S') = P(\phi(Ah_t + Bu_t) \subset \phi(-S')) = P(h_{t+1} \subset -S).
\]

(B.6)
Theorem B.5 (State-vector lower bound). Consider the nonlinear state equation \( \frac{dh_t}{dt} = Ah_t + Bu_t \) with \( \{u_t\}_{t \geq 0} \overset{i.i.d.}{\sim} \mathcal{N}(0, I_n) \). Suppose \( \phi \) is a \( \beta \)-increasing function for some constant \( \beta > 0 \). For any \( t \geq 1 \), the state vector obeys

\[
\Sigma[h_t] \geq \beta^2 s_{\min}(BB^T)I_n.
\]

Proof. The proof is an application of Lemma B.7. The main idea is to write \( h_t \) as sum of two independent vectors, one of which has independent entries. Consider a multivariate Gaussian vector \( g \sim \mathcal{N}(0, \Sigma) \). \( g \) is statistically identical to \( g_1 + g_2 \) where \( g_1 \sim \mathcal{N}(0, s_{\min}(\Sigma)I_d) \) and \( g_2 \sim \mathcal{N}(0, \Sigma - s_{\min}(\Sigma)I_d) \) are independent multivariate Gaussians.

Since \( Bu_t \sim \mathcal{N}(0, BB^T) \), setting \( \Sigma = BB^T \) and \( s_{\min} = s_{\min}(\Sigma) \), we have that \( Bu_t \sim g_1 + g_2 \) where \( g_1, g_2 \) are independent and \( g_1 \sim \mathcal{N}(0, s_{\min}I_n) \) and \( g_2 \sim \mathcal{N}(0, \Sigma - s_{\min}I_n) \). Consequently, we may write

\[
Bu_t + Ah_t \sim g_1 + g_2 + Ah_t.
\]

For lower bound, the crucial component will be the \( g_1 \) term; which has i.i.d. entries. Applying Lemma B.7 by setting \( x = g_1 \), \( y = g_2 + Ah_t \), and using the fact that \( h_t, g_1, g_2 \) are all independent of each other, we find the advertised bound, for all \( t \geq 0 \), via

\[
\Sigma[h_{t+1}] = \Sigma[\phi(g_1 + g_2 + Ah_t)] \geq \beta^2 s_{\min}I_n.
\]

The next theorem applies to multiple-input-single-output (MISO) systems where \( A \) is a scalar and \( B \) is a row vector. The goal is refining the lower bound of Theorem B.5.

Theorem B.6 (MISO lower bound). Consider the setup of Theorem B.5 with single output i.e. \( n = 1 \). For any \( t \geq 1 \), the state vector obeys

\[
\text{var}[h_t] \geq \beta^2 \|B\|_2^2 \frac{1 - (|A|)^{2t}}{1 - \beta^2 |A|^2}.
\]

Proof. For any random variable \( X \), applying Lemma B.7 we have \( \text{var}[\phi(X)] \geq \beta^2 \text{var}[X] \). Recursively, this yields

\[
\text{var}[h_{t+1}] = \text{var}[\phi(Ah_t + Bu_t)] \geq \beta^2 \text{var}[Ah_t + Bu_t] = \beta^2 (|A|^2 \text{var}[h_t] + \|B\|_2^2).
\]

Expanding these inequalities till \( h_0 \), we obtain the desired bound

\[
\text{var}[h_t] \geq \sum_{i=1}^{t} (|A|)^{i-1} \|B\|_2^2.
\]

Lemma B.7 (Vector lower bound). Suppose \( \phi \) is a \( \beta \)-increasing function. Let \( x = [x_1 \ldots x_n]^T \) be a vector with i.i.d. entries distributed as \( x_i \sim X \). Let \( y \) be a random vector independent of \( x \). Then,

\[
\Sigma[\phi(x + y)] \geq \beta^2 \text{var}[X]I_n.
\]

Proof. We first apply law of total covariance (e.g. Lemma B.8) to simplify the problem using the following lower bound based on the independence of \( x \) and \( y \).

\[
\Sigma[\phi(x + y)] \geq \mathbb{E}_y[\Sigma[\phi(x + y) \mid y]] = \mathbb{E}_y[\Sigma_x[\phi(x + y)]] (B.7)
\]

Now, focusing on the covariance \( \Sigma_x[\phi(x + y)] \), fixing a realization of \( y \), and using the fact that \( x \) has i.i.d. entries; \( \phi(x + y) \) has independent entries as \( \phi \) applies entry-wise. This implies that \( \Sigma_x[\phi(x + y)] \) is a diagonal matrix. Consequently, its lowest eigenvalue is the minimum variance over all entries,

\[
\Sigma_x[\phi(x + y)] \geq \min_{1 \leq i \leq n} \text{var}[\phi(x_i + y_i)]I_n.
\]

Fortunately, Lemma B.9 provides the lower bound \( \text{var}[\phi(x_i + y_i)] \geq \beta^2 \text{var}[X] \). Since this lower bound holds for any fixed realization of \( y \), it still holds after taking expectation over \( y \); which concludes the proof.

The next two lemmas are helper results for Lemma B.7 and are provided for the sake of completeness.

Lemma B.8 (Law of total covariance). Let \( x, y \) be two random vectors and assume \( y \) has finite covariance. Then

\[
\Sigma[y] = \mathbb{E}[\Sigma(y \mid x)] + \Sigma[\mathbb{E}[y \mid x]].
\]
Proof. First, write $\Sigma[y] = \mathbb{E}[yy^T] - \mathbb{E}[y]\mathbb{E}[y^T]$. Then, applying the law of total expectation to each term, 
\[ \Sigma[y] = \mathbb{E}[\mathbb{E}[yy^T \mid x]] - \mathbb{E}[\mathbb{E}[y \mid x]]\mathbb{E}[\mathbb{E}[y^T \mid x]]. \]

Next, we can write the conditional expectation as $\mathbb{E}[\mathbb{E}[yy^T \mid x]] = \mathbb{E}[(1^T)\Sigma (1)] = \mathbb{E}[y]\mathbb{E}[y^T]$. To conclude, we obtain the covariance of $\mathbb{E}[y \mid x]$ via the difference,
\[ \mathbb{E}[\mathbb{E}[y \mid x]]\mathbb{E}[\mathbb{E}[y^T \mid x]] = \Sigma[\mathbb{E}[y \mid x]], \]

which yields the desired bound. □

Lemma B.9 (Scalar lower bound). Suppose $\phi$ is a $\beta$-increasing function with $\beta > 0$ as defined in Definition 3.1. Given a random variable $X$ and a scalar $y$, we have
\[ \text{var}[\phi(X + y)] \geq \beta^2 \text{var}[X]. \]

Proof. Since $\phi$ is $\beta$-increasing, it is invertible and $\phi^{-1}$ is strictly increasing. Additionally, $\phi^{-1}$ is $1/\beta$ Lipschitz since,
\[ |\phi(a) - \phi(b)| \geq \beta|a - b| \implies |a - b| \geq \beta|\phi^{-1}(a) - \phi^{-1}(b)|. \]

Using this observation and the fact that $\mathbb{E}[X]$ minimizes $\mathbb{E}(X - \alpha)^2$ over $\alpha$, $\text{var}[\phi(X + y)]$ can be lower bounded as follows
\[
\text{var}[\phi(X + y)] = \mathbb{E}[(\phi(X + y) - \mathbb{E}[\phi(X + y)])^2]
\geq \beta^2 \mathbb{E}((X + y) - \phi^{-1}(\mathbb{E}[\phi(X + y)]))^2
\geq \beta^2 \mathbb{E}(X + y - \mathbb{E}[X + y])^2
= \beta^2 \mathbb{E}(X - \mathbb{E}[X])^2 = \beta^2 \text{var}[X].
\]

Note that, the final line is the desired conclusion. □

C Truncating Stable Systems

One of the challenges in analyzing dynamical systems is the fact that samples from the same trajectory have temporal dependence. This section shows that, for stable systems, the impact of the past states decay exponentially fast and the system can be approximated by using the recent inputs only. We first define the truncation of the state vector.

Definition C.1 (Truncated state vector). Suppose $\phi(0) = 0$, initial condition $h_0 = 0$, and consider the state equation (1.1). Given a timestamp $t$, $L$-truncation of the state vector $h_t$ is denoted by $h_{t, L}$ and is equal to $q_t$ where
\[ q_{t+1} = \phi(Aq_t + Bu_t) \quad , \quad q_0 = 0 \quad \text{(C.1)} \]
is the state vector generated by the inputs $u_t$ satisfying
\[ u_t = \begin{cases} 0 & \text{if } t < L, \\ u_t & \text{else} \end{cases}. \]

In words, $L$ truncated state vector $h_{t, L}$ is obtained by unrolling $h_t$ until time $t - L$ and setting the contribution of the state vector $h_{t-L}$ to 0. This way, $h_{t, L}$ depends only on the variables $\{u_{t-L}\}_{t-L \leq t \leq t}$. The following lemma states that impact of truncation can be made fairly small for stable systems ($\|A\| < 1$).

Lemma C.2 (Truncation impact – deterministic). Consider the state vector $h_t$ and its $L$-truncation $h_{t, L}$ from Definition C.1. Suppose $\phi$ is $1$-Lipschitz. We have that
\[ \|h_t - h_{t, L}\|_{\ell_2} \leq \begin{cases} 0 & \text{if } t \leq L, \\ \|A\|^{t-L} \|h_{t-L}\|_{\ell_2} & \text{else} \end{cases}. \]

Proof. When $t \leq L$, Definition C.1 implies $u_t = u_L$ hence $h_t = q_t = h_{t, L}$. When $t > L$, we again use Definition C.1 and recall that $u_t = 0$ until time $\tau = t - L - 1$. For all $t - L < \tau \leq t$, using 1-Lipschitzness of $\phi$, we have that
\[
\|h_{\tau} - q_{\tau}\|_{\ell_2} = \|\phi(Ah_{\tau-1} + Bu_{\tau-1}) - \phi(Aq_{\tau-1} + Bu_{\tau-1})\|_{\ell_2}
\leq \|(Ah_{\tau-1} + Bu_{\tau-1}) - (Aq_{\tau-1} + Bu_{\tau-1})\|_{\ell_2}
\leq \|A(h_{\tau-1} - q_{\tau-1})\|_{\ell_2} \leq \|A\| \|h_{\tau-1} - q_{\tau-1}\|_{\ell_2}.
\]
Applying this recursion between $t - L < \tau \leq t$ and using the fact that $q_{t-L} = 0$ implies the advertised result
\[
\|h_t - q_t\|_{\ell_2} \leq \|A\|^{t-L} \|h_{t-L} - q_{t-L}\|_{\ell_2} 
\leq \|A\|^{L} \|h_{t-L}\|_{\ell_2}. \]

□
C.1 Near Independence of Sub-trajectories

We will now argue that, for stable systems, a single trajectory can be split into multiple nearly independent trajectories. First, we describe how the sub-trajectories are constructed.

**Definition C.3** (Sub-trajectory). Let sampling rate \( L \geq 1 \) and offset \( 1 \leq \tau \leq L \) be two integers. Let \( N = \bar{N}_\tau \) be the largest integer obeying \( (N - 1)L + \tau \leq N \). We sample the trajectory \( \{h_t, u_t\}_{t=0}^N \) at the points \( \tau, \tau + L, \ldots, \tau + (N - 1)L + \tau \) and define the \( \tau \)th sub-trajectory as

\[
(h^{(i)}, u^{(i)}) := (h^{(i,0)}, u^{(i,0)}) = (h_{(i-1)L+\tau}, u_{(i-1)L+\tau})
\]

**Definition C.4** (Truncated sub-trajectory). Consider the state equation \( \text{(1.1)} \) and recall Definition C.3. Given offset \( \tau \) and sampling rate \( L \), for \( 1 \leq i \leq \bar{N} \), the \( i \)th truncated sub-trajectory states are \( \{\bar{h}^{(i)}\}_{i=1}^\bar{N} \) where the \( i \)th state is defined as

\[
\bar{h}^{(i)} = h_{(i-1)L+\tau,L-1}.
\]

The truncated samples are independent of each other as shown in the next lemma.

**Lemma C.5.** Consider the truncated states of Definition C.3. If \( \{h_t\}_{t=0}^N \) is generated by independent vectors \( \{u_t\}_{t=0}^N \), for any offset \( \tau \) and sampling rate \( L \), the vectors \( \{\bar{h}^{(i)}\}_{i=1}^\bar{N} \), \( \{\bar{u}^{(i)}\}_{i=1}^\bar{N} \) are all independent of each other.

**Proof.** By construction \( \bar{h}^{(i)} \) only depends on the vectors \( \{u_t\}_{t=L(i-1)+\tau}^{L(i-1)+\tau-1} \). Note that the dependence ranges \( [L(i - 2) + \tau + 1, L(i - 1) + \tau - 1] \) are disjoint intervals for different \( i \)'s; hence \( \{\bar{h}^{(i)}\}_{i=1}^\bar{N} \) are independent of each other. To show the independence of \( \bar{u}^{(i)} \) and \( \bar{h}^{(i)} \), observe that inputs \( u^{(i)} \) and \( h^{(i)} \) have timestamp \( \tau \) modulo \( L \), which is not covered by the dependence range of \( \{\bar{h}^{(i)}\}_{i=1}^\bar{N} \).

If the input is randomly generated, Lemma C.2 can be combined with a probabilistic bound on \( h_t \), to show that truncated states \( \bar{h}^{(i)} \) are fairly close to the actual states \( h^{(i)} \).

**Lemma C.6** (Truncation impact – random). Given offset \( \tau \) and sampling rate \( L \), consider the state vectors of the sub-trajectory \( \{h^{(i)}\}_{i=1}^\bar{N} \) and \( L - 1 \)-truncations \( \{\bar{h}^{(i)}\}_{i=1}^\bar{N} \). Suppose \( \{u_t\}_{t=0}^N \sim N(0, I) \), \( \|A\| < 1 \), \( h_0 = 0 \), \( \phi = 1 \)-Lipschitz, and \( \phi(0) = 0 \). Also suppose upper bound (4.3) of Assumption [7] holds for some \( \theta \leq \sqrt{n} \gamma_i > 0 \). There exists an absolute constant \( c > 0 \) such that with probability at least \( 1 - 2\bar{N} \exp(-100n) \), for all \( 1 \leq i \leq \bar{N} \), the following bound holds

\[
\|h^{(i)} - \bar{h}^{(i)}\|_2 \leq c\sqrt{n}\|A\|^{L-1}\sqrt{\gamma_i}.
\]

In particular, we can always pick \( \gamma_i = B^2_\infty \) (via Lemma B.3).

**Proof.** Using Assumption [7] we can apply Lemma F.3 on vectors \( \{h_{(i-1)L+\tau,L-1}\}_{i=1}^\bar{N} \). Using a union bound, with desired probability, all vectors obey

\[
\|h_{(i-1)L+\tau,L-1} - \mathbb{E}[h_{(i-1)L+\tau,L-1}]\|_2 \leq (c-1)\sqrt{n}\gamma_i,
\]

for sufficiently large \( c \). Since \( \theta \leq \sqrt{n} \), triangle inequality implies \( \|h_{(i-1)L+\tau,L-1}\|_2 \leq c\sqrt{n}\gamma_i \). Now, applying Lemma C.2 for all \( 1 \leq i \leq \bar{N} \), we find

\[
\|h^{(i)} - \bar{h}^{(i)}\|_2 = \|h_{(i-1)L+\tau,L-1} - \bar{h}_{(i-1)L+\tau,L-1}\|_2 \\
\leq \|A\|^{L-1}\|h_{(i-2)L+\tau+1}\|_2 \\
\leq c\|A\|^{L-1}\sqrt{\gamma_i}.
\]

\[\square\]

D Properties of the Data Matrix

This section utilizes the probabilistic estimates from Section 6 to provide bounds on the condition number of data matrices obtained from the RNN trajectory \( \{u_t\}_{t=0}^N \). Following [12], these matrices \( H, U \) and \( X \) are defined as

\[
H = [h_1 \ldots h_N]^T, \quad U = H = [u_1 \ldots u_N]^T, \quad X = [x_1 \ldots x_N]^T.
\]

(D.1)

The challenge is that, the state matrix \( H \) has dependent rows; which will be addressed by carefully splitting the trajectory \( \{u_t, h_t\}_{t=0}^N \) into multiple sub-trajectories which are internally weakly dependent as discussed in Section 6. We first define the matrices obtained from these sub-trajectories.
Lemma D.2 (Handling perturbation). Given sampling rate \( L > 0 \) and offset \( \bar{\tau} \), consider the matrices \( \tilde{H}, \tilde{H}, \tilde{X} \) of Definition D.1 and let \( Q = [\gamma_+^{-1/2} \tilde{H} \tilde{U}] \). Suppose Assumption 1 holds, \( \phi \) is \( \beta \)-increasing, and \( u_i \sim N(0,I_p) \). There exists an absolute constant \( C > 0 \) such that if \( N \geq C \frac{\gamma_+^2}{\bar{\tau}} \), with probability \( 1 - 8 \exp(-c \frac{\bar{\tau}}{\gamma_+}) \), for all matrices \( M \) obeying \[ \|M - \tilde{H}\| \leq \frac{\sqrt{\gamma - N}}{\bar{\tau}} \], the perturbed \( Q \) matrices given by,
\[
\tilde{Q} = [\gamma_+^{-1/2} M \tilde{U}],
\]
\[
(\Theta + \sqrt{2})^2 \geq \tilde{Q}^T Q \geq \frac{\gamma_+ - 2}{2 \gamma_+}.
\]  
Proof. This result is a direct application of Theorem 4.1 after determining minimum/maximum eigenvalues of population covariance. The cross covariance obeys \( E[\tilde{H}^T \tilde{U}] = 0 \) due to independence. Also, for \( i > 1 \), the truncated state vector \( \tilde{h}_i \) is statistically identical to \( h_{L-1} \) hence \( \Sigma[\tilde{h}_i] \geq \gamma_+ I_n \). Consequently, \( \Sigma[u_i] = I_p \), \( \frac{1}{\gamma_+} \Sigma[\tilde{h}_i] \preceq I_n \), and \( \frac{\gamma_+ - 1}{\gamma_+} I_n \preceq \Sigma[q_i] \preceq I_n \).

Set the matrix \( \tilde{Q} = [q_1, \ldots, q_N]^T \) and note that \( Q = [q_i, \tilde{Q}^T]^T \). Applying Theorem 4.1 on \( Q \) and Corollary 4.2 on \( E \) on \( Q \), we find that, with the desired probability,
\[
\Theta + \sqrt{2} \geq \frac{1}{\sqrt{N}} \|Q\| \geq \frac{1}{\sqrt{N}} \text{ min}(Q) \geq \frac{1}{\sqrt{N}} \text{ min}(\tilde{Q}) \geq \sqrt{\frac{N - 1}{N}} \sqrt{\frac{2 \gamma_+}{3 \gamma_+}} \geq 0.99 \times \sqrt{\frac{2 \gamma_+}{3 \gamma_+}}.
\]

Setting \( E = M - \tilde{H} \) and observing \( \tilde{Q} = Q + [\gamma_+^{-1/2} E 0] \), the impact of the perturbation \( E \) can be bounded naively via \( \text{ min}(Q) - \gamma_+^{-1/2} \|E\| \leq \text{ min}(\tilde{Q}) \leq \|\tilde{Q}\| \leq \|Q\| + \gamma_+^{-1/2} \|E\| \). Using the assumed bound on \( \|E\| \), this yields
\[
\Theta + \sqrt{2} \geq \frac{1}{\sqrt{N}} \|Q\| \geq \frac{1}{\sqrt{N}} \text{ min}(\tilde{Q}) \geq \frac{\gamma_+ - 2}{2 \gamma_+}.
\]

This final inequality is identical to the desired bound (D.3).

Theorem D.3 (Data matrix condition). Consider the nonlinear state equation (\ref{eq:state}). Given \( \gamma_+ \geq \gamma_- > 0 \), define the condition number \( \rho = \frac{\gamma_+}{\gamma_-} \). For some absolute constants \( c, C > 0 \), pick a trajectory length \( N \) where
\[
L = \left[ 1 - \frac{\log (cN\rho)}{\log \|A\|} \right], \quad N_0 = \frac{N}{L}, \quad C \rho^2 (n + p),
\]
and pick scaling \( \mu = \frac{1}{\sqrt{N_0}} \). Suppose \( \|A\| < 1 \), \( \phi \) is \( \beta \)-increasing, \( u_i \overset{i.i.d.}{\sim} N(0,I_p) \), and Assumption 7 holds with \( \gamma_+, \gamma_- \geq \Theta, L \). Matrix \( X = [x_1, \ldots, x_N]^T \) of (\ref{eq:state}) satisfies the following with probability \( 1 - 4N \exp(-100n) - 8L \exp(-c(N_0/(\rho^2))):

- Each row of \( X \) has \( \ell_2 \) norm at most \( c_0 \sqrt{p + n} \) where \( c_0 \) is an absolute constant.
- \( X^T X \) obeys the bound
\[
(\Theta + \sqrt{2})^2 I_{n+p} \geq \frac{X^T X}{\gamma_-} \geq \rho^{-1} I_{n+p}/2.
\]  
Proof. The first statement on \( \ell_2 \)-norm bound can be concluded from Lemma D.4 and holds with probability \( 1 - 2N \exp(-100(n + p)) \). To show the second statement, for a fixed offset \( 1 \leq \bar{\tau} \leq \bar{L} \), consider Definition D.1 and the matrices \( \tilde{H}^{(r)}, \tilde{U}^{(r)}, \tilde{X}^{(r)} \). Observe that \( X \) is obtained by merging multiple sub-trajectory matrices
\( \{ \tilde{X}^{(t)} \}_{t=1} \). We will first show the advertised bound for an individual \( \tilde{X}^{(t)} \) by applying Lemma D.2 and then apply Lemma A.1 to obtain the bound on the combined matrix \( X \).

Recall that \( N_r \) is the length of the \( t \)th sub-trajectory i.e. number of rows of \( \tilde{X}^{(t)} \). By construction \( 2N_0 \geq N_r \geq N_0 \) for all \( 1 \leq \tau \leq L \). Given \( 1 \leq \tau \leq L \) and triple \( \tilde{H}^{(t)}, \tilde{H}^{(t)}, \tilde{U}^{(t)} \), set \( Q = [\mu \tilde{H}^{(t)}, \tilde{U}^{(t)}] \). Since \( N_0 \) is chosen to be large enough, applying Theorem D.2 with \( \mu = 1/\sqrt{\gamma} \) choice, and noting \( \rho = \gamma/\gamma \), we find that, with probability \( 1 - 4 \exp(-c_1 N_0/\rho^2) \), all matrices \( M \) satisfying \( \| M - \tilde{H}^{(t)} \| \leq \sqrt{\gamma N_0}/10 \) and \( Q \) as in (D.2) obeys

\[
(\Theta + \sqrt{2})^2 \geq \frac{\bar{Q}^T \bar{Q}}{N} \geq \rho^{-1}/2. \tag{D.5}
\]

Let us call this Event 1. To proceed, we will argue that with high probability \( \| \tilde{H}^{(t)} - \tilde{H}^{(t)} \| \) is small so that the bound above is applicable with \( M = \tilde{H}^{(t)} \) choice; which sets \( \bar{Q} = \tilde{X}^{(t)} \) in (D.5). Applying Lemma C.6 we find that, with probability \( 1 - 2N_r \exp(-100n) \),

\[
\| \tilde{H}^{(t)} - \tilde{H}^{(t)} \| \leq \sqrt{2N_0} \max \{ \| h^{(i)} - \tilde{h}^{(i)} \|_2 \} \leq c_0 \sqrt{2N_0} \sqrt{n} \| A \|^{L-1}.
\]

Let us call this Event 2. We will show that our choice of \( L \) ensures right hand side is small enough and guarantees \( \| \tilde{H}^{(t)} - \tilde{H}^{(t)} \| \leq \sqrt{\gamma N_0}/10 \). Set \( c = \max \{ 200c_0^2, 1 \} \). Desired claim follows by taking logarithms of upper/lower bounds and cancelling out \( \sqrt{N_0} \) terms as follows

\[
c_0 \sqrt{n} \| A \|^{L-1} \sqrt{\gamma} \leq \sqrt{\gamma}/10 \sqrt{2} \iff (L - 1) \log \| A \| + \log \sqrt{c_0 n} \rho \leq 0 \tag{D.6}
\]

\[
\iff - \frac{\log cn \rho}{2 \log \| A \|} \leq L - 1 \tag{D.7}
\]

\[
\iff L = \left\lceil \frac{1 - \log (cn \rho)}{\log \| A \|} \right\rceil. \tag{D.8}
\]

Here we use the fact that \( \log \| A \| < 0 \) since \( \| A \| < 1 \) and \( cn \rho \geq 0 \). Consequently, both Event 1 and Event 2 hold with probability \( 1 - 4 \exp(-c_1 N_0/\rho^2) - 2N_r \exp(-100n) \), implying (D.5) holds with \( \bar{Q} = \tilde{X}^{(t)} \). Union bounding this over \( 1 \leq \tau \leq L \), (D.5) uniformly holds with \( \bar{Q} = \tilde{X}^{(t)} \) and all rows of \( X \) are \( \ell_2 \)-bounded with probability \( 1 - 4N \exp(-100n) - 8L \exp(-c_1 N_0/\rho^2) \). Applying Lemma A.1 on \( \tilde{X}^{(t)} \), we conclude with the bound (D.4) on the merged matrix \( X \).

**Lemma D.4** (\( \ell_2 \)-bound on rows). Consider the setup of Theorem D.3. With probability \( 1 - 2N \exp(-100(n + p)) \), each row of \( X \) has \( \ell_2 \)-norm at most \( c_0 \sqrt{p + \bar{n}} \) for some constant \( c > 0 \).

**Proof.** The \( t \)th row of \( X \) is equal to \( x_t = \left[ \frac{h_t^T}{\sqrt{\gamma}} \ u_t^T \right]^T \). Since \( \| h_t - E[h_t] \|_{\infty} \leq O(\sqrt{\gamma}) \) and \( \| u_t \|_{\infty} \leq O(1) \), we have that \( \| x_t - E[x_t] \|_{\infty} \leq O(1) \). Now, applying Lemma B.5 on all rows \( x_t \), and using a union bound, with probability at least \( 1 - 2N \exp(-100(n + p)) \), we have that \( \| x_t - E[x_t] \|_{\infty} \leq c_0 \sqrt{n + p} \) for all \( t \). To conclude, note that \( \| E[x_t] \|_{\infty} \leq \| E[h_t] \|_{\infty}/\sqrt{\gamma} \leq \theta \leq 3\sqrt{n} \) via Assumption [4].

## E Proofs of Main Results

### E.1 Proof of Lemma 3.2

The statement follows from upper bound Lemma B.3 and lower bound Lemma B.5.

### E.2 Proof of Theorem 4.2

**Proof.** To prove this theorem, we combine Theorem D.3 with deterministic SGD convergence result of Theorem 4.1. Applying Theorem D.3 with the desired probability, inequality (D.4) holds and for all \( i \), input data satisfies the bound \( \| x_i \|_{\ell_2} \leq \sqrt{(n + p)/(2c_0)} \) for a sufficiently small constant \( c_0 > 0 \). As the next step, we will argue that these two events imply the convergence of SGD.

Let \( \theta^{(i)}, e^{(i)} \in \mathbb{R}^{n+p} \) denote the \( i \)th rows of \( \Theta, C \) respectively. Observe that the square-loss is separable along the rows of \( C \) via \( \| \Theta - C \|_{\phi} = \sum_{i=1}^{n} \| \theta^{(i)} - e^{(i)} \|_{\phi} \). Hence, SGD updates each row \( e^{(i)} \) via its own state equation

\[
y_{t,i} = \phi(y_{t,i}),
\]

where \( y_{t,i} \) is the \( i \)th entry of \( y_t \). Consequently, we can establish the convergence result for an individual row of \( C \). Convergence of all individual rows will imply the convergence of the overall matrix \( \Theta_T \) to the ground.
truth $C$. Pick a row index $i$ ($1 \leq i \leq n$), set $c = e^{(i)}$ and denote the $i$th row of $\Theta$, by $\theta_i$. Also denote the label corresponding to $i$th row by $y_i = y_{i,t}$. With this notation, SGD over (2.3) runs SGD over the $i$th row with equations $y_t = \phi((c, x_t))$ and with loss functions

$$L(\theta) = N^{-1} \sum_{t=1}^{N} L_i(\theta), \quad L_i(\theta) = \frac{1}{2} (y_t - \phi((\theta_i, x_t)))^2.$$  

Substituting our high-probability bounds on $x_t$ (e.g., (3.4)) into Theorem 4.1, we can set $B = (n + p)/(2c_0)$, $\gamma_+ = (\theta + \sqrt{2})^2$, and $\gamma_- = \rho^{-1}/2$. Consequently, using the learning rate $\eta = c_0 \beta^2 \rho^{-1}$ for all $\tau \geq 0$, the $\tau$th SGD iteration $\theta_\tau$ obeys

$$\mathbb{E}||\theta_\tau - c||_2^2 \leq ||\theta_0 - c||_2^2 \left(1 - c_0 \frac{\beta^4 \rho^{-2}}{2(\theta + \sqrt{2})^2(n + p)}\right)^\tau,$$

where the expectation is over the randomness of the SGD updates. This establishes the convergence for a particular row of $C$. Summing up these inequalities (E.1) over all rows $\theta^{(1)}_\tau, \ldots, \theta^{(n)}_\tau$ (which converge to $e^{(1)}, \ldots, e^{(n)}$ respectively) yields the targeted bound (4.4). \hfill $\Box$

### E.3 Proofs of Main Results on Stable Systems

#### E.3.1 Proof of Theorem 3.3

**Proof.** Applying Lemmas B.3 and B.2 independent of $L$, Assumption 1 holds with parameters

$$\gamma_+ = B_\infty^2, \quad \gamma_- = \beta^2 s_{\min}(B)^2, \quad \theta = \sqrt{6n} - \sqrt{2} \geq \sqrt{n}.$$  

This yields $(\theta + \sqrt{2})^2 = 6n$. Hence, we can apply Theorem 4.1 with the learning rate $\eta = c_0 \beta^2 s_{\min}(B)^2$ where

$$\rho = \frac{B_\infty^2}{\beta^2 s_{\min}(B)^2} = \frac{\gamma_+}{\gamma_-},$$

(E.2)

and convergence rate $1 - \frac{\beta^2 s_{\min}(B)^2}{2\rho}$ to conclude with the stated result, we use the change of variable $c_0/6 \to c_0$. \hfill $\Box$

#### E.3.2 Proof of Theorem 3.4

**Proof.** The proof is similar to that of Theorem 3.3. Applying Lemmas B.3, B.4 and B.2 independent of $L$, Assumption 1 holds with parameters

$$\gamma_+ = B_\infty^2, \quad \gamma_- = s_{\min}(B)^2, \quad \theta = 0.$$  

Hence, we again apply Theorem 4.1 with the learning rate $\eta = c_0 \beta^2 s_{\min}(B)^2$ where $\rho$ is given by (E.2). Use the change of variable $c_0/2 \to c_0$ to conclude with the stated result. \hfill $\Box$

### E.4 Learning Unstable Systems

In a similar fashion to Section 4, we provide a more general result on unstable systems that makes a parametric assumption on the statistical properties of the state vector.

**Assumption 2** (Well-behaved state vector – single timestamp). Given timestamp $T_0 > 0$, there exists positive scalars $\gamma_+, \gamma_-, \theta$ and an absolute constant $C > 0$ such that $\theta \leq 3\sqrt{n}$ and the following holds

$$\gamma_+ I_n \geq \Sigma h_{\tau} \geq \gamma_- I_n, \quad \|h_{\tau} - \mathbb{E}[h_{\tau}]\|_\psi \leq C \sqrt{T_0} \quad \text{and} \quad \|h_{\tau}\|_\Sigma \leq \theta \sqrt{T_0}. \tag{E.3}$$

The next theorem provides the parametrized result on unstable systems based on this assumption.

**Theorem E.1** (Unstable system – general). Suppose we are given $N$ independent trajectories $(h_{\tau}^{(i)}, u_{\tau}^{(i)})$ for $1 \leq i \leq N$. Each trajectory at time $T_0$ to obtain $N$ samples $(y_i, h_i, u_i)_{i=1}^{N}$ where ith sample is

$$(y_i, h_i, u_i) = (h_{T_0}^{(i)}, h_{T_0}^{(i)} + u_{T_0}^{(i)}).$$

Let $C, c_0 > 0$ be absolute constants. Suppose Assumption 1 holds with $T_0$ and sample size satisfies $N \geq C p^2(n + p)$ where $\rho = \gamma_+ / \gamma_-$. Assume $\phi$ is $\beta$-increasing, zero initial state conditions, and $u_i^{(i)} \sim N(0, I_p)$. Set scaling to be $\mu = 1/\sqrt{T_0}$ and learning rate to be $\eta = c_0 \beta^2 \rho^{-1}$. Starting from $\Theta_0$, we run SGD over the equations described in (2.2) and (2.3). With probability $1 - 2N \exp(-100(n + p)) - 4 \exp(-O(\frac{N}{p}))$, all iterates satisfy

$$\mathbb{E}[\|\Theta_i - C\|_F^2] \leq (1 - c_0 \rho^{2}(\theta + \sqrt{2})^2(n + p))^{\tau} \|\Theta_0 - C\|_F^2,$$

where the expectation is over the randomness of the SGD updates.
Applying Lemma B.3, we can substitute \( \gamma \). Theorem F.1. Suppose \( \hat{Y} \) is a functionals (see Talagrand (2014)) obey
\[
\gamma_{\psi}(K) \geq 20\gamma_{\psi}(K)
\]
functionals (see Talagrand (2014)).

E.5 Proof of Theorem F.1

Proof. The proof is a corollary of Theorem E.1. We need to substitute the proper values in Assumption 2. Applying Lemma B.3, we can substitute \( \gamma_{\psi} = B_0 \) and \( \theta = \sqrt{\gamma_{\psi}} \sqrt{2} \geq \sqrt{n} \). Next, we need to find a lower bound. Applying Lemma F.2 for \( n > 1 \) and Lemma F.6 for \( n = 1 \), we can substitute \( \gamma_{\psi} = \gamma_{+} / \rho \) with the \( \rho \) definition of F.2. With these, the result follows as an immediate corollary of Theorem E.1.

F Supplementary Statistical Results

The following theorem bounds the empirical covariance of matrices with independent subgaussian rows. Given a random vector \( x \), define the de-biasing operation as \( zm(x) = x - E[x] \).

Theorem F.1. Let \( A \in \mathbb{R}^{n \times d} \) be a matrix with independent subgaussian rows \( \{a_i\}_{i=1}^n \) satisfying \( \|zm(a_i)\|_{\ell_2} \leq O(K) \) and \( \sum a_i \leq K^2 \) for some \( K > 0 \) and \( \|E[a_i]\|_{\ell_2} \leq \theta \). Suppose \( \sum a_i \geq \lambda I_d \). Suppose \( n \geq O(K^d / \lambda^2) \). Then, with probability at least \( 1 - 4 \exp(-cK^2 \lambda^2 n) \),
\[
\theta + \sqrt{3/2K} \geq \frac{1}{\sqrt{n}} \|A\| \geq \frac{1}{\sqrt{n}} s_{\text{min}}(A) \geq \sqrt{2\lambda/3}.
\]

Proof. Let \( E = E[A], \tilde{A} = A - E[A], \tilde{a}_i = zm(a_i) \). Observe that
\[
\|A\|_{\ell_2}^2 - \|E[A]\|_{\ell_2}^2 = \|\tilde{A}v\|_{\ell_2}^2 + 2v^T \tilde{A} E v + \|E v\|_{\ell_2}^2 - \|A v\|_{\ell_2}^2 = \|\tilde{A}v\|_{\ell_2}^2 + 2v^T \tilde{A} E v + \|E v\|_{\ell_2}^2.
\]
Define the random process \( X_v = \|\tilde{A}v\|_{\ell_2}^2 \) and \( X_v = X_v - E[X_v] \). First, we provide a deviation bound for the quantity \( \sup_{v \in S^{d-1}} X_v \). To achieve this, we will utilize Talagrand’s mixed tail bound and show that increments of \( X_v \) are subexponential. Pick two unit vectors \( v, u \in \mathbb{R}^d \). Write \( x = u + v, y = u - v \). We have that
\[
X_u - X_v = \|\tilde{A}u\|_{\ell_2}^2 - \|\tilde{A}v\|_{\ell_2}^2 = \|\tilde{A}(x + y) / 2\|_{\ell_2}^2 - \|\tilde{A}(x - y) / 2\|_{\ell_2}^2 = x^T \tilde{A}^T \tilde{A} y + \sum_{i=1}^n \tilde{a}_i (x)(\tilde{a}_i)^T y.
\]
Letting \( \tilde{x} = x / \|x\|_{\ell_2}, \tilde{y} = y / \|y\|_{\ell_2} \), observe that, multiplication of subgaussians \( x^T \tilde{a}_i, y^T \tilde{a}_i \) obey
\[
\|x^T \tilde{a}_i\|_{\ell_2} \leq O(\|x\|_{\ell_2} \|y\|_{\ell_2} K^2) \leq O(K^2 \|y\|_{\ell_2}).
\]
Centering this subexponential variable around zero introduces a factor of 2 when bounding subexponential norm and yields \( \|x^T \tilde{a}_i\|_{\ell_2} \leq O(\|x^T \tilde{a}_i\|_{\ell_2} \|y^T \tilde{a}_i\|_{\ell_2}) \). Now, using the fact that \( X_u - X_v \) is sum of \( n \) independent zero-mean subexponential random variables, we have the tail bound
\[
P(n^{-1} |Y_u - Y_v| \geq t) \leq 2 \exp(-c' n \min\{t^2 / K^2 \|y\|_{\ell_2}^2, t^2 / K^2 \|y\|_{\ell_2}^2\}).
\]
Applying Talagrand’s chaining bound for mixed tail processes with distance metrics \( p_2 = K^2 \|z\|_{\ell_2}^2, p_1 = \sqrt{\lambda} \|z\|_{\ell_2}^2 \) (Theorem 3.5 of [Dirksen 2013] or Theorem 2.2.23 of [Talagrand 2014]) and using the fact that for unit sphere \( S^{d-1} \), Talagrand’s \( \gamma \) functional (see [Talagrand 2014]) obey \( \gamma_1(S^{d-1}), \gamma_2(S^{d-1}) \leq O(d), \).
\[
n^{-1} \sup_{u \in S^{d-1}} |Y_v| \leq cK^2 (\sqrt{d/n} + d/n + t / \sqrt{n}),
\]
(F.1)
Then, for any $m \geq 1$. To conclude, use the change of variable

$$T = \frac{1}{\log K^2} \sqrt{\lambda n},$$

in line (F.2) and only focusing on the spectral norm bound. The following lemma is fairly standard and is proved for the sake of completeness.

We can pick a bounding over all $v$. This leads to the following inequalities

$$\frac{1}{n} \| \tilde{A}^T \tilde{A} - \mathbb{E}[\tilde{A}^T \tilde{A}] \| \leq \frac{\lambda}{8} \implies \frac{9K^2}{8} I_d \geq \frac{1}{n} \tilde{A}^T \tilde{A} \geq \frac{7}{8} I_d.$$

(F.2)

$$\implies \frac{9}{8} K \geq \frac{1}{\sqrt{n}} \| \tilde{A} \| \geq s_{\min}(\tilde{A}) \geq \sqrt{\frac{\lambda}{8}}.

Denote the size $n$ all ones vector by $1_n$. Next, we define the process $Z_v = \frac{1}{\sqrt{n}} 1_n^T \tilde{A} v$. Observe that $\tilde{A}^T 1_n = \sum_{i=1}^n \tilde{a}_i \in \mathbb{R}^d$ is a vector satisfying $\| \tilde{A}^T 1_n / \sqrt{n} \|_{\psi_2} \leq \mathcal{O}(K)$. Hence, again using $n \geq K^2 \lambda^{-2} d$ for sufficiently large $C > 0$, applying Lemma (F.3) with $m = c_0 K^{-4} \lambda^2 n > d$ by picking a sufficiently small constant $c_0 > 1/C$, with probability at least $1 - 2 \exp(-100c_0 K^{-4} \lambda^2 n)$

$$\frac{1}{\sqrt{n}} \sup_{|v|_{\ell_2} = 1} |Z_v| = \frac{1}{n} \| \tilde{A}^T 1_n \|_{\ell_2} \leq \frac{1}{12} K K^{-2} \lambda \leq \sqrt{\frac{\lambda}{12}}.

Let $P = I_n - \frac{1}{n} 1_n 1_n^T$ be the projection onto the orthogonal complement of the all ones vector. Note that $P \mathbb{E} v = 0$ as the rows of $\mathbb{E}$ are equal. With this observation, with desired probability, for any unit length $v$,

$$\| A v \|_{\ell_2} \geq \| P A v \|_{\ell_2} \geq \| A v \|_{\ell_2} - |Z_v| \geq s_{\min}(A) - \sup_{v \in S^{d-1}} |Z_v| \geq (\sqrt{\frac{\lambda}{8}} - 1/12) \sqrt{n},

(F.3)

which implies $s_{\min}(A)/\sqrt{n} \geq 2\lambda/3$. For spectral norm of $A$, we use the naive bound

$$\frac{1}{\sqrt{n}} \| A \| \leq \frac{1}{\sqrt{n}} (\| \mathbb{E} \| + \| A \|) \leq \max_{1 \leq i \leq n} \| \mathbb{E}[a_i] \|_{\ell_2} + 9K/8 \leq \theta + \sqrt{3}/2K.

\square

The corollary below is obtained by slightly modifying the proof above by using $\frac{1}{n} \| \tilde{A}^T \tilde{A} - \mathbb{E}[\tilde{A}^T \tilde{A}] \| \leq \frac{K^2}{8}$ in line (F.2) and only focusing on the spectral norm bound.

**Corollary F.2.** Let $A \in \mathbb{R}^{n \times d}$ be a matrix with independent $\{a_i\}_{i=1}^n$ subgaussian rows satisfying $\| z a_i \|_{\psi_2} \leq \mathcal{O}(K)$ and $\Sigma[a_i] \preceq K^2 I_d$ for some $K > 0$ and $\| \mathbb{E}[a_i] \|_{\ell_2} \leq \theta$. Suppose $\Sigma[a_i] \preceq \lambda I_d$. Suppose $n \geq \mathcal{O}(K^2 d)$. Then, with probability at least $1 - 4 \exp(-c K^{-2} n)$,

$$\theta + \sqrt{3}/2K \geq \frac{1}{\sqrt{n}} \| A \|.

The following lemma is fairly standard and is proved for the sake of completeness.

**Lemma F.3** (Subgaussian vector length). Let $a \in \mathbb{R}^n$ be a zero-mean subgaussian vector with $\| a \|_{\psi_2} \leq L$. Then, for any $m \geq n$, there exists $C > 0$ such that

$$\mathbb{P}(\| a \|_{\ell_2} \leq C L \sqrt{m}) \geq 1 - 2 \exp(-100m).

**Proof.** We can pick a $1/2$ cover $C$ of the unit $\ell_2$-sphere with size $\log |C| \leq 2n$. For any $v \in C$, subgaussianity implies, $\mathbb{P}(\| v \|_{\ell_2} \geq t) \leq 2 \exp(-\frac{t^2}{2L^2})$. Setting $t = C L \sqrt{m}$ for sufficiently large constant $C > 0$, and union bounding over all $v \in C$, we find

$$\mathbb{P}(\bigcap_{v \in C} \| v \|_{\ell_2} \leq C L \sqrt{m}) \geq 1 - 2 \exp(2n - \frac{c L^2 m}{2L^2}) \leq 1 - 2 \exp(-100m).

To conclude, let $v(a) \in C$ be $a$’s neighbor satisfying $\| v - \frac{a}{\| a \|_{\ell_2}} \|_{\ell_2} \leq 1/2$. Hence, we have

$$\| a \|_{\ell_2} \leq \| (a - v(a))^T a \|_{\ell_2} + \| v(a) a \|_{\ell_2} \leq \| a \|_{\ell_2} / 2 + C L \sqrt{m} \implies \| a \|_{\ell_2} \leq 2 C L \sqrt{m}.

To conclude, use the change of variable $C \rightarrow C/2$. \square