Journal of Machine Learning Research 21 (2020) 1-20 Submitted 8/19; Published 10/20

Convex Programming for Estimation in Nonlinear Recurrent Models

Sohail Bahmani SOHAIL.BAHMANI@ECE.GATECH.EDU
School of Electrical & Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332

Justin Romberg JROM@ECE.GATECH.EDU
School of Electrical & Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332

Editor: Amir Golberson

Abstract

We propose a formulation for nonlinear recurrent models that includes simple parametric models of recurrent neural networks as a special case. The proposed formulation leads to a natural estimator in the form of a convex program. We provide a sample complexity for this estimator in the case of stable dynamics, where the nonlinear recursion has a certain contraction property, and under certain regularity conditions on the input distribution. We evaluate the performance of the estimator by simulation on synthetic data. These numerical experiments also suggest the extent at which the imposed theoretical assumptions may be relaxed.

Keywords: recurrent neural networks, convex programming, dynamical systems, VC dimension

1. Introduction

Given a *differentiable* and *convex* function $f : \mathbb{R}^n \to \mathbb{R}$ with $\nabla f(\mathbf{0}) = \mathbf{0}$, we consider the dynamics described by the recursion

$$\boldsymbol{x}_t = \nabla f\left(\boldsymbol{A}_\star \boldsymbol{x}_{t-1} + \boldsymbol{B}_\star \boldsymbol{u}_{t-1}\right) , \tag{1}$$

where $\boldsymbol{u}_0, \boldsymbol{u}_1, \dots$ are i.i.d. copies of a random vector $\boldsymbol{u} \in \mathbb{R}^p$, the initial state $\boldsymbol{x}_0$ is zero, and the matrices $\boldsymbol{A}_\star \in \mathbb{R}^{n\times n}$ and $\boldsymbol{B}_\star \in \mathbb{R}^{n\times p}$ are the parameters of the model. In the setup described above, we want to address the following problem.

Problem. Given a time horizon T, estimate the model parameters $\boldsymbol{A}_\star$ and $\boldsymbol{B}_\star$, from a single observed trajectory $(\boldsymbol{u}_0, \boldsymbol{x}_0 = \mathbf{0}), (\boldsymbol{u}_1, \boldsymbol{x}_1), \dots, (\boldsymbol{u}_T, \boldsymbol{x}_T)$.

The specific form of the nonlinearity in (1) might seem strange at first, but many common choices of nonlinearities used in practice are special cases of this formulation. For instance, increasing nonlinearities that act coordinate-wise can be modeled by choosing the appropriate *separable* convex function f in (1). Particularly, the (parameterized) ReLU function $x \mapsto x_+ + c(-x)_+$ for some constant $c \geq 0$, which is popular in neural network models, corresponds to the choice of $f(\boldsymbol{x}) = \sum_{i=1}^n (x_i)_+^2/2 + c(-x_i)_+^2/2$ in our proposed model.

Intuitively, $\boldsymbol{A}_\star$ and $\boldsymbol{B}_\star$ affect the conditioning of the problem differently; $\boldsymbol{A}_\star$ applies to the state variable that may have low temporal variation, whereas $\boldsymbol{B}_\star$ applies to the input variable that is constantly changing. For example, if $\boldsymbol{B}_\star$ is dominated by $\boldsymbol{A}_\star$, say, in the Frobenius norm, one may need to observe the system for a much longer time to collect enough information about $\boldsymbol{A}_\star$. Therefore, to estimate $\boldsymbol{A}_\star$ and $\boldsymbol{B}_\star$ in tandem, it is reasonable to scale one component to avoid error in one component dominating the error in the other. We introduce $\beta > 0$ as a sufficiently large normalizing constant to address a potential imbalance between the components $\boldsymbol{A}_\star$ and $\boldsymbol{B}_\star$. We collect the ground truth parameters in $\boldsymbol{C}_\star = \begin{bmatrix}\boldsymbol{A}_\star & \beta^{-1}\boldsymbol{B}_\star\end{bmatrix}$ and for $t = 0, 1, \dots,$ we set

$$\boldsymbol{z}_t = \begin{bmatrix}\boldsymbol{x}_t \\ \beta\,\boldsymbol{u}_t\end{bmatrix}.$$

Therefore, the dynamics can be equivalently expressed as

$$\boldsymbol{z}_0 = \begin{bmatrix}\boldsymbol{0} \\ \beta\,\boldsymbol{u}_0\end{bmatrix}, \text{ and} \qquad \boldsymbol{z}_t = \begin{bmatrix}\nabla f\left(\boldsymbol{C}_\star \boldsymbol{z}_{t-1}\right) \\ \beta\,\boldsymbol{u}_t\end{bmatrix}, \text{ for } t \geq 1\,.$$

Our goal is effectively to estimate $\boldsymbol{C}_\star$, from the observations $\boldsymbol{z}_0, \boldsymbol{z}_1, \dots, \boldsymbol{z}_T$.

Not surprisingly, further model assumptions are needed to exclude inherently intractable instances of the problem. Below in Section 1.3 we state the assumptions we make to analyze the problem. Under these assumptions, by Theorem 1 we show that if T (i.e., the observation horizon) scales with $n + p$ up to some polylogarithmic factors, the estimator described in Section 1.2 recovers $\boldsymbol{C}_\star$ with high probability. As explained in that section, our argument effectively reduces to establishing a positive lower bound for the eigenvalues of the "sample correlation matrix" of the vectors $\boldsymbol{z}_t$.

In this paper we provide statistical guarantees for the parameter estimation under the nonlinear recurrent model (1), which itself generalizes the most prevalent recurrent models. Specifically, we propose a novel estimator that, unlike the few existing competitors, is formulated as a convex program. More importantly, we show that this estimator is accurate, with a sample complexity that scales gracefully with the dimensions of the problem, even for heavy-tailed input distributions.

1.1 Related work

Recurrent Neural Networks (RNN) and similar models of random dynamical systems have become the main tool in machine learning applications dealing with sequential data. In this section we briefly review some recent results that provide theoretical analysis for these models.

Parameter estimation in discrete-time *linear* dynamical systems whose state variable are generally governed by the recursion

$$\boldsymbol{x}_t = \boldsymbol{A}_\star \boldsymbol{x}_{t-1} + \boldsymbol{B}_\star \boldsymbol{u}_{t-1} + \boldsymbol{\xi}_t\,, \tag{2}$$

with ξ_t denoting an additive observation noise, are studied in (Hardt et al., 2018; Faradonbeh et al., 2018; Simchowitz et al., 2018; Du et al., 2018; Sarkar and Rakhlin, 2019). The difference in the mentioned results stem from variations to the model such as

- observing the state variable *indirectly*, through the sequence

$$\boldsymbol{y}_t = \boldsymbol{A}'_\star \boldsymbol{x}_t + \boldsymbol{B}'_\star \boldsymbol{u}_t + \boldsymbol{\xi}'_t\,, \tag{3}$$

  with ξ'_t denoting an additive output noise,

- observing *single versus multiple trajectories*,
- restricting $\boldsymbol{A}_\star$ (e.g., $\max_i |\lambda_i(\boldsymbol{A}_\star)| < 1$ in the *stable model* versus $\min_i |\lambda_i(\boldsymbol{A}_\star)| > 1$ in the *explosive model*), and
- choosing to have an input (i.e., $\boldsymbol{B}_\star = \boldsymbol{0}$ versus $\boldsymbol{B}_\star \neq \boldsymbol{0}$) with a certain distribution.

Hardt et al. (2018) consider a prediction problem in *controllable* linear dynamical systems with indirect observations as in (3). Specifically, formulating the prediction problem naturally as a (non-convex) least squares, the prediction error achieved by stochastic gradient descent (SGD) is analyzed under some technical assumptions. In a stable *single input single output* setting, it is shown that with N trajectories of length T observed, the prediction error can be bounded as $O(\sqrt{(n^5 + \sigma^2 n^3)/(TN)})$ where n is the number of controllable parameters, and σ^2 is the variance of the zero-mean noise terms ξ'_t in (3).

Under technical assumptions, Faradonbeh et al. (2018) establish a sample complexity for estimation of $\boldsymbol{A}_\star$ in the explosive regime (i.e., $\min_i |\lambda_i(\boldsymbol{A}_\star)| > 1$) with heavy-tailed noise and deactivated input (i.e., $\boldsymbol{B}_\star = \boldsymbol{0}$).

Simchowitz et al. (2018) analyze the ordinary least squares (OLS) in estimation of $\boldsymbol{A}_\star$ from a single trajectory of observations $\boldsymbol{x}_0, \boldsymbol{x}_1, \dots$ where there is no input (i.e., $\boldsymbol{B}_\star = \boldsymbol{0}$) and the process noise is i.i.d. samples of a zero-mean isotropic Gaussian random variable. It is shown in Simchowitz et al. (2018) that for "marginally stable" systems (i.e., $\max_i |\lambda_i(\boldsymbol{A}_\star)| \le 1$), the estimate $\widehat{\boldsymbol{A}}$ produced by the OLS, with high probability, achieves the natural error rate of $\left\|\widehat{\boldsymbol{A}} - \boldsymbol{A}_\star\right\| \lesssim \sqrt{n/T}$ up to some constants and log factors depending implicitly on $\boldsymbol{A}_\star$. Remarkably, this result applies to systems where the spectral radius of $\boldsymbol{A}_\star$ equals one (i.e., $\max_i |\lambda_i(\boldsymbol{A}_\star)| = 1$) where the more standard arguments based on mixing time which require stability of the system do not apply.

Oymak and Ozay (2018) considers estimation from a single trajectory of input/output observation pairs $(\boldsymbol{u}_0, \boldsymbol{y}_0), (\boldsymbol{u}_1, \boldsymbol{y}_1), \dots$ where the output sequence $\boldsymbol{y}_0, \boldsymbol{y}_1, \dots$ is generated by the recursion (3). Assuming the input, the state noise, and the output noise each to have i.i.d. samples from zero-mean isotropic Gaussian distributions, (Oymak and Ozay, 2018) studies accuracy of a least squares approach in estimation of the parameter matrix $\boldsymbol{G}_\star = \begin{bmatrix} \boldsymbol{B}'_\star & \boldsymbol{A}'_\star \boldsymbol{B}_\star & \boldsymbol{A}'_\star \boldsymbol{A}_\star \boldsymbol{B}_\star & \cdots & \boldsymbol{A}'_\star \boldsymbol{A}_\star^{T-2} \boldsymbol{B}_\star \end{bmatrix}$ that characterizes the dynamics.

Similar to (Simchowitz et al., 2018), Sarkar and Rakhlin (2019) establish the estimation error rate for OLS in the single observation trajectory regime under the model (2) with deactivated input (i.e., $\boldsymbol{B}_\star = \boldsymbol{0}$) and sub-Gaussian noise $\boldsymbol{\xi}_t$. Particularly, in the three regimes of stable or marginally stable systems (i.e. $\max_i |\lambda_i(\boldsymbol{A}_\star)| < 1 + O(1/T)$), marginally stable systems (i.e. $\max_i |\lambda_i(\boldsymbol{A}_\star)| < 1 - O(1/T)$), and explosive systems (in the sense that $\min_i |\lambda_i(\boldsymbol{A}_\star)| > 1 + O(1/T)$) the operator norm of the error roughly decays as $1/\sqrt{T}$, $1/T$, and e^{-T}, respectively.

Du et al. (2018) study the minimax rate of estimation from multiple trajectories in simple linear recurrent neural networks (and convolutional neural networks). Considering the state variable to be linearly collapsed to a scalar in the output, under a subGaussian model for the input sequence as well as the output noise, the mentioned paper provides upper and lower bounds for the minimax risk of the mean squared error. In particular, it is shown that the minimax rate of estimating from T trajectories of length L, is orderwise between $\sqrt{\min\{n,L\}p/T}$ and $\sqrt{(p+L)\min\{n,p\}\log(Lp)/T}$.

From a technical point of view, linearity in recurrent models typically provides the convenience of "unfolding" the state recursion into explicit equations in terms of the past input. This convenient feature disappears immediately as nonlinearities are introduced in the recursion as in (1). Miller and Hardt (2019) showed that in the stable regime nonlinear RNNs can be approximated by "truncated" RNNs. Furthermore, they showed that, for unstable RNNs, gradient descent does not necessarily converge. Oymak (2019) studies the parameter estimation under (1) when the nonlinearity ∇f is replaced by an activation function that is *strictly increasing* and applies coordinatewise. Formulating the problem as nonconvex least squares, (Oymak, 2019) establishes a sample complexity for the convergence of in Frobenius norm. Basically, (Oymak, 2019) shows that if $T \gtrsim \rho(n+p)$ with ρ being a certain notion of condition number of $\boldsymbol{B}_\star$, then with high probability, SGD converges at a linear, albeit dimension dependent, rate. Allen-Zhu and Li (2019) analyzed the performance of SGD applied to the Elman's model of RNNs. However, their result is not immediately comparable to our results because of the differences in the setup, e.g.,they consider prediction performance from multiple independent trajectories of observations whereas we consider parameter estimation from a single trajectory.

In this paper, we generalize the results of (Oymak, 2019) in two directions. First, our formulation of the recurrence (1) admits a broader class of nonlinearities, and, as will be seen in the sequel, it enables us to formulate a *convex program* as the estimator. Second, the analysis of (Oymak, 2019) relies critically on the assumption that the input distribution is Gaussian. This is partly due to the use of the Gaussian concentration inequality for Lipschitz functions. At the cost of having a stricter form of nonlinearity, we relax the requirement on the input distribution by allowing the random input to have heavier tail.

1.2 Proposed Estimator

Our proposed estimator is formulated as a convex program as follows

$$\widehat{\boldsymbol{C}} \in \underset{\boldsymbol{C}\in\mathbb{R}^{n\times(n+p)}}{\operatorname{argmin}} \sum_{t=1}^{T} f(\boldsymbol{C}\boldsymbol{z}_{t-1}) - \langle \boldsymbol{x}_t, \boldsymbol{C}\boldsymbol{z}_{t-1}\rangle\,. \tag{4}$$

Readers familiar with convex analysis may observe that if f^*, the *convex conjugate* of f, is smooth, then $\nabla f^* \equiv (\nabla f)^{-1}$ and (1) is equivalent to

$$\nabla f^*(\boldsymbol{x}_{t+1}) = \boldsymbol{A}_\star \boldsymbol{x}_{t-1} + \boldsymbol{B}_\star \boldsymbol{u}_{t-1}\,.$$

Should ∇f^* be easy to compute, it is evident that the resulting system of linear equations can be solved by the common least squares approach to estimate $\boldsymbol{A}_\star$ and $\boldsymbol{B}_\star$. However, we prefer (4) as the estimator, since it can be implemented regardless of f^* and its properties.

In view of (1) and convexity of f, it is straightforward to verify that $\boldsymbol{C}_\star$ is a minimizer for (4). Under the assumptions specified below in Section 1.3, we will show that the minimizer of (4) is unique and therefore $\widehat{\boldsymbol{C}} = \boldsymbol{C}_\star$. In particular, with f assumed to be λ-strongly convex, we have

$$f(\boldsymbol{C}\boldsymbol{z}_{t-1}) - \langle \boldsymbol{x}_t, \boldsymbol{C}\boldsymbol{z}_{t-1}\rangle \geq f(\boldsymbol{C}_\star\boldsymbol{z}_{t-1}) - \langle \boldsymbol{x}_t, \boldsymbol{C}_\star\boldsymbol{z}_{t-1}\rangle + \frac{\lambda}{2}\|(\boldsymbol{C}-\boldsymbol{C}_\star)\boldsymbol{z}_{t-1}\|_2^2\,.$$

Therefore, to guarantee uniqueness of the minimizer in (4), it suffices to show that, with high probability, the smallest eigenvalue of

$$\boldsymbol{\Sigma} \stackrel{\text{def}}{=} \sum_{t=0}^{T-1} \boldsymbol{z}_t\boldsymbol{z}_t^\mathsf{T}\,, \tag{5}$$

is strictly positive with high probability.

1.3 Assumptions

With no restricting conditions imposed on the observation model in (1), the posed estimation problem is not meaningful. For instance, any affine function f is permitted in the core model above, but clearly its corresponding trajectory conveys no information about $\boldsymbol{C}_\star$.

Assumption 1 (regularity of f)**.** *The function f has the following properties:*

1. *The function f is λ-strongly convex and Λ-smooth in the usual sense, i.e.,*

$$\frac{\lambda}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 \leq f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle \leq \frac{\Lambda}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2\,, \tag{6}$$

holds for all $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$.

2. *There exist a matrix-valued function $F\colon \mathbb{R}^n \to \mathbb{R}^{n\times n}$ and some constant $\varepsilon > 0$ such that, for all $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$, we have*

$$\left\|\frac{1}{2}\left(\nabla f(\boldsymbol{x}+\boldsymbol{y}) - \nabla f(\boldsymbol{x}-\boldsymbol{y})\right) - F(\boldsymbol{x})\boldsymbol{y}\right\|_2 \leq \varepsilon\|\boldsymbol{y}\|_2\,. \tag{7}$$

Perhaps the simplest example of the functions that meet the conditions of Assumption 1 is the convex quadratic functions. Let $\boldsymbol{Q}$ be a positive semidefinite matrix that satisfies $\lambda\boldsymbol{I} \preceq \boldsymbol{Q} \preceq \Lambda\boldsymbol{I}$. Then $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\mathsf{T}\boldsymbol{Q}\boldsymbol{x}$ clearly satisfies (6), and also satisfies (7) for $F(\boldsymbol{x}) \equiv \boldsymbol{Q}$ and $\varepsilon = 0$. Using the mean value theorem, we can easily generalize this example to all twice-differentiable convex functions f whose Hessian obeys $\lambda\boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq \Lambda\boldsymbol{I}$ for all $\boldsymbol{x}$. For such functions, F and ε can be chosen respectively as $F(\boldsymbol{x}) = (\lambda+\Lambda)/2\,\boldsymbol{I}$ and $\varepsilon = (\Lambda-\lambda)/2$.

Another important example of the function f that meets the above conditions, is the piecewise quadratic function

$$f(\boldsymbol{x}) = \frac{1}{2}\sum_{i=1}^{n} \max\{\lambda(-x_i)_+^2, \Lambda(x_i)_+^2\}\,.$$

The gradient of this function, which can be used as the nonlinearity in (1), is given by

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\Lambda+\lambda}{2}x_1 + \frac{\Lambda-\lambda}{2}|x_1| \\ \vdots \\ \frac{\Lambda+\lambda}{2}x_n + \frac{\Lambda-\lambda}{2}|x_n| \end{bmatrix}\,,$$

whose coordinates happen to be the (parameterized) ReLU functions. For this specific f, the mapping F can be chosen as

$$F(\boldsymbol{x}) = \begin{bmatrix} \frac{\varLambda+\lambda}{2} + \frac{\varLambda-\lambda}{2}\,\mathrm{sgn}(x_1) & & & \mathbf{0} \\ & \frac{\varLambda+\lambda}{2} + \frac{\varLambda-\lambda}{2}\,\mathrm{sgn}(x_2) & & \\ & & \ddots & \\ \mathbf{0} & & & \frac{\varLambda+\lambda}{2} + \frac{\varLambda-\lambda}{2}\,\mathrm{sgn}(x_n) \end{bmatrix},$$

for which (7) holds with $\varepsilon = (\varLambda - \lambda)/2$.

An immediate consequence of Assumption 1 is the following.

Lemma 1. *Under Assumption 1, the mapping F obeys*

$$(\lambda - \varepsilon)\|\boldsymbol{y}\|_2 \leq \|F(\boldsymbol{x})\boldsymbol{y}\|_2 \leq (\varLambda + \varepsilon)\|\boldsymbol{y}\|_2\,,$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.

Proof. Using the standard equivalent definitions of strong convexity and smoothness (Nesterov, 2013, Theorem 2.1.5), we have

$$\lambda\|\boldsymbol{y} - \boldsymbol{x}\|_2 \leq \|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|_2 \leq \varLambda\|\boldsymbol{y} - \boldsymbol{x}\|_2\,.$$

Rewriting these inequalities, in terms of the pair $(-\boldsymbol{x} + \boldsymbol{y}, \boldsymbol{x} + \boldsymbol{y})$ in place of $(\boldsymbol{x}, \boldsymbol{y})$, we can obtain

$$2\lambda\|\boldsymbol{y}\| \leq \|\nabla f(\boldsymbol{x} + \boldsymbol{y}) - \nabla f(\boldsymbol{x} - \boldsymbol{y})\|_2 \leq 2\varLambda\|\boldsymbol{y}\|_2\,.$$

Furthermore, by (7) and the triangle inequality we have

$$\|F(\boldsymbol{x})\boldsymbol{y}\|_2 - \varepsilon\|\boldsymbol{y}\|_2 \leq \frac{1}{2}\|\nabla f(\boldsymbol{x} + \boldsymbol{y}) - \nabla f(\boldsymbol{x} - \boldsymbol{y})\|_2 \leq \|F(\boldsymbol{x})\boldsymbol{y}\|_2 + \varepsilon\|\boldsymbol{y}\|_2\,.$$

The lemma easily follows from the latter two lines of inequalities. □

We make the following assumption on the input $\boldsymbol{u}$.

Assumption 2 (regularity of the input distribution)**.** *The input $\boldsymbol{u}$ has the following properties:*

1. *The input $\boldsymbol{u} \in \mathbb{R}^p$ is a zero-mean isotropic random variable, i.e.,*

   $$\mathbb{E}(\boldsymbol{u}) = \mathbf{0}, \text{ and } \mathbb{E}(\boldsymbol{u}\boldsymbol{u}^\intercal) = \boldsymbol{I}\,.$$

2. *The coordinates of $\boldsymbol{u}$ have independent symmetric distributions, i.e., for all measurable subsets $\mathcal{A} = \mathcal{A}_1 \times \ldots \times \mathcal{A}_p$ of $\mathbb{R}^p$, we have*

   $$\mathbb{P}(\boldsymbol{u} \in \mathcal{A}) = \prod_{i=1}^{p} \mathbb{P}(u_i \in \mathcal{A}_i) = \prod_{i=1}^{p} \mathbb{P}(-u_i \in \mathcal{A}_i)\,.$$

3. *For a certain $\alpha \geq 1$, the input $\boldsymbol{u}$ has a bounded directional Orlicz ψ_α norm, i.e., there exists a finite absolute constant $K > 0$ such that*

   $$\sup_{\boldsymbol{h} \in \mathbb{S}^{p-1}} \mathbb{E}\left(e^{|\langle \boldsymbol{h}, \boldsymbol{u}\rangle|^\alpha / K^\alpha}\right) \leq 2\,. \tag{8}$$

The following lemma is an immediate consequence of (8).

Lemma 2. *Let $\boldsymbol{u}$ be the random variable under the Assumption 2 and $\boldsymbol{u}'$ be an independent copy $\boldsymbol{u}$. The vector $\boldsymbol{u}$ has a bounded directional fourth moment, i.e., there exist $\eta \in [1,\, 2(4/\alpha)^{4/\alpha} K^4]$ such that*

$$\mathbb{E}\left(\left(\langle \boldsymbol{h}, \boldsymbol{u}\rangle\right)^4\right) \le \eta\,, \tag{9}$$

holds for all $\boldsymbol{h} \in \mathbb{S}^{p-1}$. Furthermore, for all $\boldsymbol{h}, \boldsymbol{h}' \in \mathbb{S}^{p-1}$ we have

$$\mathbb{E}\left(\left(\langle \boldsymbol{h}, \boldsymbol{u}\rangle + \langle \boldsymbol{h}', \boldsymbol{u}'\rangle\right)^4\right) \le \max\{\eta, 3\}\left(\|\boldsymbol{h}\|_2^2 + \|\boldsymbol{h}'\|_2^2\right)^2 .$$

Proof. Clearly, existence of the exponential moments guarantees that $\mathbb{E}\left(|\langle \boldsymbol{h}, \boldsymbol{u}\rangle|^4\right) < \infty$ for all $\boldsymbol{h} \in \mathbb{S}^{p-1}$. To prove the first part, we show that (9) holds for $\eta = 2(4/\alpha)^{4/\alpha} K^4$. For all $\boldsymbol{h} \in \mathbb{S}^{p-1}$ we have

$$\begin{aligned}\mathbb{E}\left(|\langle \boldsymbol{h}, \boldsymbol{u}\rangle|^4\right) &= \frac{\eta}{2}\,\mathbb{E}\left(\left(\frac{|\langle \boldsymbol{h}, \boldsymbol{u}\rangle|^\alpha}{(\eta/2)^{\alpha/4}}\right)^{4/\alpha}\right)\\ &\le \frac{\eta}{2}\,\mathbb{E}\left(\exp\left(\frac{4}{\alpha}\,\frac{|\langle \boldsymbol{h}, \boldsymbol{u}\rangle|^\alpha}{(\eta/2)^{\alpha/4}}\right)\right).\end{aligned}$$

For the prescribed η we have $(\eta/2)^{\alpha/4}\alpha/4 = K^\alpha$. Thus, in view of (8), we obtain

$$\mathbb{E}\left(|\langle \boldsymbol{h}, \boldsymbol{u}\rangle|^4\right) \le \frac{\eta}{2} 2 = \eta\,,$$

as desired. Since $\boldsymbol{u}$ and $\boldsymbol{u}'$ are zero-mean, isotropic, i.i.d, and further obey (9), we have

$$\begin{aligned}\mathbb{E}\left(\left(\langle \boldsymbol{h}, \boldsymbol{u}\rangle + \langle \boldsymbol{h}', \boldsymbol{u}'\rangle\right)^4\right) &= \mathbb{E}\left(\left(\langle \boldsymbol{h}, \boldsymbol{u}\rangle\right)^4 + 6\left(\langle \boldsymbol{h}, \boldsymbol{u}\rangle\right)^2\left(\langle \boldsymbol{h}', \boldsymbol{u}'\rangle\right)^2 + \left(\langle \boldsymbol{h}', \boldsymbol{u}'\rangle\right)^4\right)\\ &\le \eta\|\boldsymbol{h}\|_2^4 + 6\|\boldsymbol{h}\|_2^2\|\boldsymbol{h}'\|_2^2 + \eta\|\boldsymbol{h}'\|_2^4\\ &\le \max\{\eta, 3\}\left(\|\boldsymbol{h}\|_2^2 + \|\boldsymbol{h}'\|_2^2\right)^2 ,\end{aligned}$$

which proves the second part. □

In addition to the assumptions made above, our analysis crucially depends on a form of contraction that can be ensured by the following assumption. Note that Λ can be taken as $\Lambda = \mathrm{Lip}(\nabla f)$, i.e., the Lipschitz constant of ∇f with respect to the usual Euclidean metric.

Assumption 3 (conctractive dynamics)**.** *The nonlinearity ∇f and the matrix $\boldsymbol{A}_\star$ induce a contraction in the sense that*

$$\Lambda\,\|\boldsymbol{A}_\star\| < 1\,.$$

2. Main result

Our main theorem below, effectively guarantees that $T = \widetilde{O}(n+p)$ is sufficient for the matrix $\boldsymbol{\Sigma}$ to be (strictly) positive definite. Throughout, we denote the matrix $\boldsymbol{M}$ with its ith column replaced by the zero vector as $\boldsymbol{M}_{\setminus i}$. Furthermore, we denote the $\ell_1 \to \ell_2$ induced norm, or equivalently the largest column ℓ_2 norm, of $\boldsymbol{M}$, by

$$\|\boldsymbol{M}\|_{1\to 2} \overset{\text{def}}{=} \max\{\|\boldsymbol{M}\boldsymbol{z}\|_2 \colon \|z\|_1 \le 1\}\,.$$

Theorem 1. *Suppose that the energy of $\boldsymbol{B}_\star$ is well-spread among its columns in the sense that $\|\boldsymbol{B}_\star\|_{1\to 2}/\|\boldsymbol{B}_\star\|_{\mathrm{F}} = O(p^{-1/2})$. Furthermore, suppose that the constant*

$$\begin{aligned}\theta = \theta_{\alpha,\beta,\varepsilon,\lambda,K,\boldsymbol{B}_\star} \overset{\text{def}}{=} & -\varepsilon K \log^{\frac{1}{\alpha}}\left(10\max\{\eta,3\}\right)\|\boldsymbol{B}_\star\|_{1\to 2} \\ & + 0.6 \min_{i=1,\dots,p} \min\left\{\beta, (\lambda-\varepsilon)\lambda_{\min}^{1/2}\left(\boldsymbol{B}_{\star\setminus i}\boldsymbol{B}_{\star\setminus i}^\top\right)\right\}\,,\end{aligned} \tag{10}$$

is strictly positive. Furthermore, suppose that L satisfies

$$L \geq 1 + \frac{\log\left(\frac{c^2}{\theta^2}\log\left(\frac{2(T-1)(p+1)}{\delta}\right)\left(\frac{\mathit{\Lambda}\|\boldsymbol{B}_\star\|_{\mathrm{F}}}{1-\mathit{\Lambda}\|\boldsymbol{A}_\star\|}\right)^2\right)}{\log\frac{1}{\mathit{\Lambda}\|\boldsymbol{A}_\star\|}}\,, \tag{11}$$

for a sufficiently large constant $c > 0$. Then, for

$$T \gtrsim \max\{\eta^2, 9\}\,(n+p)L\log\left(\frac{eT/L}{n+p}\right) + \log\left(\frac{8L}{\delta}\right)\,, \tag{12}$$

we have

$$\lambda_{\min}(\boldsymbol{\Sigma}) \gtrsim \frac{\theta^2}{\max\{\eta,3\}}T\,,$$

with probability $\geq 1-\delta$. Consequently, on the same event, (4) recovers $\boldsymbol{C}_\star$, exactly.

A critical condition of Theorem 1 is that θ is strictly positive. This condition implicitly requires ε in (7) to be sufficiently small, which in turn implies the condition number of f is sufficiently close to 1 (i.e. ∇f is nearly linear). Furthermore, it is needed that the energy of $\boldsymbol{B}_\star$ to be well-spread not only among its columns, but also in a "spectral" sense. More precisely, we need the quantity

$$\max_{i=1,\dots,p} \frac{\|\boldsymbol{B}_\star\|_{1\to 2}}{\lambda_{\min}^{1/2}(\boldsymbol{B}_{\star\setminus i}\boldsymbol{B}_{\star\setminus i}^\top)}\,,$$

to be sufficiently small. The equation (10) also suggests that a reasonable choice of the normalizing constant β should satisfy $\beta \approx (\lambda-\varepsilon)\lambda_{\min}^{1/2}(\boldsymbol{B}_\star\boldsymbol{B}_\star^\top)$.

3. Simulation

We evaluated the proposed estimator numerically on synthetic data in a setup similar to the experiments of (Oymak, 2019). In all of the experiments, we consider the dimensions to be $n = 50$, $p = 100$, and the time horizon to be $T = 500$. For $\alpha \in \{0.2, 0.8\}$ we choose $\boldsymbol{A}_\star = \alpha\boldsymbol{R}$ with $\boldsymbol{R}$ being a uniformly distributed $n \times n$ orthogonal matrix. Furthermore, $\boldsymbol{B}_\star \in \mathbb{R}^{n\times p}$ is generated randomly with i.i.d. standard normal entries. The normalizing factor β is chosen as prescribed in (Oymak, 2019). We consider two different models for the input $\boldsymbol{u}$. Let $g \sim \mathrm{Normal}(0,1)$ denote a standard Normal random scalar. The first model is similar to the model of (Oymak, 2019) where the entries of $\boldsymbol{u}$ are i.i.d. copies of g, whereas in the second model takes i.i.d. copies of g^3 as the entries of $\boldsymbol{u}$. We refer to these models as the Gaussian model and the heavy-tailed model, respectively. The nonlinearity in (1) is described by one of the functions

$$f(\boldsymbol{x}) = \frac{1-\rho}{2}\sum_{i=1}^{n}(x_i)_+^2 + \frac{\rho}{2}\sum_{i=1}^{n} x_i^2\,,$$

at $\rho = 1$ (i.e., linear activation), $\rho = 0.5$ (i.e., leaky ReLU activation with slope 0.5 over $\mathbb{R}_{\le 0}$), $\rho = 0.3$ (i.e., leaky ReLU activation with slope 0.3 over $\mathbb{R}_{\le 0}$), and $\rho = 0$ (i.e., ReLU activation).

For each choice of α and ρ, we solved (4) using Nesterov's Accelerated Gradient Method (AGM) (Nesterov, 1983; Nesterov, 2013, Section 2.2), for 100 randomly generated instances of the problem. For the Gaussian model the step-size is set to 10^{-3}, whereas for the heavy-tailed model the step-size is set to 10^{-4}. In each trial, the AGM is run for a maximum of 500 iterations and terminated only if the relative error dropped below 10^{-8} (i.e., $\left\|\widehat{\boldsymbol{C}} - \boldsymbol{C}_\star\right\|_{\mathrm{F}}^2 / \|\boldsymbol{C}_\star\|_{\mathrm{F}}^2 \le 10^{-8}$). The optimization task can be solved by the SGD as well. However, slower convergence of the SGD is only tolerable for large-scale problems where lower memory load is crucial. Nevertheless, because the estimator (4) is formulated as a convex program, we can apply the SGD methods with variance reduction (see e.g., Johnson and Zhang, 2013; Schmidt et al., 2017; Defazio et al., 2014) and rely on their theoretical guarantees.

Figures 1 and 2 depict the achieved relative error under the Gaussian model and the heavy-tailed model for the chosen values of α and ρ, respectively. The solid lines show the median of the achieved relative error, whereas the dashed lines show the 0.1 and 0.9 quantiles of the relative error. Perhaps, the result that might strike as counter intuitive at first, is that the estimation performance is not monotonic with respect to the strength of stability. For instance, the plots in the first two rows of Figure 1 suggest that convergence is faster for the less stable system (i.e., $\alpha = 0.8$). A similar conclusion can be made regarding the plots in the first row of Figure 2 corresponding to linear activation functions. However, it appears that this behavior is sensitive to the level of nonlinearity, particularly in the case of the heavy-tailed input distributions.

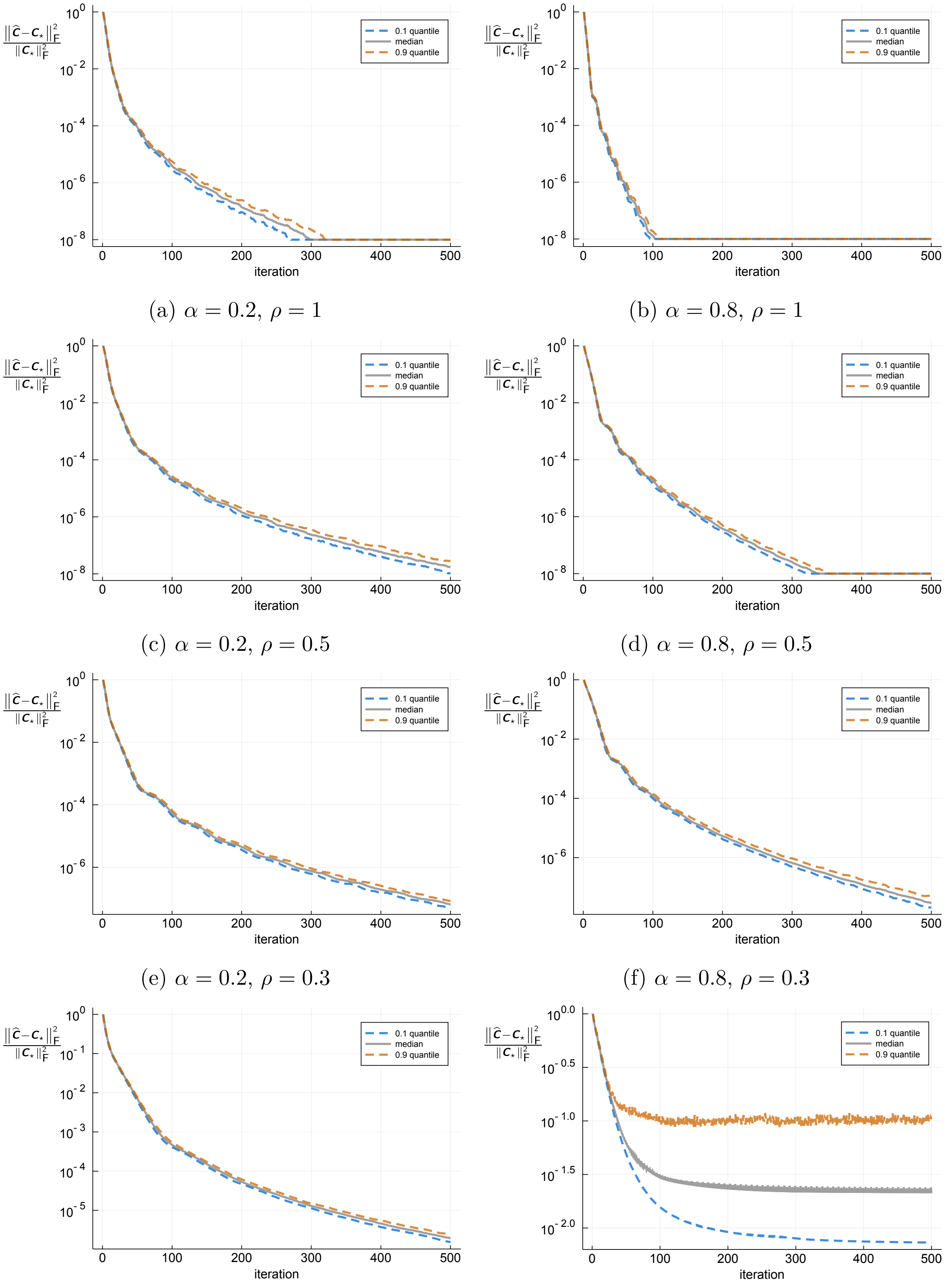
4. Proof of the main result

Proof of Theorem 1. Recall the definition of $\boldsymbol{\Sigma}$ in (5). We would like to find a lower bound for the smallest eigenvalue of $\boldsymbol{\Sigma}$ that holds with high probability. Consider a sufficiently large integer L as a *stride* parameter and for $\ell = 0, \dots, L-1$ let

$$\mathcal{T}_\ell = \{t\,:\, L \le t < T \text{ and } t = \ell \bmod L\}\,,$$

which partition $\{L, \dots, T-1\}$ to sets of subsampled time indices with stride L. For each $\ell = 0, \dots, L-1$, we define the "restarted" state variables $\boldsymbol{x}_t^{(\ell)}$ through the recursion

$$\boldsymbol{x}_{t+1}^{(\ell)} = \begin{cases} \boldsymbol{0} & , t = \ell \bmod L \\ \nabla f\left(\boldsymbol{A}_\star \boldsymbol{x}_t^{(\ell)} + \boldsymbol{B}_\star \boldsymbol{u}_t\right) & , t \ne \ell \bmod L\,, \end{cases}$$



(a) $\alpha = 0.2,\ \rho = 1$

(b) $\alpha = 0.8,\ \rho = 1$

(c) $\alpha = 0.2,\ \rho = 0.5$

(d) $\alpha = 0.8,\ \rho = 0.5$

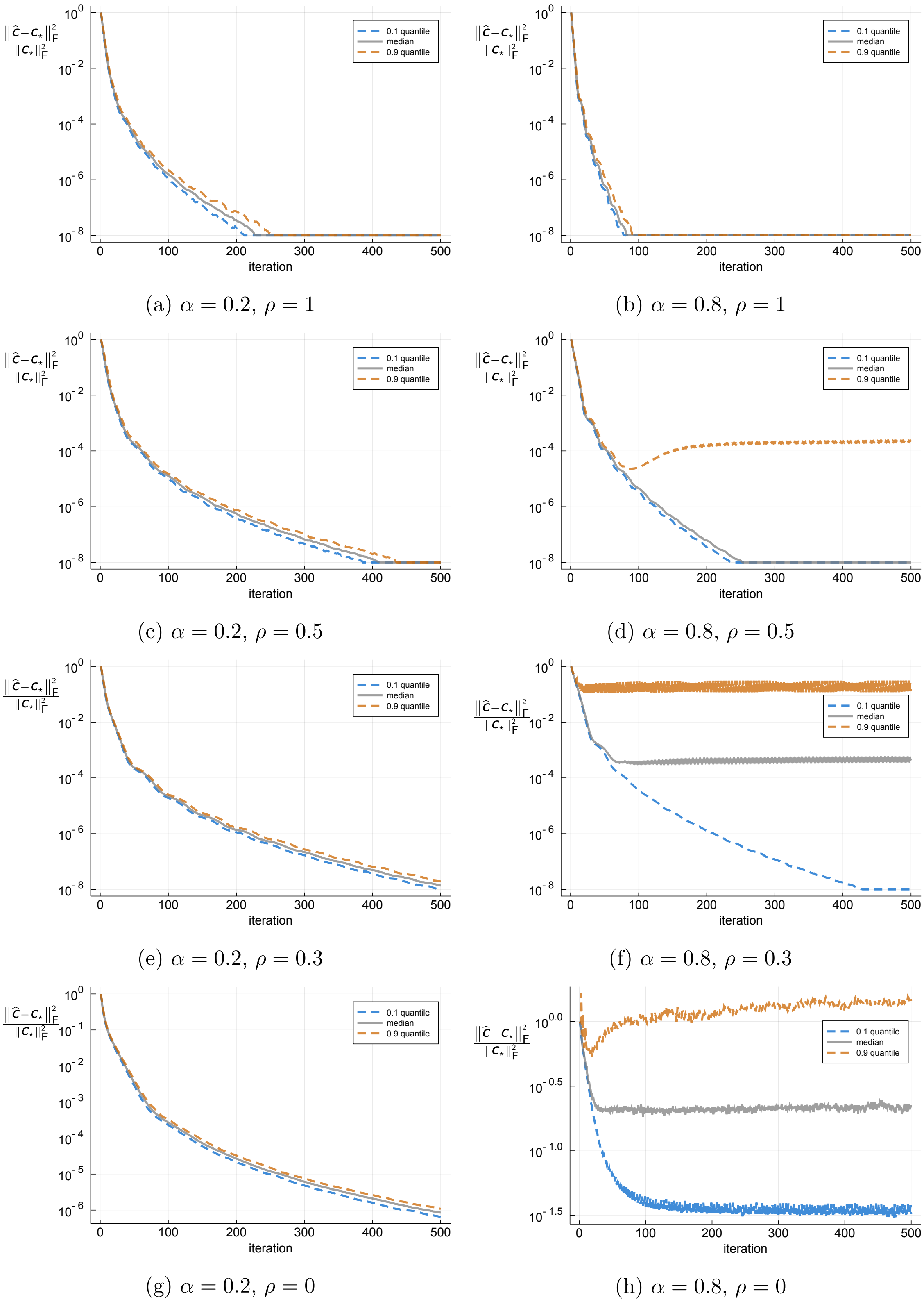
(e) $\alpha = 0.2,\ \rho = 0.3$

(f) $\alpha = 0.8,\ \rho = 0.3$

(g) $\alpha = 0.2,\ \rho = 0$

(h) $\alpha = 0.8,\ \rho = 0$

Figure 1: Gaussian model



(a) $\alpha = 0.2,\ \rho = 1$

(b) $\alpha = 0.8,\ \rho = 1$

(c) $\alpha = 0.2,\ \rho = 0.5$

(d) $\alpha = 0.8,\ \rho = 0.5$

(e) $\alpha = 0.2,\ \rho = 0.3$

(f) $\alpha = 0.8,\ \rho = 0.3$

(g) $\alpha = 0.2,\ \rho = 0$

(h) $\alpha = 0.8,\ \rho = 0$

Figure 2: Heavy-tailed model

and the corresponding restarted version of $\boldsymbol{z}_t$ as

$$\boldsymbol{z}_t^{(\ell)} = \begin{bmatrix} \boldsymbol{x}_t^{(\ell)} \\ \beta\, \boldsymbol{u}_t \end{bmatrix} . \tag{13}$$

For any $\boldsymbol{w} \in \mathbb{S}^{n+p-1}$ we have

$$\boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w} \geq \sum_{t=L}^{T-1} \left(\boldsymbol{w}^\top \boldsymbol{z}_t\right)^2 = \sum_{\ell=0}^{L-1} \sum_{t \in \mathcal{T}_\ell} \left(\boldsymbol{w}^\top \boldsymbol{z}_t\right)^2 .$$

To find a lower bound for $\sum_{t \in \mathcal{T}_\ell} \left(\boldsymbol{w}^\top \boldsymbol{z}_t\right)^2$, the strategy is to approximate this summation by its corresponding restarted version. Aggregating the obtained bounds for all $\ell = 0, \dots, L-1$ then yields the desired lower bound for $\boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w}$.

By the Cauchy-Schwarz inequality we have

$$\left(\boldsymbol{w}^\top \boldsymbol{z}_t\right)^2 + \left(\boldsymbol{w}^\top \left(\boldsymbol{z}_t - \boldsymbol{z}_t^{(\ell)}\right)\right)^2 \geq \frac{1}{2} \left(\boldsymbol{w}^\top \boldsymbol{z}_t^{(\ell)}\right)^2 .$$

Summing over $t \in \mathcal{T}_\ell$ and rearranging the terms then yields

$$\sum_{t \in \mathcal{T}_\ell} \left(\boldsymbol{w}^\top \boldsymbol{z}_t\right)^2 \geq \frac{1}{2} \underbrace{\sum_{t \in \mathcal{T}_\ell} \left(\boldsymbol{w}^\top \boldsymbol{z}_t^{(\ell)}\right)^2}_{\overset{\text{def}}{=} S_\ell(\boldsymbol{w})} - \underbrace{\sum_{t \in \mathcal{T}_\ell} \left(\boldsymbol{w}^\top \left(\boldsymbol{z}_t - \boldsymbol{z}_t^{(\ell)}\right)\right)^2}_{\overset{\text{def}}{=} \widetilde{S}_\ell(\boldsymbol{w})} .$$

Observe that the term $S_\ell\left(\boldsymbol{w}\right)$ is a sum of independent random quadratic functions. Therefore, deriving a uniform lower bound for $S_\ell\left(\boldsymbol{w}\right)$ is amenable to standard techniques. We also need to establish a uniform upper bound for the term $\widetilde{S}_\ell\left(\boldsymbol{w}\right)$ for which we leverage the contraction assumption.

Recall that we denote the matrix $\boldsymbol{M}$ with its ith column replaced by the zero vector as $\boldsymbol{M}_{\setminus i}$. The following lemma, whose proof is relegated to the appendix, provides a uniform lower bound on $\sum_{\ell=0}^{L-1} S_\ell(\boldsymbol{w})$. The proof for this lemma is also provided in the appendix.

Lemma 3 (uniform lower bound for $S_\ell\left(\boldsymbol{w}\right)$)**.** *With probability $\geq 1 - \delta$, for all $\boldsymbol{w} \in \mathbb{S}^{n+p-1}$ we have*

$$\sum_{\ell=0}^{L-1} S_\ell(\boldsymbol{w}) \geq \theta^2 L |\mathcal{T}_\ell| \left(\frac{0.1}{\max\{\eta, 3\}} - \sqrt{\frac{2(n+p)\log\frac{eT/L}{(n+p)} + \log\frac{4L}{\delta}}{|\mathcal{T}_\ell|}} \right) ,$$

where θ is defined as in (10).

Furthermore, we have the following lemma that establishes a uniform upper bound for $\sum_{\ell=0}^{L-1} \widetilde{S}_\ell(\boldsymbol{w})$.

Lemma 4 (uniform upper bound for $\widetilde{S}_\ell\left(\boldsymbol{w}\right)$)**.** *Suppose that $\mu \overset{\text{def}}{=} p^{1/2}\|\boldsymbol{B}_\star\|_{1\to 2}/\|\boldsymbol{B}_\star\|_{\mathrm{F}} = O(1)$ and let $\epsilon > 0$ be a parameter. If for a certain absolute constant $c > 0$, we have*

$$L \geq 1 + \frac{\log\left(\frac{c^2 T}{\epsilon} \log\left(\frac{2(T-1)(p+1)}{\delta}\right) \left(\frac{\Lambda\|\boldsymbol{B}_\star\|_{\mathrm{F}}}{1-\Lambda\|\boldsymbol{A}_\star\|}\right)^2\right)}{\log\frac{1}{\Lambda\|\boldsymbol{A}_\star\|}} ,$$

then with probability $\geq 1-\delta$, *we can guarantee*

$$\sum_{\ell=0}^{L-1} \widetilde{S}_\ell(\boldsymbol{w}) \leq \epsilon\,.$$

Consequently, under (11) and (12), it follows from Lemmas 3 and 4 that

$$\boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w} \gtrsim \frac{\theta^2}{\max\{\eta, 3\}} T\,.$$

holds uniformly for all $\boldsymbol{w} \in \mathbb{S}^{n+p-1}$ with probability $\geq 1-\delta$. □

Acknowledgements

This work was supported in part by the Semiconductor Research Corporation (SRC) and DARPA.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Can sgd learn recurrent neural networks with provable generalization? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 10331–10341. Curran Associates, Inc., 2019.

Víctor H de la Peña and Evarist Giné. *Decoupling: From dependence to independence.* Probability and its Applications. Springer-Verlag, New York, 1999.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems 27*, pages 1646–1654. 2014.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Simon S Du, Yining Wang, Xiyu Zhai, Sivaraman Balakrishnan, Ruslan Salakhutdinov, and Aarti Singh. How many samples are needed to estimate a convolutional neural network? preprint arXiv:1805.07883 [stat.ML], 2018.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, 2018.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26*, pages 315–323. 2013.

V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems.* Lecture Notes in Mathematics: École d'Été de Probabilités de Saint-Flour XXXVIII-2008. Springer-Verlag Berlin Heidelberg, 2011.

John Miller and Moritz Hardt. Stable recurrent models. In *International Conference on Learning Representations*, 2019.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2013.

Yurii E Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

Samet Oymak. Stochastic gradient descent learns state equations with nonlinear activations. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2551–2579, Phoenix, USA, 25–28 Jun 2019. PMLR.

Samet Oymak and Necmiye Ozay. Non-asymptotic identification of LTI systems from a single trajectory. preprint arXiv:1806.05722 [cs.LG], 2018.

R. E. A. C. Paley and A. Zygmund. A note on analytic functions in the unit circle. *Mathematical Proceedings of the Cambridge Philosophical Society*, 28(3):266–272, 1932.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5610–5618, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112, Mar 2017.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I. Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 439–473. PMLR, 06–09 Jul 2018.

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

Appendix A. Proofs for technical lemmas

Proof of Lemma 3. For each $\ell = 0, \dots, L-1$, the vectors $\boldsymbol{z}_t^{(\ell)}$ with $t \in \mathcal{T}_\ell$ are independent and identically distributed. Let $\theta > 0$ be a parameter to be specified later. Using a simple truncation we can write

$$\sum_{t\in\mathcal{T}_\ell} \left(\boldsymbol{w}^\intercal \boldsymbol{z}_t^{(\ell)}\right)^2 \ge \theta^2 \sum_{t\in\mathcal{T}_\ell} \mathbb{1}\left(\left|\boldsymbol{w}^\intercal \boldsymbol{z}_t^{(\ell)}\right| \ge \theta\right) .$$

To bound the right-hands side of the inequality above uniformly with respect to the set of binary functions

$$\mathcal{F}_\ell \stackrel{\text{def}}{=} \left\{\boldsymbol{z} \mapsto \mathbb{1}\left(\left|\boldsymbol{w}^\intercal \boldsymbol{z}\right| \ge \theta\right) \,:\, \boldsymbol{w} \in \mathbb{S}^{n+p-1}\right\} ,$$

we can resort to classic VC bounds (Vapnik and Chervonenkis, 1971; see also Devroye et al., 2013, chapters 13 & 14). Particularly, because the VC dimension of $\mathcal{F}_\ell$ is no more than $2\,(n+p)$, with probability $\ge 1 - \delta/L$ we have

$$\frac{1}{|\mathcal{T}_\ell|} \sum_{t\in\mathcal{T}_\ell} \mathbb{1}\left(\left|\boldsymbol{w}^\intercal \boldsymbol{z}_t^{(\ell)}\right| \ge \theta\right) \ge \mathbb{P}\left(\left|\boldsymbol{w}^\intercal \boldsymbol{z}_t^{(\ell)}\right| \ge \theta\right) - \sqrt{\frac{2(n+p)\log\frac{e|\mathcal{T}_\ell|}{n+p} + \log\frac{4L}{\delta}}{|\mathcal{T}_\ell|}} ,$$

for all $\boldsymbol{w} \in \mathbb{S}^{n+p-1}$. It only remains to find appropriate lower bounds for the probability in the summation. Lemma 6 below provides the needed lower bound.

Taking the union bound over ℓ then shows that with probability $\geq 1-\delta$ we obtain

$$\sum_{\ell=0}^{L-1} \frac{1}{|\mathcal{T}_\ell|} \sum_{t \in \mathcal{T}_\ell} \mathbb{1}\left(\left|\boldsymbol{w}^\mathsf{T} \boldsymbol{z}_t^{(\ell)}\right| \geq \theta\right) \geq L\left(\frac{0.1}{\max\{\eta, 3\}} - \sqrt{\frac{2(n+p)\log\frac{e|\mathcal{T}_\ell|}{n+p} + \log\frac{4L}{\delta}}{|\mathcal{T}_\ell|}}\right),$$

which yields the desired bound. □

Proof of Lemma 4. Recall the definition of $\boldsymbol{z}_t^{(\ell)}$ in (13). For every $t \in \mathcal{T}_\ell$ and $\boldsymbol{w} \in \mathbb{S}^{n+p-1}$ we have

$$\begin{aligned}\left(\boldsymbol{w}^\mathsf{T}\left(\boldsymbol{z}_t - \boldsymbol{z}_t^{(\ell)}\right)\right)^2 &\leq \left\|\boldsymbol{z}_t - \boldsymbol{z}_t^{(\ell)}\right\|_2^2 \\ &= \left\|\nabla f\left(\boldsymbol{C}_\star \boldsymbol{z}_{t-1}\right) - \nabla f\left(\boldsymbol{C}_\star \boldsymbol{z}_{t-1}^{(\ell)}\right)\right\|_2^2.\end{aligned}$$

Furthermore, we can write

$$\begin{aligned}\left\|\nabla f\left(\boldsymbol{C}_\star \boldsymbol{z}_{t-1}\right) - \nabla f\left(\boldsymbol{C}_\star \boldsymbol{z}_{t-1}^{(\ell)}\right)\right\|_2^2 &\leq \Lambda^2 \left\|\boldsymbol{C}_\star\left(\boldsymbol{z}_{t-1} - \boldsymbol{z}_{t-1}^{(\ell)}\right)\right\|_2^2 \\ &= \Lambda^2 \left\|\boldsymbol{A}_\star\left(\nabla f\left(\boldsymbol{C}_\star \boldsymbol{z}_{t-2}\right) - \nabla f\left(\boldsymbol{C}_\star \boldsymbol{z}_{t-2}^{(\ell)}\right)\right)\right\|_2^2 \\ &\leq \left(\Lambda\|\boldsymbol{A}_\star\|\right)^2 \left\|\left(\nabla f\left(\boldsymbol{C}_\star \boldsymbol{z}_{t-2}\right) - \nabla f\left(\boldsymbol{C}_\star \boldsymbol{z}_{t-2}^{(\ell)}\right)\right)\right\|_2^2.\end{aligned}$$

Using the above inequality recursively yields

$$\begin{aligned}\left\|\nabla f\left(\boldsymbol{C}_\star \boldsymbol{z}_{t-1}\right) - \nabla f\left(\boldsymbol{C}_\star \boldsymbol{z}_{t-1}^{(\ell)}\right)\right\|_2^2 &\leq \left(\Lambda\|\boldsymbol{A}_\star\|\right)^{2(L-2)} \left\|\nabla f\left(\boldsymbol{C}_\star \boldsymbol{z}_{t-L+1}\right) - \nabla f\left(\boldsymbol{C}_\star \boldsymbol{z}_{t-L+1}^{(\ell)}\right)\right\|_2^2 \\ &\leq \left(\Lambda\|\boldsymbol{A}_\star\|\right)^{2(L-1)} \left\|\boldsymbol{x}_{t-L}\right\|_2^2.\end{aligned}$$

Therefore, we deduce that

$$\left(\boldsymbol{w}^\mathsf{T}\left(\boldsymbol{z}_t - \boldsymbol{z}_t^{(\ell)}\right)\right)^2 \leq \left(\Lambda\|\boldsymbol{A}_\star\|\right)^{2(L-1)} \left\|\boldsymbol{x}_{t-L}\right\|_2^2. \tag{14}$$

Furthermore, for any time index $s \geq 1$ we have

$$\begin{aligned}\|\boldsymbol{x}_s\|_2 &\leq \Lambda\|\boldsymbol{A}_\star \boldsymbol{x}_{s-1} + \boldsymbol{B}_\star \boldsymbol{u}_{s-1}\|_2 \\ &\leq \Lambda\|\boldsymbol{A}_\star\|\|\boldsymbol{x}_{s-1}\|_2 + \Lambda\|\boldsymbol{B}_\star \boldsymbol{u}_{s-1}\|_2.\end{aligned}$$

Therefore, we can write

$$\max_{1\leq s\leq T-1} \|\boldsymbol{x}_s\|_2 \leq \Lambda\|\boldsymbol{A}_\star\| \max_{1\leq s\leq T-1} \|\boldsymbol{x}_{s-1}\|_2 + \Lambda \max_{1\leq s\leq T-1} \|\boldsymbol{B}_\star \boldsymbol{u}_{s-1}\|_2,$$

which implies

$$\max_{1\leq s\leq T-1} \|\boldsymbol{x}_s\|_2 \leq \frac{\Lambda}{1-\Lambda\|\boldsymbol{A}_\star\|} \max_{1\leq s\leq T-1} \|\boldsymbol{B}_\star \boldsymbol{u}_{s-1}\|_2.$$

Since $\mu = p^{1/2}\|\boldsymbol{B}_\star\|_{1\to 2}/\|\boldsymbol{B}_\star\|_{\mathrm{F}} = O(1)$ by assumption, using the matrix Bernstein inequality, stated in Lemma 5 below, for each $s = 1,\dots,T-1$, with probability $\geq 1-\delta/(T-1)$ we have

$$\|\boldsymbol{B}_\star \boldsymbol{u}_{s-1}\|_2 \leq c\|\boldsymbol{B}_\star\|_{\mathrm{F}} \log^{\frac{1}{2}}\left(\frac{2(T-1)(p+1)}{\delta}\right) ,$$

for some absolute constant $c > 0$. It then follows from a simple union bound that

$$\max_{1\leq s\leq T-1} \|\boldsymbol{B}_\star \boldsymbol{u}_{s-1}\|_2 \leq c\|\boldsymbol{B}_\star\|_{\mathrm{F}} \log^{\frac{1}{2}}\left(\frac{2(T-1)(p+1)}{\delta}\right) ,$$

holds with probability $\geq 1-\delta$. Consequently,

$$\max_{1\leq s\leq T-1} \|\boldsymbol{x}_s\|_2 \leq c\log^{\frac{1}{2}}\left(\frac{2(T-1)(p+1)}{\delta}\right)\frac{\Lambda\|\boldsymbol{B}_\star\|_{\mathrm{F}}}{1-\Lambda\|\boldsymbol{A}_\star\|} ,$$

holds with probability $\geq 1-\delta$. Under the same event and in view of (14) we have

$$\left(\boldsymbol{w}^\mathsf{T}\left(\boldsymbol{z}_t - \boldsymbol{z}_t^{(\ell)}\right)\right)^2 \leq (\Lambda\|\boldsymbol{A}_\star\|)^{2(L-1)}\, c^2 \log\left(\frac{2(T-1)(p+1)}{\delta}\right)\left(\frac{\Lambda\|\boldsymbol{B}_\star\|_{\mathrm{F}}}{1-\Lambda\|\boldsymbol{A}_\star\|}\right)^2 ,$$

for all $\boldsymbol{w}\in\mathbb{S}^{n+p-1}$, $0\leq \ell\leq L-1$, and $t\in\mathcal{T}_\ell$. Summation over $t\in\mathcal{T}_\ell$ then yields

$$\begin{aligned}\widetilde{S}_\ell(\boldsymbol{w}) &= \sum_{t\in\mathcal{T}_\ell}\left(\boldsymbol{w}^\mathsf{T}\left(\boldsymbol{z}_t - \boldsymbol{z}_t^{(\ell)}\right)\right)^2\\ &\leq \frac{T}{L}(\Lambda\|\boldsymbol{A}_\star\|)^{2(L-1)}\, c^2 \log\left(\frac{2(T-1)(p+1)}{\delta}\right)\left(\frac{\Lambda\|\boldsymbol{B}_\star\|_{\mathrm{F}}}{1-\Lambda\|\boldsymbol{A}_\star\|}\right)^2 .\end{aligned}$$

Therefore, for $\epsilon > 0$ if

$$L \geq 1 + \frac{\log\left(\frac{c^2 T}{\epsilon}\log\left(\frac{2(T-1)(p+1)}{\delta}\right)\left(\frac{\Lambda\|\boldsymbol{B}_\star\|_{\mathrm{F}}}{1-\Lambda\|\boldsymbol{A}_\star\|}\right)^2\right)}{\log\frac{1}{\Lambda\|\boldsymbol{A}_\star\|}} ,$$

then with probability $\geq 1-\delta$ for all $\boldsymbol{w}\in\mathbb{S}^{n+p-1}$ we have

$$\sum_{\ell=0}^{L-1}\widetilde{S}_\ell(\boldsymbol{w}) \leq \epsilon .$$

□

Appendix B. Auxiliary lemmas

We use a special case of a matrix Bernstein inequality (Koltchinskii, 2011, Corollary 2.1). For reference, the following lemma states the special inequality we need; we omit the proof and refer the reader to (Koltchinskii, 2011) for the general Bernstein inequality.

Lemma 5. *Suppose that $\boldsymbol{u}$ obeys the Assumption 2. Furthermore, define a coherence parameter for $\boldsymbol{B}_\star$ as $\mu \overset{\text{def}}{=} p^{1/2}\|\boldsymbol{B}_\star\|_{1\to 2}/\|\boldsymbol{B}_\star\|_{\mathrm{F}}$. Then, for some absolute constant $c > 0$, and any $\gamma \in (0, 1]$, the bound*

$$\begin{aligned}&\|\boldsymbol{B}_\star\boldsymbol{u}\|_2\\ &\le \max\left\{c^{\frac{1}{2}}\log^{\frac{1}{2}}\left(2\gamma^{-1}(p+1)\right),\ c\max\{K,2\}\mu\,\log^{\frac{1}{\alpha}}\left(\max\{K,2\}\mu\right)\,\frac{\log\left(2\gamma^{-1}(p+1)\right)}{p^{1/2}}\right\}\|\boldsymbol{B}_\star\|_{\mathrm{F}}\,,\end{aligned}$$

holds with probability $\ge 1-\gamma$. In particular, if $\mu = O(1)$, meaning that the weight of $\boldsymbol{B}_\star$ is distributed almost evenly across its columns, and p is sufficiently large, the bound stated above effectively reduces to

$$\|\boldsymbol{B}_\star\boldsymbol{u}\|_2 \le c\|\boldsymbol{B}_\star\|_{\mathrm{F}}\log^{\frac{1}{2}}\left(2\gamma^{-1}(p+1)\right)\,,$$

for some absolute constant $c > 0$.

In general, the coherence parameter μ defined in Lemma 5 obeys $1 \le \mu \le p^{1/2}$. However, we assume we operate in the scenario that $\mu = O(1)$ so that we apply the simpler bound stated in the lemma. Therefore, choosing $\gamma = 1/p$ and for a sufficiently large p we have

$$\begin{aligned}&\mathbb{P}\left(\left|\boldsymbol{w}^\intercal\begin{bmatrix}F(\boldsymbol{A}_\star\boldsymbol{x}_{t-1}^{(\ell)})\boldsymbol{B}_\star\boldsymbol{u}_{t-1}\\ \beta\,\boldsymbol{u}_t\end{bmatrix}\right| - \varepsilon\|\boldsymbol{B}_\star\boldsymbol{u}_{t-1}\|_2 \ge \theta\right)\\ &\ge \mathbb{P}\left(\left|\boldsymbol{w}^\intercal\begin{bmatrix}F(\boldsymbol{A}_\star\boldsymbol{x}_{t-1}^{(\ell)})\boldsymbol{B}_\star\boldsymbol{u}_{t-1}\\ \beta\,\boldsymbol{u}_t\end{bmatrix}\right| \ge \theta + c\varepsilon\log^{\frac{1}{2}}\left(2p(p+1)\right)\|\boldsymbol{B}_\star\|_{\mathrm{F}}\right) - \frac{1}{p}\,.\end{aligned}$$

for some absolute constant $c > 0$.

Lemma 6 (lower bound for the probabilities)**.** *With θ defined as in (10), for each $\ell \in \{0, 1, \ldots, L-1\}$, and every $t \in \mathcal{T}_\ell$ we have*

$$\mathbb{P}\left(\left|\boldsymbol{w}^\intercal\boldsymbol{z}_t^{(\ell)}\right| \ge \theta\right) \ge \frac{0.1}{\max\{\eta, 3\}}$$

Proof. For $t = 0, 1, \ldots$, let i_t be i.i.d. integers uniformly distributed over $\{1, \ldots, p\}$, independent of everything else. For any vector $\boldsymbol{v}$, we use the notation $\boldsymbol{v}^{-i}$ to denote the vector obtained by flipping the sign of the ith coordinate of $\boldsymbol{v}$. Furthermore, for $t \in \mathcal{T}_\ell$ let

$$\overline{\boldsymbol{z}}_t^{(\ell)} = \begin{bmatrix}\nabla f(\boldsymbol{A}_\star\boldsymbol{x}_{t-1}^{(\ell)} + \boldsymbol{B}_\star\boldsymbol{u}_{t-1}^{-i_{t-1}}\\ -\beta\,\boldsymbol{u}_t\end{bmatrix}.$$

Recall that, by assumption, $\boldsymbol{u}_{t-1}$ and $\boldsymbol{u}_t$ have coordinates with independent symmetric distributions. Therefore, it is straightforward to show that $\boldsymbol{z}_t^{(\ell)}$ and $\overline{\boldsymbol{z}}_t^{(\ell)}$ are identically distributed, and for any $\theta > 0$ we can write

$$\begin{aligned}\mathbb{P}\left(\left|\boldsymbol{w}^\intercal\boldsymbol{z}_t^{(\ell)}\right| \ge \theta\right) &= \frac{1}{2}\mathbb{P}\left(\left|\boldsymbol{w}^\intercal\boldsymbol{z}_t^{(\ell)}\right| \ge \theta\right) + \frac{1}{2}\mathbb{P}\left(\left|\boldsymbol{w}^\intercal\overline{\boldsymbol{z}}_t^{(\ell)}\right| \ge \theta\right)\\ &\ge \frac{1}{2}\mathbb{P}\left(\left|\boldsymbol{w}^\intercal\boldsymbol{z}_t^{(\ell)}\right| + \left|\boldsymbol{w}^\intercal\overline{\boldsymbol{z}}_t^{(\ell)}\right| \ge 2\theta\right).\end{aligned}$$

Then, it follows from the triangle inequality, and the assumption (7), that

$$\begin{aligned}\mathbb{P}\left(\left|\boldsymbol{w}^\top\boldsymbol{z}_t^{(\ell)}\right|\ge\theta\right) &\ge \frac{1}{2}\,\mathbb{P}\left(\left|\boldsymbol{w}^\top\left(\boldsymbol{z}_t^{(\ell)}-\overline{\boldsymbol{z}}_t^{(\ell)}\right)\right|\ge 2\theta\right)\\
&\ge \frac{1}{2}\,\mathbb{P}\left(\left|\boldsymbol{w}^\top\left[\underset{2\beta\,\boldsymbol{u}_t}{\nabla f(\boldsymbol{A}_\star\boldsymbol{x}_{t-1}^{(\ell)}+\boldsymbol{B}_\star\boldsymbol{u}_{t-1})-\nabla f(\boldsymbol{A}_\star\boldsymbol{x}_{t-1}^{(\ell)}+\boldsymbol{B}_\star\boldsymbol{u}_{t-1}^{-i_{t-1}})}\right]\right|\ge 2\theta\right)\\
&\ge \frac{1}{2}\,\mathbb{P}\left(\left|\boldsymbol{w}^\top\left[\underset{\beta\,\boldsymbol{u}_t}{F(\boldsymbol{A}_\star\boldsymbol{x}_{t-1}^{(\ell)})\boldsymbol{B}_\star\left(\tfrac{1}{2}\boldsymbol{u}_{t-1}+\tfrac{1}{2}\boldsymbol{u}_{t-1}^{-i_{t-1}}\right)}\right]\right|-\varepsilon\left\|\boldsymbol{B}_\star\left(\frac{1}{2}\boldsymbol{u}_{t-1}-\frac{1}{2}\boldsymbol{u}_{t-1}^{-i_{t-1}}\right)\right\|_2\ge\theta\right).\end{aligned}\tag{15}$$

Furthermore, for any $\gamma\in(0,1]$ we can write

$$\begin{aligned}&\mathbb{P}\left(\left|\boldsymbol{w}^\top\left[\underset{\beta\,\boldsymbol{u}_t}{F(\boldsymbol{A}_\star\boldsymbol{x}_{t-1}^{(\ell)})\boldsymbol{B}_\star\left(\tfrac{1}{2}\boldsymbol{u}_{t-1}+\tfrac{1}{2}\boldsymbol{u}_{t-1}^{-i_{t-1}}\right)}\right]\right|-\varepsilon\left\|\boldsymbol{B}_\star\left(\frac{1}{2}\boldsymbol{u}_{t-1}-\frac{1}{2}\boldsymbol{u}_{t-1}^{-i_{t-1}}\right)\right\|_2\ge\theta\right)\\
&+\mathbb{P}\left(\left\|\boldsymbol{B}_\star\left(\frac{1}{2}\boldsymbol{u}_{t-1}-\frac{1}{2}\boldsymbol{u}_{t-1}^{-i_{t-1}}\right)\right\|_2\ge K\log^{\frac{1}{\alpha}}\left(\frac{2}{\gamma}\right)\|\boldsymbol{B}_\star\|_{1\to2}\right)\\
&\ge\mathbb{P}\left(\left|\boldsymbol{w}^\top\left[\underset{\beta\,\boldsymbol{u}_t}{F(\boldsymbol{A}_\star\boldsymbol{x}_{t-1}^{(\ell)})\boldsymbol{B}_\star\left(\tfrac{1}{2}\boldsymbol{u}_{t-1}+\tfrac{1}{2}\boldsymbol{u}_{t-1}^{-i_{t-1}}\right)}\right]\right|\ge\theta+\varepsilon K\log^{\frac{1}{\alpha}}\left(\frac{2}{\gamma}\right)\|\boldsymbol{B}_\star\|_{1\to2}\right).\end{aligned}\tag{16}$$

Observe that $(\boldsymbol{v}-\boldsymbol{v}^{-i})/2=\boldsymbol{v}|_i$ and $(\boldsymbol{v}+\boldsymbol{v}^{-i})/2=\boldsymbol{v}|_{\backslash i}$ are respectively the selectors of the ith coordinate and its complement. With this convention, on one hand we can write

$$\begin{aligned}&\mathbb{P}\left(\left\|\boldsymbol{B}_\star\left(\frac{1}{2}\boldsymbol{u}_{t-1}-\frac{1}{2}\boldsymbol{u}_{t-1}^{-i_{t-1}}\right)\right\|_2\ge K\log^{\frac{1}{\alpha}}\left(\frac{2}{\gamma}\right)\|\boldsymbol{B}_\star\|_{1\to2}\right)\\
&=\mathbb{P}\left(\left\|\boldsymbol{B}_\star\left(\boldsymbol{u}_{t-1}\,|_{i_{t-1}}\right)\right\|_2\ge K\log^{\frac{1}{\alpha}}\left(\frac{2}{\gamma}\right)\|\boldsymbol{B}_\star\|_{1\to2}\right)\\
&\le\gamma\,,\end{aligned}\tag{17}$$

where the third line follows from the fact that $\|\boldsymbol{B}_\star\|_{1\to2}$ is equal to the greatest ℓ_2 norm of the columns of $\boldsymbol{B}_\star$, and that under the assumption (8) we have

$$\mathbb{P}\left(\left|(\boldsymbol{u}_{t-1})_{i_{t-1}}\right|\ge K\log^{\frac{1}{\alpha}}\left(\frac{2}{\gamma}\right)\right)\le\gamma\,.$$

On the other hand, we can write

$$\boldsymbol{B}_\star\left(\frac{1}{2}\boldsymbol{u}_{t-1}+\frac{1}{2}\boldsymbol{u}_{t-1}^{-i_{t-1}}\right)=\boldsymbol{B}_{\star\backslash i_{t-1}}\boldsymbol{u}_{t-1}$$

and invoke Lemma 7 below to obtain

$$\begin{aligned}&\mathbb{P}\left(\left|\boldsymbol{w}^\top\left[\underset{\beta\,\boldsymbol{u}_t}{F(\boldsymbol{A}_\star\boldsymbol{x}_{t-1}^{(\ell)})\boldsymbol{B}_{\star\backslash i_{t-1}}\boldsymbol{u}_{t-1}}\right]\right|\ge 0.36\min_{i=1,\dots,p}\min\left\{\beta,(\lambda-\varepsilon)\lambda_{\min}^{1/2}\left(\boldsymbol{B}_{\star\backslash i}\boldsymbol{B}_{\star\backslash i}^\top\right)\right\}\right)\\
&\ge\frac{0.4}{\max\{\eta,3\}}\end{aligned}\tag{18}$$

Therefore, recalling the assumed condition (7), by choosing

$$\theta = \theta_{\alpha,\beta,\varepsilon,\lambda,K,\boldsymbol{B}_\star}\,,$$

and

$$\gamma = \frac{0.2}{\max\{\eta,3\}}\,,$$

and in view of (15), (16), (17), and (18) we obtain the desired bound

$$\mathbb{P}\left(\left|\boldsymbol{w}^\top \boldsymbol{z}_t^{(\ell)}\right| \geq \theta_{\alpha,\beta,\varepsilon,\lambda,K,\boldsymbol{B}_\star}\right) \geq \frac{0.1}{\max\{\eta,3\}}\,.$$

□

Lemma 7. *With the notation and conditions as in Lemma 6 we have*

$$\mathbb{P}\left(\left|\boldsymbol{w}^\top \begin{bmatrix} F(\boldsymbol{A}_\star \boldsymbol{x}_{t-1}^{(\ell)})\boldsymbol{B}_{\star\backslash i_{t-1}}\boldsymbol{u}_{t-1} \\ \beta\,\boldsymbol{u}_t \end{bmatrix}\right| \geq 0.36 \min_{i=1,\dots,p}\,\min\left\{\beta, (\lambda-\varepsilon)\lambda_{\min}^{1/2}\left(\boldsymbol{B}_{\star\backslash i}\boldsymbol{B}_{\star\backslash i}^\top\right)\right\}\right)$$
$$\geq \frac{0.4}{\max\{\eta,3\}}$$

Proof. By conditioning on $\boldsymbol{x}_{t-1}^{(\ell)}$ and applying the Paley-Zygmund inequality (Paley and Zygmund, 1932; de la Peña and Giné, 1999, Corollary 3.3.2) we have

$$\mathbb{P}\left(\left|\boldsymbol{w}^\top \begin{bmatrix} F(\boldsymbol{A}_\star \boldsymbol{x}_{t-1}^{(\ell)})\boldsymbol{B}_{\star\backslash i_{t-1}}\boldsymbol{u}_{t-1} \\ \beta\,\boldsymbol{u}_t \end{bmatrix}\right|^2 \geq 0.36\,\mathbb{E}\left(\left|\boldsymbol{w}^\top \begin{bmatrix} F(\boldsymbol{A}_\star \boldsymbol{x}_{t-1}^{(\ell)})\boldsymbol{B}_{\star\backslash i_{t-1}}\boldsymbol{u}_{t-1} \\ \beta\,\boldsymbol{u}_t \end{bmatrix}\right|^2 \,\middle|\, \boldsymbol{x}_{t-1}^{(\ell)}\right) \,\middle|\, \boldsymbol{x}_{t-1}^{(\ell)}\right)$$
$$\geq 0.4\,\frac{\left(\mathbb{E}\left(\left|\boldsymbol{w}^\top \begin{bmatrix} F(\boldsymbol{A}_\star \boldsymbol{x}_{t-1}^{(\ell)})\boldsymbol{B}_{\star\backslash i_{t-1}}\boldsymbol{u}_{t-1} \\ \beta\,\boldsymbol{u}_t \end{bmatrix}\right|^2 \,\middle|\, \boldsymbol{x}_{t-1}^{(\ell)}\right)\right)^2}{\mathbb{E}\left(\left|\boldsymbol{w}^\top \begin{bmatrix} F(\boldsymbol{A}_\star \boldsymbol{x}_{t-1}^{(\ell)})\boldsymbol{B}_{\star\backslash i_{t-1}}\boldsymbol{u}_{t-1} \\ \beta\,\boldsymbol{u}_t \end{bmatrix}\right|^4 \,\middle|\, \boldsymbol{x}_{t-1}^{(\ell)}\right)} \tag{19}$$

Using the assumption that $\boldsymbol{u}_{t-1}$ and $\boldsymbol{u}_t$ are independent, zero-mean, and isotropic we obtain

$$\mathbb{E}\left(\left|\boldsymbol{w}^\top \begin{bmatrix} F(\boldsymbol{A}_\star \boldsymbol{x}_{t-1}^{(\ell)})\boldsymbol{B}_{\star\backslash i_{t-1}}\boldsymbol{u}_{t-1} \\ \beta\,\boldsymbol{u}_t \end{bmatrix}\right|^2 \,\middle|\, \boldsymbol{x}_{t-1}^{(\ell)}\right) = \left\|\begin{bmatrix} \left(F(\boldsymbol{x}_{t-1}^{(\ell)})\boldsymbol{B}_{\star\backslash i_{t-1}}\right)^\top & \boldsymbol{0} \\ \boldsymbol{0} & \beta\,\boldsymbol{I} \end{bmatrix}\boldsymbol{w}\right\|_2^2.$$

Furthermore, in view of Lemma 2, the denominator in (19) can be bounded from above as

$$\mathbb{E}\left(\left|\boldsymbol{w}^\top \begin{bmatrix} F(\boldsymbol{A}_\star \boldsymbol{x}_{t-1}^{(\ell)})\boldsymbol{B}_{\star\backslash i_{t-1}}\boldsymbol{u}_{t-1} \\ \beta\,\boldsymbol{u}_t \end{bmatrix}\right|^4 \,\middle|\, \boldsymbol{x}_{t-1}^{(\ell)}\right) \leq \max\{\eta,3\}\left\|\begin{bmatrix} \left(F(\boldsymbol{x}_{t-1}^{(\ell)})\boldsymbol{B}_{\star\backslash i_{t-1}}\right)^\top & \boldsymbol{0} \\ \boldsymbol{0} & \beta\,\boldsymbol{I} \end{bmatrix}\boldsymbol{w}\right\|_2^4.$$

Therefore, (19) reduces to

$$\begin{aligned}&\mathbb{P}\left(\left|\boldsymbol{w}^\mathsf{T}\begin{bmatrix}F(\boldsymbol{A}_\star\boldsymbol{x}_{t-1}^{(\ell)})\boldsymbol{B}_{\star\backslash i_{t-1}}\boldsymbol{u}_{t-1}\\ \beta\,\boldsymbol{u}_t\end{bmatrix}\right|^2 \geq 0.36\,\mathbb{E}\left(\left|\boldsymbol{w}^\mathsf{T}\begin{bmatrix}F(\boldsymbol{A}_\star\boldsymbol{x}_{t-1}^{(\ell)})\boldsymbol{B}_{\star\backslash i_{t-1}}\boldsymbol{u}_{t-1}\\ \beta\,\boldsymbol{u}_t\end{bmatrix}\right|^2 \,\middle|\, \boldsymbol{x}_{t-1}^{(\ell)}\right) \,\middle|\, \boldsymbol{x}_{t-1}^{(\ell)}\right)\\ &\geq \frac{0.4}{\max\{\eta,3\}}\,. \qquad (20)\end{aligned}$$

It follows from Lemma 1 that

$$\lambda_{\min}\left(\begin{bmatrix}F(\boldsymbol{y})\boldsymbol{B}_{\star\backslash i} & \boldsymbol{0}\\ \boldsymbol{0} & \beta\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\left(F(\boldsymbol{y})\boldsymbol{B}_{\star\backslash i}\right)^\mathsf{T} & \boldsymbol{0}\\ \boldsymbol{0} & \beta\boldsymbol{I}\end{bmatrix}\right) \geq \min\{\beta^2, (\lambda-\varepsilon)^2\lambda_{\min}\left(\boldsymbol{B}_{\star\backslash i}\boldsymbol{B}_{\star\backslash i}^\mathsf{T}\right)\}$$

for all $\boldsymbol{y}$. In particular,

$$\min\{\beta^2, (\lambda-\varepsilon)^2\,\lambda_{\min}^2\left(\boldsymbol{B}_{\star\backslash i}\boldsymbol{B}_{\star\backslash i}^\mathsf{T}\right)\} \leq \left\|\begin{bmatrix}\left(F(\boldsymbol{x}_{t-1}^{(\ell)})\boldsymbol{B}_{\star\backslash i}\right)^\mathsf{T} & \boldsymbol{0}\\ \boldsymbol{0} & \beta\boldsymbol{I}\end{bmatrix}\boldsymbol{w}\right\|_2^2.$$

Therefore, the conditional expectation in (20) can be replaced by

$$\min_{i=1,\ldots,p}\min\{\beta^2, (\lambda-\varepsilon)^2\,\lambda_{\min}\left(\boldsymbol{B}_{\star\backslash i}\boldsymbol{B}_{\star\backslash i}^\mathsf{T}\right)\}\,.$$

Finally, taking the expectation with respect to $\boldsymbol{x}_{t-1}^{(\ell)}$ completes the proof. □