Gradient Descent Learns Linear Dynamical Systems

Moritz Hardt   M@MRTZ.ORG
Department of Electrical Engineering and Computer Science
University of California, Berkeley

Tengyu Ma   TENGYUMA@STANFORD.EDU
Facebook AI Research

Benjamin Recht   BRECHT@BERKELEY.EDU
Department of Electrical Engineering and Computer Science
University of California, Berkeley

Abstract

We prove that stochastic gradient descent efficiently converges to the global optimizer of the maximum likelihood objective of an unknown linear time-invariant dynamical system from a sequence of noisy observations generated by the system. Even though the objective function is non-convex, we provide polynomial running time and sample complexity bounds under strong but natural assumptions. Linear systems identification has been studied for many decades, yet, to the best of our knowledge, these are the first polynomial guarantees for the problem we consider.

Key-words: non-convex optimization, linear dynamical system, stochastic gradient descent, generalization bounds, time series, over-parameterization

1. Introduction

Many learning problems are by their nature sequence problems where the goal is to fit a model that maps a sequence of input words $x_1, \ldots, x_T$ to a corresponding sequence of observations $y_1, \ldots, y_T$. Text translation, speech recognition, time series prediction, video captioning and question answering systems, to name a few, are all sequence to sequence learning problems. For a sequence model to be both expressive and parsimonious in its parameterization, it is crucial to equip the model with memory thus allowing its prediction at time t to depend on previously seen inputs.

Recurrent neural networks form an expressive class of non-linear sequence models. Through their many variants, such as long-short-term-memory Hochreiter and Schmidhuber (1997), recurrent neural networks have seen remarkable empirical success in a broad range of domains. At the core, neural networks are typically trained using some form of (stochastic) gradient descent. Even though the training objective is non-convex, it is widely observed in practice that gradient descent quickly approaches a good set of model parameters. Understanding the effectiveness of gradient descent for non-convex objectives on theoretical grounds is a major open problem in this area.

If we remove all non-linear state transitions from a recurrent neural network, we are left with the state transition representation of a linear dynamical system. Notwithstanding, the natural training objective for linear systems remains non-convex due to the composition of multiple linear operators in the system. If there is any hope of eventually understanding recurrent neural networks, it will be inevitable to develop a solid understanding of this special case first.

To be sure, linear dynamical systems are important in their own right and have been studied for many decades independently of machine learning within the control theory community. Control theory provides a rich set techniques for identifying and manipulating linear systems. The learning problem in this context corresponds to "linear dynamical system identification". Maximum likelihood estimation with gradient descent is a popular heuristic for dynamical system identification Ljung (1998). In the context of machine learning, linear systems play an important role in numerous tasks. For example, their estimation arises as subroutines of reinforcement learning in robotics Levine and Koltun (2013), location and mapping estimation in robotic systems Durrant-Whyte and Bailey (2006), and estimation of pose from video Rahimi et al. (2005).

In this work, we show that gradient descent efficiently minimizes the maximum likelihood objective of an unknown linear system given noisy observations generated by the system. More formally, we receive noisy observations generated by the following *time-invariant linear system*:

$$h_{t+1} = Ah_t + Bx_t \tag{1.1}$$
$$y_t = Ch_t + Dx_t + \xi_t$$

Here, A, B, C, D are linear transformations with compatible dimensions and we denote by $\Theta = (A, B, C, D)$ the parameters of the system. The vector h_t represents the hidden state of the model at time t. Its dimension n is called the *order* of the system. The stochastic noise variables $\{\xi_t\}$ perturb the output of the system which is why the model is called an *output error model* in control theory. We assume the variables are drawn i.i.d. from a distribution with mean 0 and variance σ^2.

Throughout the paper we focus on *controllable* and *externally stable* systems. A linear system is externally stable (or *equivalently bounded-input bounded-output stable*) if and only if the spectral radius of A, denoted $\rho(A)$, is strictly bounded by 1. Controllability is a mild non-degeneracy assumption that we formally define later. Without loss of generality, we further assume that the transformations B, C and D have bounded Frobenius norm. This can be achieved by a rescaling of the output variables. We assume we have N pairs of sequences (x, y) as training examples,

$$S = \left\{ (x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)}) \right\} .$$

Each input sequence $x \in \mathbb{R}^T$ of length T is drawn from a distribution and y is the corresponding output of the system above generated from an unknown initial state h. We allow the unknown initial state to vary from one input sequence to the next. This only makes the learning problem more challenging.

Our goal is to fit a linear system to the observations. We parameterize our model in exactly the same way as (1.1). That is, for linear mappings $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$, the trained model

is defined as:

$$\hat{h}_{t+1} = \hat{A}\hat{h}_t + \hat{B}x_t, \qquad \hat{y}_t = \hat{C}\hat{h}_t + \hat{D}x_t \tag{1.2}$$

The *(population) risk* of the model is obtained by feeding the learned system with the correct initial states and comparing its predictions with the ground truth in expectation over inputs and errors. Denoting by $\hat{y}_t$ the t-th prediction of the trained model starting from the correct initial state that generated y_t, and using $\widehat{\Theta}$ as a short hand for $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$, we formally define population risk as:

$$f(\widehat{\Theta}) = \mathop{\mathbb{E}}_{\{x_t\},\{\xi_t\}} \left[\frac{1}{T} \sum_{t=1}^{T} \|\hat{y}_t - y_t\|^2 \right] \tag{1.3}$$

Note that even though the prediction $\hat{y}_t$ is generated from the correct initial state, the learning algorithm does not have access to the correct initial state for its training sequences.

While the squared loss objective turns out to be non-convex, it has many appealing properties. Assuming the inputs x_t and errors ξ_t are drawn independently from a Gaussian distribution, the corresponding population objective corresponds to maximum likelihood estimation. In this work, we make the weaker assumption that the inputs and errors are drawn independently from possibly different distributions. The independence assumption is certainly idealized for some learning applications. However, in control applications the inputs can often be chosen by the controller rather than by nature. Moreover, the outputs of the system at various time steps are correlated through the unknown hidden state and therefore not independent even if the inputs are.

1.1 Our results

We show that we can efficiently minimize the population risk using projected stochastic gradient descent. The bulk of our work applies to single-input single-output (SISO) systems meaning that inputs and outputs are scalars $x_t, y_t \in \mathbb{R}$. However, the hidden state can have arbitrary dimension n. Every controllable SISO admits a convenient canonical form called *controllable canonical form* that we formally introduce later in Section 1.7. In this canonical form, the transition matrix A is governed by n parameters $a_1, \ldots, a_n$ which coincide with the coefficients of the characteristic polynomial of A. The minimal assumption under which we might hope to learn the system is that the spectral radius of A is smaller than 1. However, the set of such matrices is non-convex and does not have enough structure for our analysis.[1] We will therefore make additional assumptions. The assumptions we need differ between the case where we are trying to learn A with n parameter system, and the case where we allow ourselves to over-specify the trained model with $n' > n$ parameters. The former is sometimes called proper learning, while the latter is called improper learning. In the improper case, we are essentially able to learn any system with spectral radius less than 1 under a mild separation condition on the roots of the characteristic polynomial. Our assumption in the proper case is stronger and we introduce it next.

---

1. In both the controllable canonical form and the standard parameterization of the matrix A, the set of matrices with spectral radius less than 1 is not convex.

1.2 Proper learning

Suppose the state transition matrix A has characteristic polynomial $\det(zI - A) = z^n + a_1 z^{n-1} + \cdots + a_n$, which turns out to by and large decide the difficulty of the learning according to our analysis. (In fact, we will parameterize A in a way so that the coefficients of the characteristic polynomials are the parameters of learning problem. See Section 1.7 for the detailed setup.) Consider the corresponding polynomial $q(z) = 1 + a_1 z + a_2 z^2 + \cdots + a_n z^n$ over the complex numbers $\mathbb{C}$.

We will require the state transition matrix satisfy that the image of the unit circle on the complex plane under the polynomial q is contained in the cone of complex numbers whose real part is larger than their absolute imaginary part. Formally, for all $z \in \mathbb{C}$ such that $|z| = 1$, we require that $\Re(q(z)) > |\Im(q(z))|$. Here, $\Re(z)$ and $\Im(z)$ denote the real and imaginary part of z, respectively. We illustrate this condition in Figure 1 on the right for a degree 4 system.

Our assumption has three important implications. First, it implies (via Rouché's theorem) that the spectral radius of A is smaller than 1 and therefore ensures the stability of the system. Second, the vectors satisfying our assumption form a convex set in $\mathbb{R}^n$. Finally, it ensures that the objective function is *weakly quasi-convex*, a condition we introduce later when we show that it enables stochastic gradient descent to make sufficient progress.

We note in passing that our assumption can be satisfied via the ℓ_1-norm constraint $\|a\|_1 \leq \sqrt{2}/2$. Moreover, if we pick random Gaussian coefficients with expected norm bounded by $o(1/\sqrt{\log n})$, then the resulting vector will satisfy our assumption with probability $1 - o(1)$. Roughly speaking, the assumption requires the roots of the characteristic polynomial $p(z) = z^n + a_1 z^{n-1} + \cdots + a_n$ are relatively dispersed inside the unit circle. (For comparison, on the other end of the spectrum, the polynomial $p(z) = (z - 0.99)^n$ have all its roots colliding at the same point and doesn't satisfy the assumption.)

Figure 1: An example of polynomial q that satisfies our assumption. The unit circle is the collection of the inputs of q and the other curve shows the corresponding outputs (with the corresponding colors.) We see the image of the polynomial stays in the wedge which contains all the complex number z satisfying $\Re(q(z)) > |\Im(q(z))|$.

Theorem 1.1 (Informal) *Under our assumption, projected stochastic gradient descent, when given N sample sequence of length T, returns parameters $\widehat{\Theta}$ with population risk*

$$\mathbb{E}\, f(\widehat{\Theta}) \leq f(\Theta) + O\left(\sqrt{\frac{n^5 + \sigma^2 n^3}{TN}}\right).$$

Here the expectation on LHS is with respect to the randomness of the algorithm. Recall that $f(\Theta)$ is the population risk of the optimal system, and σ^2 refers to the variance of the noise variables. We also assume that the inputs x_t are drawn from a pairwise independent distribution with mean 0 and variance 1. Note, however, that this does not imply
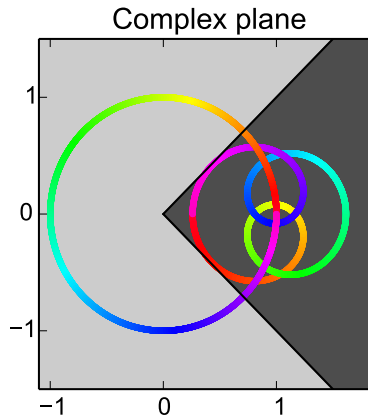
independence of the outputs as these are correlated by a common hidden state. The stated version of our result glosses over the fact that we need our assumption to hold with a small amount of slack; a precise version follows in Section 4. Our theorem establishes a polynomial convergence rate for stochastic gradient descent. Since each iteration of the algorithm only requires a sequence of matrix operations and an efficient projection step, the running time is polynomial, as well. Likewise, the sample requirements are polynomial since each iteration requires only a single fresh example. An important feature of this result is that the error decreases with both the length T and the number of samples N. The dependency on the dimension n, on the other hand, is likely to be quite loose, and tighter bounds are left for future work.

The algorithm requires a (polynomial-time) projection step to a convex set at every iteration (formally defined in Section 4 and Algorithm 1). Computationally, it can be a bottleneck, although it is unlikely to be required in practice and may be an artifact of our analysis.

1.3 The power of over-parameterization

Endowing the model with additional parameters compared to the ground truth turns out to be surprisingly powerful. We show that we can essentially remove the assumption we previously made in proper learning. The idea is simple. If p is the characteristic polynomial of A of degree n. We can find a system of order $n' > n$ such that the characteristic polynomial of its transition matrix becomes $p \cdot p'$ for some polynomial p' of order $n' - n$. This means that to apply our result we only need the polynomial $p \cdot p'$ to satisfy our assumption. At this point, we can choose p' to be an approximation of the inverse p^{-1}. For sufficiently good approximation, the resulting polynomial $p \cdot p'$ is close to 1 and therefore satisfies our assumption. Such an approximation exists generically for $n' = O(n)$ under mild non-degeneracy assumptions on the roots of p. In particular, any small random perturbation of the roots would suffice.

Theorem 1.2 (Informal) *Under a mild non-degeneracy assumption, stochastic gradient descent returns parameters* $\widehat{\Theta}$ *corresponding to a system of order* $n' = O(n)$ *with population risk*

$$f(\widehat{\Theta}) \leq f(\Theta) + O\left(\sqrt{\frac{n^5 + \sigma^2 n^3}{TN}}\right),$$

when given N sample sequences of length T.

We remark that the idea we sketched also shows that, in the extreme, improper learning of linear dynamical systems becomes easy in the sense that the problem can be solved using linear regression against the outputs of the system. However, our general result interpolates between the proper case and the regime where linear regression works. We discuss in more details in Section 6.3.

1.4 Multi-input multi-output systems

Both results we saw immediately extend to single-input multi-output (SIMO) systems as the dimensionality of C and D are irrelevant for us. The case of multi-input multi-output

(MIMO) systems is more delicate. Specifically, our results carry over to a broad family of multi-input multi-output systems. However, in general MIMO systems no longer enjoy canonical forms like SISO systems. In Section 7, we introduce a natural generalization of controllable canonical form for MIMO systems and extend our results to this case.

1.5 Related work

System identification is a core problem in dynamical systems and has been studied in depth for many years. The most popular reference on this topic is the text by Ljung Ljung (1998). Nonetheless, the list of non-asymptotic results on identifying linear systems from noisy data is surprisingly short. Several authors have recently tried to estimate the sample complexity of dynamical system identification using machine learning tools Vidyasagar and Karandikar (2008); Campi and Weyer (2002); Weyer and Campi (1999). All of these result are rather pessimistic with sample complexity bounds that are exponential in the degree of the linear system and other relevant quantities. Contrastingly, we prove that gradient descent has an associated polynomial sample complexity in all of these parameters. Moreover, all of these papers only focus on how well empirical risk approximates the true population risk and do not provide guarantees about any algorithmic schemes for minimizing the empirical risk.

The only result to our knowledge which provides polynomial sample complexity for identifying linear dynamical systems is in Shah *et al* Shah et al. (2012). Here, the authors show that if certain *frequency domain* information about the linear dynamical system is observed, then the linear system can be identified by solving a second-order cone programming problem. This result is about improper learning only, and the size of the resulting system may be quite large, scaling as the $(1 - \rho(A))^{-2}$. As we describe in this work, very simple algorithms work in the improper setting when the system degree is allowed to be polynomial in $(1 - \rho(A))^{-1}$. Moreover, it is not immediately clear how to translate the frequency-domain results to the time-domain identification problem discussed above.

Our main assumption about the image of the polynomial $q(z)$ is an appeal to the theory of passive systems. A system is passive if the dot product between the input sequence u_t and output sequence y_t are strictly positive. Physically, this notion corresponds to systems that cannot create energy. For example, a circuit made solely of resistors, capacitors, and inductors would be a passive electrical system. If one added an amplifier to the internals of the system, then it would no longer be passive. The set of passive systems is a subset of the set of stable systems, and the subclass is somewhat easier to work with mathematically. Indeed, Megretski used tools from passive systems to provide a relaxation technique for a family of identification problems in dynamical systems Megretski (2008). His approach is to lower bound a nonlinear least-squares cost with a convex functional. However, he does not prove that his technique can identify any of the systems, even asymptotically. Stoica and Söderström Söderström and Stoica (1982); Stoica and Söderström (1982, 1984) and later Bazanella *et al.* Bazanella et al. (2008); Eckhard and Bazanella (2011) prove the quasi-convexity of a cost function under a passivity condition in the context of system identification, but no sample complexity or global convergence proofs are provided.

1.6 Proof overview

The first important step in our proof is to develop population risk in Fourier domain where it is closely related to what we call *idealized risk*. Idealized risk essentially captures the ℓ_2-difference between the *transfer function* of the learned system and the ground truth. The transfer function is a fundamental object in control theory. Any linear system is completely characterized by its transfer function $G(z) = C(zI - A)^{-1}B$. In the case of a SISO, the transfer function is a rational function of degree n over the complex numbers and can be written as $G(z) = s(z)/p(z)$. In the canonical form introduced in Section 1.7, the coefficients of $p(z)$ are precisely the parameters that specify A. Moreover, $z^n p(1/z) = 1 + a_1 z + a_2 z^2 + \cdots + a_n z^n$ is the polynomial we encountered in the introduction. Under the assumption illustrated earlier, we show in Section 3 that the idealized risk is *weakly quasi-convex* (Lemma 3.3). Quasi-convexity implies that gradients cannot vanish except at the optimum of the objective function; we review this (mostly known) material in Section 2. In particular, this lemma implies that in principle we can hope to show that gradient descent converges to a global optimum. However, there are several important issues that we need to address. First, the result only applies to idealized risk, not our actual population risk objective. Therefore it is not clear how to obtain unbiased gradients of the idealized risk objective. Second, there is a subtlety in even defining a suitable empirical risk objective. The reason is that risk is defined with respect to the correct initial state of the system which we do not have access to during training. We overcome both of these problems. In particular, we design an almost unbiased estimator of the gradient of the idealized risk in Lemma 5.4 and give variance bounds of the gradient estimator (Lemma 5.5).

Our results on improper learning in Section 6 rely on a surprisingly simple but powerful insight. We can extend the degree of the transfer function $G(z)$ by extending both numerator and denominator with a polynomial $u(z)$ such that $G(z) = s(z)u(z)/p(z)u(z)$. While this results in an equivalent system in terms of input-output behavior, it can dramatically change the geometry of the optimization landscape. In particular, we can see that only $p(z)u(z)$ has to satisfy the assumption of our proper learning algorithm. This allows us, for example, to put $u(z) \approx p(z)^{-1}$ so that $p(z)u(z) \approx 1$, hence trivially satisfying our assumption. A suitable inverse approximation exists under light assumptions and requires degree no more than $d = O(n)$. Algorithmically, there is almost no change. We simply run stochastic gradient descent with $n + d$ model parameters rather than n model parameters.

1.7 Preliminaries

For complex matrix (or vector, number) C, we use $\Re(C)$ to denote the real part and $\Im(C)$ the imaginary part, and $\bar{C}$ the conjugate and $C^* = \bar{C}^\top$ its conjugate transpose . We use $|\cdot|$ to denote the absolute value of a complex number c. For complex vector u and v, we use $\langle u, v \rangle = u^* v$ to denote the inner product and $\|u\| = \sqrt{u^* u}$ is the norm of u. For complex matrix A and B with same dimension, $\langle A, B \rangle = \text{tr}(A^* B)$ defines an inner product, and $\|A\|_F = \sqrt{\text{tr}(A^* A)}$ is the Frobenius norm. For a square matrix A, we use $\rho(A)$ to denote the spectral radius of A, that is, the largest absolute value of the elements in the spectrum of A. We use I_n to denote the identity matrix with dimension $n \times n$, and we drop the subscript when it's clear from the context.We let e_i denote the i-th standard basis vector.

A SISO of order n is in *controllable canonical form* if A and B have the following form

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_1 \end{bmatrix} \qquad B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \qquad (1.4)$$

We will parameterize $\hat{A}, \hat{B}, \hat{C}, \hat{D}$ accordingly. We will write $A = \mathrm{CC}(a)$ for brevity, where a is used to denote the unknown last row $[-a_n, \ldots, -a_1]$ of matrix A. We will use $\hat{a}$ to denote the corresponding training variables for a. Since here B is known, so $\hat{B}$ is no longer a trainable parameter, and is forced to be equal to B. Moreover, C is a row vector and we use $[c_1, \cdots, c_n]$ to denote its coordinates (and similarly for $\hat{C}$).

A SISO is *controllable* if and only if the matrix $[B \mid AB \mid A^2 B \mid \cdots \mid A^{n-1}B]$ has rank n. This statement corresponds to the condition that all hidden states should be reachable from some initial condition and input trajectory. Any controllable system admits a controllable canonical form Heij et al. (2007). For vector $a = [a_n, \ldots, a_1]$, let $p_a(z)$ denote the polynomial

$$p_a(z) = z^n + a_1 z^{n-1} + \cdots + a_n. \qquad (1.5)$$

When a defines the matrix A that appears in controllable canonical form, then p_a is precisely the characteristic polynomial of A. That is, $p_a(z) = \det(zI - A)$.

2. Gradient descent and quasi-convexity

It is known that under certain mild conditions (stochastic) gradient descent converges even on non-convex functions to local minimum Ge et al. (2015); Lee et al. (2016). Though usually for concrete problems the challenge is to prove that there is no spurious local minimum other than the target solution. Here we introduce a condition similar to the quasi-convexity notion in Hazan et al. (2015), which ensures that any point with vanishing gradient is the optimal solution . Roughly speaking, the condition says that at any point θ the negative of the gradient $-\nabla f(\theta)$ should be positively correlated with direction $\theta^* - \theta$ pointing towards the optimum. Our condition is slightly weaker than that in Hazan et al. (2015) since we only require quasi-convexity and smoothness with respect to the optimum, and this (simple) extension will be necessary for our analysis.

Definition 2.1 (Weak quasi-convexity) *We say an objective function f is τ-weakly-quasi-convex (τ-WQC) over a domain $\mathcal{B}$ with respect to global minimum θ^* if there is a positive constant $\tau > 0$ such that for all $\theta \in \mathcal{B}$,*

$$\nabla f(\theta)^\top (\theta - \theta^*) \geq \tau(f(\theta) - f(\theta^*)). \qquad (2.1)$$

We further say f is Γ-weakly-smooth if for for any point θ, $\|\nabla f(\theta)\|^2 \leq \Gamma(f(\theta) - f(\theta^*))$.

Note that indeed any Γ-smooth function in the usual sense (that is, $\|\nabla^2 f\| \leq \Gamma$) is $O(\Gamma)$-weakly-smooth. For a random vector $X \in \mathbb{R}^n$, we define it's variance to be $\mathrm{Var}[X] = \mathbb{E}[\|X - \mathbb{E}X\|^2]$.

Definition 2.2 *We call* $\mathfrak{r}(\theta)$ *an unbiased estimator of* $\nabla f(\theta)$ *with variance* V *if it satisfies* $\mathbb{E}[\mathfrak{r}(\theta)] = \nabla f(\theta)$ *and* $\mathrm{Var}[\mathfrak{r}(\theta)] \leq V$.

Projected stochastic gradient descent over some closed convex set $\mathcal{B}$ with *learning rate* $\eta > 0$ refers to the following algorithm in which $\Pi_{\mathcal{B}}$ denotes the Euclidean projection onto $\mathcal{B}$:

$$\begin{aligned}
&\textbf{for } k = 0 \text{ to } K - 1 : \\
&\qquad w_{k+1} = \theta_k - \eta \mathfrak{r}(\theta_k) \\
&\qquad \theta_{k+1} = \Pi_{\mathcal{B}}(w_{k+1}) \\
&\textbf{return } \theta_j \text{ with } j \text{ uniformly picked from } 1, \ldots, K
\end{aligned} \qquad (2.2)$$

The following Proposition is well known for convex objective functions (corresponding to 1-weakly-quasi-convex functions). We extend it (straightforwardly) to the case when τ-WQC holds with any positive constant τ.

Proposition 2.3 *Suppose the objective function* f *is* τ*-weakly-quasi-convex and* Γ*-weakly-smooth, and* $\mathfrak{r}(\cdot)$ *is an unbiased estimator for* $\nabla f(\theta)$ *with variance* V. *Moreover, suppose the global minimum* θ^* *belongs to* $\mathcal{B}$, *and the initial point* θ_0 *satisfies* $\|\theta_0 - \theta^*\| \leq R$. *Then projected gradient descent* (2.2) *with a proper learning rate returns* θ_K *in* K *iterations with expected error*

$$\mathbb{E}\, f(\theta_K) - f(\theta^*) \leq O\left(\max\left\{ \frac{\Gamma R^2}{\tau^2 K}, \frac{R\sqrt{V}}{\tau\sqrt{K}} \right\} \right).$$

Remark 2.4 *It's straightforward to see (from the proof) that the algorithm tolerates inverse exponential bias, namely bias on the order of* $\exp(-\Omega(n))$, *in the gradient estimator. Technically, suppose* $\mathbb{E}[\mathfrak{r}(\theta)] = \nabla f(\theta) \pm \zeta$ *then* $f(\theta_K) \leq O\left(\max\left\{ \frac{\Gamma R^2}{\tau^2 K}, \frac{R\sqrt{V}}{\tau\sqrt{K}} \right\} \right) + \mathrm{poly}(K) \cdot \zeta$. *Throughout the paper, we assume that the error that we are shooting for is inverse polynomial, namely* $1/n^C$ *for some absolute constant* C, *and therefore the effect of inverse exponential bias is negligible.*

We defer the proof of Proposition 2.3 to Appendix A which is a simple variation of the standard convergence analysis of stochastic gradient descent (see, for example, Bottou (1998)). Finally, we note that the sum of two quasi-convex functions may no longer be quasi-convex. However, if a sequence functions is τ-WQC with respect to a common point θ^*, then their sum is also τ-WQC. This follows from the linearity of gradient operation.

Proposition 2.5 *Suppose functions* $f_1, \ldots, f_n$ *are individually* τ*-weakly-quasi-convex in* $\mathcal{B}$ *with respect to a common global minimum* θ^*, *then for non-negative* $w_1, \ldots, w_n$ *the linear combination* $f = \sum_{i=1}^n w_i f_i$ *is also* τ*-weakly-quasi-convex with respect to* θ^* *in* $\mathcal{B}$.

3. Population risk in frequency domain

We next establish conditions under which risk is weakly-quasi-convex. Our strategy is to first approximate the risk functional $f(\widehat{\Theta})$ by what we call *idealized risk*. This approximation

of the objective function is fairly straightforward; we justify it toward the end of the section. We can show that

$$f(\widehat{\Theta}) \approx \|D - \hat{D}\|^2 + \sum_{k=0}^{\infty} \left(\hat{C}\hat{A}^k B - CA^k B\right)^2 . \tag{3.1}$$

The leading term $\|D - \hat{D}\|^2$ is convex in $\hat{D}$ which appears nowhere else in the objective. It therefore doesn't affect the convergence of the algorithm (up to lower order terms) by virtue of Proposition 2.5, and we restrict our attention to the remaining terms.

Definition 3.1 (Idealized risk) *We define the* idealized risk *as*

$$g(\hat{A}, \hat{C}) = \sum_{k=0}^{\infty} \left(\hat{C}\hat{A}^k B - CA^k B\right)^2 . \tag{3.2}$$

We now use basic concepts from control theory (see Heij et al. (2007); Hespanha (2009) for more background) to express the idealized risk (3.2) in Fourier domain. The *transfer function* of the linear system is

$$G(z) = C(zI - A)^{-1}B . \tag{3.3}$$

Note that $G(z)$ is a rational function over the complex numbers of degree n and hence we can find polynomials $s(z)$ and $p(z)$ such that $G(z) = \frac{s(z)}{p(z)}$, with the convention that the leading coefficient of $p(z)$ is 1. In controllable canonical form (1.4), the coefficients of p will correspond to the last row of the A, while the coefficients of s correspond to the entries of C. Also note that

$$G(z) = \sum_{t=1}^{\infty} z^{-t} CA^{t-1}B = \sum_{t=1}^{\infty} z^{-t} r_{t-1}$$

The sequence $r = (r_0, r_1, \ldots, r_t, \ldots) = (CB, CAB, \ldots, CA^t B, \ldots)$ is called the *impulse response* of the linear system. The behavior of a linear system is uniquely determined by the impulse response and therefore by $G(z)$. Analogously, we denote the transfer function of the learned system by $\widehat{G}(z) = \hat{C}(zI - \hat{A})^{-1}B = \hat{s}(z)/\hat{p}(z)$. The idealized risk (3.2) is only a function of the impulse response $\hat{r}$ of the learned system, and therefore it is only a function of $\widehat{G}(z)$.

Recall that $C = [c_1, \ldots, c_n]$ is defined in Section 1.7. For future reference, we note that by some elementary calculation (see Lemma B.1), we have

$$G(z) = C(zI - A)^{-1}B = \frac{c_1 + \cdots + c_n z^{n-1}}{z^n + a_1 z^{n-1} + \cdots + a_n} , \tag{3.4}$$

which implies that $s(z) = c_1 + \cdots + c_n z^{n-1}$ and $p(z) = z^n + a_1 z^{n-1} + \cdots + a_n$.

With these definitions in mind, we are ready to express idealized risk in Fourier domain.

Proposition 3.2 *Suppose $p_{\hat{a}}(z)$ has all its roots inside unit circle, then the idealized risk $g(\hat{a}, \hat{C})$ can be written in the Fourier domain as*

$$g(\hat{A}, \hat{C}) = \int_0^{2\pi} \left|\widehat{G}(e^{i\theta}) - G(e^{i\theta})\right|^2 d\theta .$$

10

Proof Note that $G(e^{i\theta})$ is the Fourier transform of the sequence $\{r_k\}$ and so is $\widehat{G}(e^{i\theta})$ the Fourier transform[2] of $\hat{r}_k$. Therefore by Parseval' Theorem, we have that

$$g(\hat{A}, \hat{C}) = \sum_{k=0}^{\infty} \|\hat{r}_k - r_k\|^2 = \int_0^{2\pi} |\widehat{G}(e^{i\theta}) - G(e^{i\theta})|^2 d\theta . \tag{3.5}$$

∎

3.1 Quasi-convexity of the idealized risk

Now that we have a convenient expression for risk in Fourier domain, we can prove that the idealized risk $g(\hat{A}, \hat{C})$ is weakly-quasi-convex when $\hat{a}$ is not so far from the true a in the sense that $p_a(z)$ and $\hat{p}_a(z)$ have an angle less than $\pi/2$ for every z on the unit circle. We will use the convention that a and $\hat{a}$ refer to the parameters that specify A and $\hat{A}$, respectively.

Lemma 3.3 *For $\tau > 0$ and every $\hat{C}$, the idealized risk $g(\hat{A}, \hat{C})$ is τ-weakly-quasi-convex over the domain*

$$\mathcal{N}_\tau(a) = \left\{ \hat{a} \in \mathbb{R}^n : \Re\left(\frac{p_a(z)}{p_{\hat{a}}(z)}\right) \geq \tau/2, \forall \ z \in \mathbb{C}, \ \text{s.t.} \ |z| = 1 \right\} . \tag{3.6}$$

Proof We first analyze a single term $h = |\widehat{G}(z) - G(z)|^2$. Recall that $\widehat{G}(z) = \hat{s}(z)/\hat{p}(z)$ where $\hat{p}(z) = p_{\hat{a}}(z) = z^n + \hat{a}_1 z^{n-1} + \cdots + \hat{a}_n$. Note that z is fixed and h is a function of $\hat{a}$ and $\hat{C}$. Then it is straightforward to see that

$$\frac{\partial h}{\partial \hat{s}(z)} = 2\Re\left\{ \frac{1}{\hat{p}(z)} \left(\frac{\hat{s}(z)}{\hat{p}(z)} - \frac{s(z)}{p(z)} \right)^* \right\} . \tag{3.7}$$

and

$$\frac{\partial h}{\partial \hat{p}(z)} = -2\Re\left\{ \frac{\hat{s}(z)}{\hat{p}(z)^2} \left(\frac{\hat{s}(z)}{\hat{p}(z)} - \frac{s(z)}{p(z)} \right)^* \right\} . \tag{3.8}$$

Since $\hat{s}(z)$ and $\hat{p}(z)$ are linear in $\hat{C}$ and $\hat{a}$ respectively, by chain rule we have that

$$\langle \frac{\partial h}{\partial \hat{a}}, \hat{a} - a \rangle + \langle \frac{\partial h}{\partial \hat{C}}, \hat{C} - C \rangle = \frac{\partial h}{\partial \hat{p}(z)} \langle \frac{\partial \hat{p}(z)}{\partial \hat{a}}, \hat{a} - a \rangle + \frac{\partial h}{\partial \hat{s}(z)} \langle \frac{\partial \hat{s}(z)}{\partial \hat{C}}, \hat{C} - C \rangle$$

$$= \frac{\partial h}{\partial \hat{p}(z)} (\hat{p}(z) - p(z)) + \frac{\partial h}{\partial \hat{s}(z)} (\hat{s}(z) - s(z)) .$$

---

2. The Fourier transform exists since $\|r_k\|^2 = \|\hat{C}\hat{A}^k\hat{B}\|^2 \leq \|\hat{C}\|\|\hat{A}^k\|\|\hat{B}\| \leq c\rho(\hat{A})^k$ where c doesn't depend on k and $\rho(\hat{A}) < 1$.

Plugging the formulas (3.7) and (3.8) for $\frac{\partial h}{\partial \hat{s}(z)}$ and $\frac{\partial h}{\partial \hat{p}(z)}$ into the equation above, we obtain that

$$
\begin{aligned}
\langle \frac{\partial h}{\partial \hat{a}}, \hat{a} - a \rangle + \langle \frac{\partial h}{\partial \hat{C}}, \hat{C} - C \rangle &= 2\Re \left\{ \frac{-\hat{s}(z)(\hat{p}(z) - p(z)) + \hat{p}(z)(\hat{s}(z) - s(z))}{\hat{p}(z)^2} \left(\frac{\hat{s}(z)}{\hat{p}(z)} - \frac{s(z)}{p(z)} \right)^* \right\} \\
&= 2\Re \left\{ \frac{\hat{s}(z)p(z) - s(z)\hat{p}(z)}{\hat{p}(z)^2} \left(\frac{\hat{s}(z)}{\hat{p}(z)} - \frac{s(z)}{p(z)} \right)^* \right\} \\
&= 2\Re \left\{ \frac{p(z)}{\hat{p}(z)} \left| \frac{\hat{s}(z)}{\hat{p}(z)} - \frac{s(z)}{p(z)} \right|^2 \right\} \\
&= 2\Re \left\{ \frac{p(z)}{\hat{p}(z)} \right\} \left| \widehat{G}(z) - G(z) \right|^2 .
\end{aligned}
$$

Hence $h = |\widehat{G}(z) - G(z)|^2$ is τ-weakly-quasi-convex with $\tau = 2\min_{|z|=1} \Re \left\{ \frac{p(z)}{\hat{p}(z)} \right\}$. This implies our claim, since by Proposition 3.2, the idealized risk g is convex combination of functions of the form $|\widehat{G}(z) - G(z)|^2$ for $|z| = 1$. Moreover, Proposition 2.5 shows that convex combination preserves weak quasi-convexity. ∎

For future reference, we also prove that the idealized risk is $O(n^2/\tau_1^4)$-weakly smooth.

Lemma 3.4 *The idealized risk $g(\hat{A}, \hat{C})$ is Γ-weakly smooth with $\Gamma = O(n^2/\tau_1^4)$.*

Proof By equation (3.8) and the chain rule we get that

$$
\frac{\partial g}{\partial \hat{C}} = \int_{\mathbb{T}} \frac{\partial |\widehat{G}(z) - G(z)|^2}{\partial p(z)} \cdot \frac{\partial p(z)}{\partial \hat{C}} dz = \int_{\mathbb{T}} 2\Re \left\{ \frac{1}{\hat{p}(z)} \left(\frac{\hat{s}(z)}{\hat{p}(z)} - \frac{s(z)}{p(z)} \right)^* \right\} \cdot [1, \ldots, z^{n-1}] dz .
$$

therefore we can bound the norm of the gradient by

$$
\left\| \frac{\partial g}{\partial \hat{C}} \right\|^2 \leq \left(\int_{\mathbb{T}} \left| \frac{\hat{s}(z)}{\hat{p}(z)} - \frac{s(z)}{p(z)} \right|^2 dz \right) \cdot \left(\int_{\mathbb{T}} 4\|[1, \ldots, z^{n-1}]\|^2 \cdot |\frac{1}{p(z)}|^2 dz \right) \leq O(n/\tau_1^2) \cdot g(\hat{A}, \hat{C}) .
$$

Similarly, we could show that $\left\| \frac{\partial g}{\partial \hat{a}} \right\|^2 \leq O(n^2/\tau_1^2) \cdot g(\hat{A}, \hat{C})$. ∎

3.2 Justifying idealized risk

We need to justify the approximation we made in Equation (3.1).

Lemma 3.5 *Assume that ξ_t and x_t are drawn i.i.d. from an arbitrary distribution with mean 0 and variance 1. Then the population risk $f(\widehat{\Theta})$ can be written as,*

$$
f(\widehat{\Theta}) = (\hat{D} - D)^2 + \sum_{k=1}^{T-1} \left(1 - \frac{k}{T}\right) \left(\hat{C}\hat{A}^{k-1}B - CA^{k-1}B \right)^2 + \sigma^2 . \tag{3.9}
$$

The idealized risk is upper bound of the population risk $f(\widehat{\Theta})$ according to equation (3.1) and (3.9). We don't have to quantify the gap between them because later in Algorithm 1, we will directly optimize the idealized risk by constructing an estimator of its gradient, and thus the optimization will guarantee a bounded idealized risk which translates to a bounded population risk. See Section 5 for details.

Proof [Proof of Lemma 3.5] Under the distributional assumptions on ξ_t and x_t, we can calculate the objective functions above analytically. We write out $y_t, \hat{y}_t$ in terms of the inputs,

$$y_t = Dx_t + \sum_{k=1}^{t-1} CA^{t-k-1}Bx_k + CA^{t-1}h_0 + \xi_t\,, \qquad \hat{y}_t = \hat{D}x_t + \sum_{k=1}^{t-1} \hat{C}\hat{A}^{t-k-1}\hat{B}x_k + CA^{t-1}h_0\,.$$

Therefore, using the fact that x_t's are independent and with mean 0 and covariance I, the expectation of the error can be calculated (formally by Claim B.2),

$$\mathbb{E}\left[\|\hat{y}_t - y_t\|^2\right] = \|\hat{D} - D\|_F^2 + \sum_{k=1}^{t-1}\left\|\hat{C}\hat{A}^{t-k-1}\hat{B} - CA^{t-k-1}B\right\|_F^2 + \mathbb{E}[\|\xi_t\|^2]\,. \qquad (3.10)$$

Using $\mathbb{E}[\|\xi_t\|^2] = \sigma^2$, it follows that

$$f(\widehat{\Theta}) = \|\hat{D} - D\|_F^2 + \sum_{k=1}^{T-1}\left(1 - \tfrac{k}{T}\right)\left\|\hat{C}\hat{A}^{k-1}\hat{B} - CA^{k-1}B\right\|_F^2 + \sigma^2\,. \qquad (3.11)$$

Recall that under the controllable canonical form (1.4), $B = e_n$ is known and therefore $\hat{B} = B$ is no longer a variable. Then the expected objective function (3.11) simplifies to

$$f(\widehat{\Theta}) = (\hat{D} - D)^2 + \sum_{k=1}^{T-1}\left(1 - \tfrac{k}{T}\right)\left(\hat{C}\hat{A}^{k-1}B - CA^{k-1}B\right)^2 + \sigma^2\,.$$

∎

The previous lemma does not yet control higher order contributions present in the idealized risk. This requires additional structure that we introduce in the next section.

4. Effective relaxations of spectral radius

The previous section showed quasi-convexity of the idealized risk. However, several steps are missing towards showing finite sample guarantees for stochastic gradient descent. In particular, we will need to control the variance of the stochastic gradient at any system that we encounter in the training. For this purpose we formally introduce our main assumption now and show that it serves as an effective relaxation of spectral radius. This results below will be used for proving convergence of stochastic gradient descent in Section 5.

Consider the following convex region $\mathcal{C}$ in the complex plane,

$$\mathcal{C} = \{z\colon \Re z \geq (1 + \tau_0)|\Im z|\} \cap \{z\colon \tau_1 < \Re z < \tau_2\}\,. \qquad (4.1)$$

where $\tau_0, \tau_1, \tau_2 > 0$ are constants that are considered as fixed constant throughout the paper. Our bounds will have polynomial dependency on these parameters. Pictorially, this convex set is pretty much the dark area in Figure 1 (with the corner chopped). This set in $\mathbb{C}$ induces a convex set in the parameter space which is a subset of the transition matrix with spectral radius less than α.

Definition 4.1 *We say a polynomial $p(z)$ is α-acquiescent if $\{p(z)/z^n \colon |z| = \alpha\} \subseteq \mathcal{C}$. A linear system with transfer function $G(z) = s(z)/p(z)$ is α-acquiescent if the denominator $p(z)$ is.*

The set of coefficients $a \in \mathbb{R}^n$ defining acquiescent systems form a convex set. Formally, for a positive $\alpha > 0$, define the convex set $\mathcal{B}_\alpha \subseteq \mathbb{R}^n$ as

$$\mathcal{B}_\alpha = \left\{ a \in \mathbb{R}^n : \{p_a(z)/z^n \colon |z| = \alpha\} \subseteq \mathcal{C} \right\}. \tag{4.2}$$

We note that definition (4.2) is equivalent to the definition $\mathcal{B}_\alpha = \big\{ a \in \mathbb{R}^n : \{z^n p(1/z) \colon |z| = 1/\alpha\} \subseteq \mathcal{C} \big\}$, which is the version that we used in introduction for simplicity. Indeed, we can verify the convexity of $\mathcal{B}_\alpha$ by definition and the convexity of $\mathcal{C}$: $a, b \in \mathcal{B}_\alpha$ implies that $p_a(z)/z^n, p_b(z)/z^n \in \mathcal{C}$ and therefore, $p_{(a+b)/2}(z)/z^n = \frac{1}{2}(p_a(z)/z^n + p_b(z)/z^n) \in \mathcal{C}$. We also note that the parameter α in the definition of acquiescence corresponds to the spectral radius of the companion matrix. In particular, an acquiescent system is stable for $\alpha < 1$.

Lemma 4.2 *Suppose $a \in \mathcal{B}_\alpha$, then the roots of polynomial $p_a(z)$ have magnitudes bounded by α. Therefore the controllable canonical form $A = \mathrm{CC}(a)$ defined by a has spectral radius $\rho(A) < \alpha$.*

Proof Define holomorphic function $f(z) = z^n$ and $g(z) = p_a(z) = z^n + a_1 z^{n-1} + \cdots + a_n$. We apply the symmetric form of Rouche's theorem Estermann (1962) on the circle $\mathcal{K} = \{z : |z| = \alpha\}$. For any point z on $\mathcal{K}$, we have that $|f(z)| = \alpha^n$, and that $|f(z) - g(z)| = \alpha^n \cdot |1 - p_a(z)/z^n|$. Since $a \in \mathcal{B}_\alpha$, we have that $p_a(z)/z^n \in \mathcal{C}$ for any z with $|z| = \alpha$. Observe that for any $c \in \mathcal{C}$ we have that $|1 - c| < 1 + |c|$, therefore we have that

$$|f(z) - g(z)| = \alpha^n |1 - p_a(z)/z^n| < \alpha^n (1 + |p_a(z)|/|z^n|) = |f(z)| + |p_a(z)| = |f(z)| + |g(z)|.$$

Hence, using Rouche's Theorem, we conclude that f and g have same number of roots inside circle $\mathcal{K}$. Note that function $f = z^n$ has exactly n roots in $\mathcal{K}$ and therefore g have all its n roots inside circle $\mathcal{K}$. ∎

The following lemma establishes the fact that $\mathcal{B}_\alpha$ is a monotone family of sets in α. The proof follows from the maximum modulo principle of the harmonic functions $\Re(z^n p(1/z))$ and $\Im(z^n p(1/z))$. We defer the short proof to Section C.1. We remark that there are larger convex sets than $\mathcal{B}_\alpha$ that ensure bounded spectral radius. However, in order to also guarantee monotonicity and the no blow-up property below, we have to restrict our attention to $\mathcal{B}_\alpha$.

Lemma 4.3 (Monotonicity of $\mathcal{B}_\alpha$) *For any $0 < \alpha < \beta$, we have that $\mathcal{B}_\alpha \subset \mathcal{B}_\beta$.*

Our next lemma entails that acquiescent systems have well behaved impulse responses.

Lemma 4.4 (No blow-up property) *Suppose $a \in \mathcal{B}_\alpha$ for some $\alpha \leq 1$. Then the companion matrix $A = \mathrm{CC}(a)$ satisfies*

$$\sum_{k=0}^{\infty} \|\alpha^{-k} A^k B\|^2 \leq 2\pi n \alpha^{-2n}/\tau_1^2. \tag{4.3}$$

14

Moreover, it holds that for any $k \geq 0$,

$$\|A^k B\|^2 \leq \min\{2\pi n/\tau_1^2, 2\pi n \alpha^{2k-2n}/\tau_1^2\}.$$

Proof [Proof of Lemma 4.4]

Let $f_\lambda = \sum_{k=0}^{\infty} e^{i\lambda k} \alpha^{-k} A^k B$ be the Fourier transform of the series $\alpha^{-k} A^k B$. Then using Parseval's Theorem, we have

$$\sum_{k=0}^{\infty} \|\alpha^{-k} A^k B\|^2 = \int_0^{2\pi} |f_\lambda|^2 d\lambda = \int_0^{2\pi} |(I - \alpha^{-1} e^{i\lambda} A)^{-1} B|^2 d\lambda$$

$$= \int_0^{2\pi} \frac{\sum_{j=1}^{n} \alpha^{2j}}{|p_a(\alpha e^{-i\lambda})|^2} d\lambda \leq \int_0^{2\pi} \frac{n}{|p_a(\alpha e^{-i\lambda})|^2} d\lambda. \qquad (4.4)$$

where at the last step we used the fact that $(I - wA)^{-1} B = \frac{1}{p_a(w^{-1})}[w^{-1}, w^{-2} \ldots, z^{-n}]^\top$ (see Lemma B.1), and that $\alpha \leq 1$. Since $a \in \mathcal{B}_\alpha$, we have that $|q_a(\alpha^{-1} e^{i\lambda})| \geq \tau_1$, and therefore $p_a(\alpha e^{-i\lambda}) = \alpha^n e^{-in\lambda} q(e^{i\lambda}/\alpha)$ has magnitude at least $\tau_1 \alpha^n$. Plugging in this into equation (4.4), we conclude that

$$\sum_{k=0}^{\infty} \|\alpha^{-k} A^k B\|^2 \leq \int_0^{2\pi} \frac{n}{|p_a(\alpha e^{-i\lambda})|^2} d\lambda \leq 2\pi n \alpha^{-2n}/\tau_1^2.$$

Finally we establish the bound for $\|A^k B\|^2$. By Lemma 4.3, we have $\mathcal{B}_\alpha \subset \mathcal{B}_1$ for $\alpha \leq 1$. Therefore we can pick $\alpha = 1$ in equation (4.3) and it still holds. That is, we have that

$$\sum_{k=0}^{\infty} \|A^k B\|^2 \leq 2\pi n/\tau_1^2.$$

This also implies that $\|A^k B\|^2 \leq 2\pi n/\tau_1^2$. ∎

4.1 Efficiently computing the projection

In our algorithm, we require a projection onto $\mathcal{B}_\alpha$. However, the only requirement of the projection step is that it projects onto a set contained inside $\mathcal{B}_\alpha$ that also contains the true linear system. So a variety of subroutines can be used to compute this projection or an approximation. First, the explicit projection onto $\mathcal{B}_\alpha$ is representable by a semidefinite program. This is because each of the three constrains can be checked by testing if a trigonometric polynomial is non-negative. A simple inner approximation can be constructed by requiring the constraints to hold on an a finite grid of size $O(n)$. One can check that this provides a tight, polyhedral approximation to the set $\mathcal{B}_\alpha$, following an argument similar to Appendix C of Bhaskar *et al* Bhaskar et al. (2013). Projection to this polyhedral takes at most $O(n^{3.5})$ time by linear programming and potentially can be made faster by using fast Fourier transform. See Section F for more detailed discussion on why projection on a polytope suffices. Furthermore, sometimes we can replace the constraint by an ℓ_1 or ℓ_2-constraint if we know that the system satisfies the corresponding assumption. Removing the projection step entirely is an interesting open problem.

5. Learning acquiescent systems

Next we show that we can learn acquiescent systems.

Theorem 5.1 *Suppose the true system Θ is α-acquiescent and satisfies $\|C\| \leq 1$. Then with N samples of length $T \geq \Omega(n + 1/(1 - \alpha))$, stochastic gradient descent (Algorithm 1) with projection set $\mathcal{B}_\alpha$ returns parameters $\widehat{\Theta} = (\hat{A}, \hat{B}, \hat{C}, \hat{D})$ with population risk*

$$\mathbb{E}\, f(\widehat{\Theta}) \leq f(\Theta) + O\left(\frac{n^2}{N} + \sqrt{\frac{n^5 + \sigma^2 n^3}{TN}}\right), \tag{5.1}$$

where $O(\cdot)$-notation hides polynomial dependencies on $1/(1-\alpha)$, $1/\tau_0, 1/\tau_1, \tau_2$, and $R = \|a\|$. The expectation is taken over the randomness of the algorithms and the examples.

Algorithm 1 Projected stochastic gradient descent with partial loss

---

For $i = 0$ to N:

1. Take a fresh sample $((x_1, \ldots, x_T), (y_1, \ldots, y_T))$. Let $\tilde{y}_t$ be the simulated outputs[3] of system $\widehat{\Theta}$ on inputs x and initial states $h_0 = 0$.

2. Let $T_1 = T/4$. Run stochastic gradient descent[4] on loss function $\ell((x, y), \widehat{\Theta}) = \frac{1}{T - T_1} \sum_{t > T_1} \|\tilde{y}_t - y_t\|^2$. Concretely, let $G_A = \frac{\partial \ell}{\partial \hat{a}}$, $G_C = \frac{\partial \ell}{\partial \hat{C}}$, and $G_D = \frac{\partial \ell}{\partial \hat{D}}$, we update

$$[\hat{a}, \hat{C}, \hat{D}] \rightarrow [\hat{a}, \hat{C}, \hat{D}] - \eta[G_A, G_C, G_D].$$

3. Project $\widehat{\Theta} = (\hat{a}, \hat{C}, \hat{D})$ to the set $\mathcal{B}_\alpha \otimes \mathbb{R}^n \otimes \mathbb{R}$.

Recall that T is the length of the sequence and N is the number of samples. The first term in the bound (5.1) comes from the smoothness of the population risk and the second comes from the variance of the gradient estimator of population risk (which will be described in detail below). An important (but not surprising) feature here is the variance scale in $1/T$ and therefore for long sequence actually we got $1/N$ convergence instead of $1/\sqrt{N}$ (for relatively small N).

Computational complexity: Step 2 in each iteration of the algorithm takes $O(Tn)$ arithmetic operations, and the projection step takes $O(n^{3.5})$ time to solve an linear programming problem. The project step is unlikely to be required in practice and may be an artifact of our analysis.

We can further balance the variance of the estimator with the number of samples by breaking each long sequence of length T into $\Theta(T/n)$ short sequences of length $\Theta(n)$, and then run back-propagation (1) on these TN/n shorter sequences. This leads us to the following bound which gives the right dependency in T and N as we expected: TN should be counted as the true number of samples for the sequence-to-sequence model.

---

4. Note that $\tilde{y}_t$ is different from $\hat{y}_t$ defined in equation (1.2) which is used to define the population risk: here $\hat{y}_t$ is obtained from the (wrong) initial state $h_0 = 0$ while $\hat{y}_t$ is obtained from the correct initial state.

4. See Algorithm Box 3 for a detailed back-propagation algorithm that computes the gradient.

Corollary 5.2 *Under the assumption of Theorem 5.1, Algorithm 2 returns parameters $\widehat{\Theta}$ with population risk*

$$\mathbb{E}\, f(\widehat{\Theta}) \le f(\Theta) + O\left(\sqrt{\frac{n^5 + \sigma^2 n^3}{TN}}\right),$$

where $O(\cdot)$-notation hides polynomial dependencies on $1/(1-\alpha)$, $1/\tau_0$, $1/\tau_1, \tau_2$, and $R = \|a\|$.

Algorithm 2 Projected stochastic gradient descent for long sequences
Input: N samples sequences of length T
Output: Learned system $\widehat{\Theta}$

1. Divide each sample of length T into $T/(\beta n)$ samples of length βn where β is a large enough constant. Then run algorithm 1 with the new samples and obtain $\widehat{\Theta}$.

We remark the the gradient computation procedure takes time linear in Tn since one can use chain-rule (also called back-propagation) to compute the gradient efficiently . For completeness, Algorithm 3 gives a detailed implementation. Finally and importantly, we remark that although we defined the population risk as the expected error with respected to sequence of length T, actually our error bound generalizes to any longer (or shorter) sequences of length $T' \gg \max\{n, 1/(1-\alpha)\}$. By the explicit formula for $f(\widehat{\Theta})$ (Lemma 3.5) and the fact that $\|CA^k B\|$ decays exponentially for $k \gg n$ (Lemma 4.4), we can bound the population risk on sequences of different lengths. Concretely, let $f_{T'}(\widehat{\Theta})$ denote the population risk on sequence of length T', we have for all $T' \gg \max\{n, 1/(1-\alpha)\}$,

$$f_{T'}(\widehat{\Theta}) \le 1.1 f(\widehat{\Theta}) + \exp(-(1-\alpha)\min\{T, T'\}) \le O\left(\sqrt{\frac{n^5 + \sigma^2 n^3}{TN}}\right).$$

We note that generalization to longer sequence does deserve attention. Indeed in practice, it's usually difficult to train non-linear recurrent networks that generalize to longer sequences than the training data.

We could hope to achieve linear convergence by showing that the empirical risk also satisfies the weakly-quasi-convexity. Then, we can re-use the samples and hope to use strong optimization tools (such as SVRG) to achieve the linear convergence. This is beyond the scope of this paper and left to future work.

Our proof of Theorem 5.1 simply consists of three parts: a) showing the idealized risk is weakly quasi-convex in the convex set $\mathcal{B}_\alpha$ (Lemma 5.3); b) designing an (almost) unbiased estimator of the gradient of the idealized risk (Lemma 5.4); c) variance bounds of the gradient estimator (Lemma 5.5).

First of all, using the theory developed in Section 3 (Lemma 3.3 and Lemma 3.4), it is straightforward to verify that in the convex set $\mathcal{B}_\alpha \otimes \mathbb{R}^n$, the idealized risk is both weakly-quasi-convex and weakly-smooth.

Lemma 5.3 *Under the condition of Theorem 5.1, the idealized risk (3.2) is τ-weakly-quasi-convex in the convex set $\mathcal{B}_\alpha \otimes \mathbb{R}^n$ and Γ-weakly smooth, where $\tau = \Omega(\tau_0 \tau_1/\tau_2)$ and $\Gamma = O(n^2/\tau_1^4)$.*

Proof [Proof of Lemma 5.3] It suffices to show that for all $\hat{a}, a \in \mathcal{B}_\alpha$, it satisfies $\hat{a} \in \mathcal{N}_\tau(a)$ for $\tau = \Omega(\tau_0 \tau_1 / \tau_2)$. Indeed, by the monotonicity of the family of sets $\mathcal{B}_\alpha$ (Lemma 4.3), we have that $\hat{a}, a \in \mathcal{B}_1$, which by definition means for every z on unit circle, $p_a(z)/z^n, p_{\hat{a}}(z)/z^n \in \mathcal{C}$. By definition of $\mathcal{C}$, for any point $w, \hat{w} \in \mathcal{C}$, the angle ϕ between w and $\hat{w}$ is at most $\pi - \Omega(\tau_0)$ and ratio of the magnitude is at least τ_1/τ_2, which implies that $\Re(w/\hat{w}) = |w|/|\hat{w}| \cdot \cos(\phi) \geq \Omega(\tau_0 \tau_1 / \tau_2)$. Therefore $\Re(p_a(z)/p_{\hat{a}}(z)) \geq \Omega(\tau_0 \tau_1 / \tau_2)$, and we conclude that $\hat{a} \in \mathcal{N}_\tau(a)$. The smoothness bound was established in Lemma 3.4. ∎

Towards designing an unbiased estimator of the gradient, we note that there is a small caveat here that prevents us to just use the gradient of the empirical risk, as commonly done for other (static) problems. Recall that the population risk is defined as the expected risk with *known* initial state h_0, while in the training we don't have access to the initial states and therefore using the naive approach we couldn't even estimate population risk from samples without knowing the initial states.

We argue that being able to handle the missing initial states is indeed desired: in most of the interesting applications h_0 is unknown (or even to be learned). Moreover, the ability of handling unknown h_0 allows us to break a long sequence into shorter sequences, which helps us to obtain Corollary 5.2. Here the difficulty is essentially that we have a supervised learning problem with missing data h_0. We get around it by simply ignoring first $T_1 = \Omega(T)$ outputs of the system and setting the corresponding errors to 0. Since the influence of h_0 to any outputs later than time $k \geq T_1 \gg \max\{n, 1/(1-\alpha)\}$ is inverse exponentially small, we could safely assume $h_0 = 0$ when the error earlier than time T_1 is not taken into account.

This small trick also makes our algorithm suitable to the cases when these early outputs are actually not observed. This is indeed an interesting setting, since in many sequence-to-sequence model Sutskever et al. (2014), there is no output in the first half fraction of iterations (of course these models have non-linear operation that we cannot handle).

The proof of the correctness of the estimator is almost trivial and deferred to Section C.

Lemma 5.4 *Under the assumption of Theorem 5.1, suppose $\hat{a}, a \in \mathcal{B}_\alpha$. Then in Algorithm 1, at each iteration, G_A, G_C are unbiased estimators of the gradient of the idealized risk (3.2) in the sense that:*

$$\mathbb{E}\left[G_A, G_C\right] = \left[\frac{\partial g}{\partial \hat{a}}, \frac{\partial g}{\partial \hat{C}}\right] \pm \exp(-\Omega((1-\alpha)T)).$$

(5.2)

Finally, we control the variance of the gradient estimator.

Lemma 5.5 *The (almost) unbiased estimator (G_A, G_C) of the gradient of $g(\hat{A}, \hat{C})$ has variance bounded by*

$$\mathrm{Var}\left[G_A\right] + \mathrm{Var}\left[G_C\right] \leq \frac{O\left(n^3 \Lambda^2 / \tau_1^6 + \sigma^2 n^2 \Lambda / \tau_1^4\right)}{T}.$$

where $\Lambda = O(\max\{n, 1/(1-\alpha) \log 1/(1-\alpha)\})$.

Note that Lemma 5.5 does not directly follow from the Γ-weakly-smoothness of the population risk, since it's not clear whether the loss function $\ell((x,y),\widehat{\Theta})$ is also Γ-smooth for every sample. Moreover, even if it could work out, from smoothness the variance bound can be only as small as Γ^2, while the true variance scales linearly in $1/T$. Here the discrepancy comes from that smoothness implies an upper bound of the expected squared norm of the gradient, which is equal to the variance plus the expected squared mean. Though typically for many other problems variance is on the same order as the squared mean, here for our sequence-to-sequence model, actually the variance decreases in length of the data, and therefore the bound of variance from smoothness is pessimistic.

We bound directly the variance instead. It's tedious but simple in spirit. We mainly need Lemma 4.4 to control various difference sums that shows up from calculating the expectation. The only tricky part is to obtain the $1/T$ dependency which corresponds to the cancellation of the contribution from the cross terms. In the proof we will basically write out the variance as a (complicated) function of $\hat{A}, \hat{C}$ which consists of sums of terms involving $(\hat{C}\hat{A}^k B - CA^k B)$ and $\hat{A}^k B$. We control these sums using Lemma 4.4. The proof is deferred to Section C.

Finally we are ready to prove Theorem 5.1. We essentially just combine Lemma 5.3, Lemma 5.4 and Lemma 5.5 with the generic convergence Proposition 2.3. This will give us low error in idealized risk and then we relate the idealized risk to the population risk.

Proof [Proof of Theorem 5.1] We consider $g'(\hat{A}, \hat{C}, \hat{D}) = (\hat{D} - D)^2 + g(\hat{A}, \hat{C})$, an extended version of the idealized risk which takes the contribution of $\hat{D}$ into account. By Lemma 5.4 we have that Algorithm 1 computes G_A, G_C which are almost unbiased estimators of the gradients of g' up to negligible error $\exp(-\Omega((1-\alpha)T))$, and by Lemma C.2 we have G_D is an unbiased estimator of g' with respect to $\hat{D}$. Moreover by Lemma 5.5, these unbiased estimator has total variance $V = \frac{O(n^5 + \sigma^2 n^3)}{T}$ where $O(\cdot)$ hides dependency on τ_1 and $(1-\alpha)$. Applying Proposition 2.3 (which only requires an unbiased estimator of the gradient of g'), we obtain that after T iterations, we converge to a point with $g'(\hat{a}, \hat{C}, \hat{D}) \leq O\left(\frac{n^2}{N} + \sqrt{\frac{n^5 + \sigma^2 n^3}{TN}}\right)$. Then, by Lemma 3.5 we have $f(\widehat{\Theta}) \leq g'(\hat{a}, \hat{C}, \hat{D}) + \sigma^2 = g'(\hat{a}, \hat{C}, \hat{D}) + f(\Theta) \leq O\left(\frac{n^2}{N} + \sqrt{\frac{n^5 + \sigma^2 n^3}{TN}}\right) + f(\Theta)$ which completes the proof. ∎

6. The power of improper learning

We observe an interesting and important fact about the theory in Section 5: it solely requires a condition on the characteristic function $p(z)$. This suggests that the geometry of the training objective function depends mostly on the denominator of the transfer function, even though the system is uniquely determined by the transfer function $G(z) = s(z)/p(z)$. This might seem to be an undesirable discrepancy between the behavior of the system and our analysis of the optimization problem.

However, we can actually exploit this discrepancy to design improper learning algorithms that succeed under much weaker assumptions. We rely on the following simple observation about the invariance of a system $G(z) = \frac{s(z)}{p(z)}$. For an arbitrary polynomial $u(z)$ of leading

coefficient 1, we can write $G(z)$ as

$$G(z) = \frac{s(z)u(z)}{p(z)u(z)} = \frac{\tilde{s}(z)}{\tilde{p}(z)},$$

where $\tilde{s} = su$ and $\tilde{p} = pu$. Therefore the system $\tilde{s}(z)/\tilde{p}(z)$ has identical behavior as G. Although this is a redundant representation of $G(z)$, it should counted as an acceptable solution. After all, learning the minimum representation[5] of linear system is impossible in general. In fact, we will encounter an example in Section 6.1.

While not changing the behavior of the system, the extension from $p(z)$ to $\tilde{p}(z)$, does affect the geometry of the optimization problem. In particular, if $\tilde{p}(z)$ is now an α-acquiescent characteristic polynomial as defined in Definition 4.1, then we could find it simply using stochastic gradient descent as shown in Section 5. Observe that we don't require knowledge of $u(z)$ but only its existence. Denoting by d the degree of u, the algorithm itself is simply stochastic gradient descent with $n + d$ model parameters instead of n.

Our discussion motivates the following definition.

Definition 6.1 *A polynomial $p(z)$ of degree n is α-acquiescent by extension of degree d if there exists a polynomial $u(z)$ of degree d and leading coefficient 1 such that $p(z)u(z)$ is α-acquiescent.*

For a transfer function $G(z)$, we define it's $\mathcal{H}_2$ norm as

$$\|G\|_{\mathcal{H}_2}^2 = \frac{1}{2\pi} \int_0^{2\pi} |G(e^{i\theta})|^2 d\theta.$$

We assume (with loss of generality) that the true transfer function $G(z)$ has bounded $\mathcal{H}_2$ norm, that is, $\|G\|_{\mathcal{H}_2} \leq 1$. This can be achieve by a rescaling[6] of the matrix C.

Theorem 6.2 *Suppose the true system has transfer function $G(z) = s(z)/p(z)$ with a characteristic function $p(z)$ that is α-acquiescent by extension of degree d, and $\|G\|_{\mathcal{H}_2} \leq 1$, then projected stochastic gradient descent with $m = n + d$ states (that is, Algorithm 2 with m states) returns a system $\widehat{\Theta}$ with population risk*

$$f(\widehat{\Theta}) \leq O\left(\sqrt{\frac{m^5 + \sigma^2 m^3}{TK}} \right).$$

where the $O(\cdot)$ notation hides polynomial dependencies on $\tau_0, \tau_1, \tau_2, 1/(1-\alpha)$.

The theorem follows directly from Corollary 5.2 (with some additional care about the scaling.

Proof [Proof of Theorem 6.2] Let $\tilde{p}(z) = p(z)u(z)$ be the acquiescent extension of $p(z)$. Since $\tau_2 \geq |u(z)p(z)| = |\tilde{p}(z)| \geq \tau_0$ on the unit circle, we have that $|\tilde{s}(z)| = |s(z)||u(z)| =$

---

5. The minimum representation of a transfer function $G(z)$ is defined as the representation $G(z) = s(z)/p(z)$ with $p(z)$ having minimum degree.

6. In fact, this is a natural scaling that makes comparing error easier. Recall that the population risk is essentially $\|\hat{G} - G\|_{\mathcal{H}_2}$, therefore rescaling C so that $\|G\|_{\mathcal{H}_2} = 1$ implies that when error $\ll 1$ we achieve non-trivial performance.

$s(z) \cdot O_\tau(1/p(z))$. Therefore we have that $\tilde{s}(z)$ satisfies that $\|\tilde{s}\|_{\mathcal{H}_2} = O_\tau(\|s(z)/p(z)\|_{\mathcal{H}_2}) = O_\tau(\|G(z)\|_{\mathcal{H}_2}) \leq O_\tau(1)$. That means that the vector C that determines the coefficients of $\tilde{s}$ satisfies that $\|C\| \leq O_\tau(1)$, since for a polynomial $h(z) = b_0 + \cdots + b_{n-1}z^{n-1}$, we have $\|h\|_{\mathcal{H}_2} = \|b\|$. Therefore we can apply Corollary 5.2 to complete the proof. ∎

In the rest of this section, we discuss in subsection 6.1 the instability of the minimum representation in subsection, and in subsection 6.2 we show several examples where the characteristic function $p(z)$ is not α-acquiescent but is α-acquiescent by extension with small degree d.

As a final remark, the examples illustrated in the following sub-sections may be far from optimally analyzed. It is beyond the scope of this paper to understand the optimal condition under which $p(z)$ is acquiescent by extension.

6.1 Instability of the minimum representation

We begin by constructing a contrived example where the minimum representation of $G(z)$ is not stable at all and as a consequence one can't hope to recover the minimum representation of $G(z)$.

Consider $G(z) = \frac{s(z)}{p(z)} := \frac{z^n - 0.8^{-n}}{(z-0.1)(z^n - 0.9^{-n})}$ and $G'(z) = \frac{s'(z)}{p'(z)} := \frac{1}{z-0.1}$. Clearly these are the minimum representations of the $G(z)$ and $G'(z)$, which also both satisfy acquiescence. On the one hand, the characteristic polynomial $p(z)$ and $p'(z)$ are very different. On the other hand, the transfer functions $G(z)$ and $G'(z)$ have almost the same values on unit circle up to exponentially small error,

$$|G(z) - G'(z)| \leq \frac{0.8^{-n} - 0.9^{-n}}{(z-0.1)(z - 0.9^{-n})} \leq \exp(-\Omega(n)).$$

Moreover, the transfer functions $G(z)$ and $\hat{G}(z)$ are on the order of $\Theta(1)$ on unit circle. These suggest that from an (inverse polynomially accurate) approximation of the transfer function $G(z)$, we cannot hope to recover the minimum representation in any sense, even if the minimum representation satisfies acquiescence.

6.2 Power of improper learning in various cases

We illustrate the use of improper learning through various examples below.

6.2.1 Example: artificial construction

We consider a simple contrived example where improper learning can help us learn the transfer function dramatically. We will show an example of characteristic function which is not 1-acquiescent but $(\alpha+1)/2$-$(\alpha+1)/2$-acquiescent by extension of degree 3.

Let n be a large enough integer and α be a constant. Let $J = \{1, n-1, n\}$ and $\omega = e^{2\pi i/n}$, and then define $p(z) = z^3 \prod_{j \in [n], j \notin J}(z - \alpha\omega^j)$. Therefore we have that

$$p(z)/z^n = z^3 \prod_{j \in [n], j \in J}(1 - \alpha\omega^j/z) = \frac{1 - \alpha^n/z^n}{(1 - \omega/z)(1 - \omega^{-1}/z)(1 - 1/z)} \tag{6.1}$$

21

Taking $z = e^{-i\pi/2}$ we have that $p(z)/z^n$ has argument (phase) roughly $-3\pi/4$, and therefore it's not in $\mathcal{C}$, which implies that $p(z)$ is *not* 1-acquiescent. On the other hand, picking $u(z) = (z - \omega)(z - 1)(z - \omega^{-1})$ as the helper function, from equation (6.1) we have $p(z)u(z)/z^{n+3} = 1 - \alpha^n/z^n$ takes values inverse exponentially close to 1 on the circle with radius $(\alpha + 1)/2$. Therefore $p(z)u(z)$ is $(\alpha + 1)/2$-acquiescent.

6.2.2 EXAMPLE: CHARACTERISTIC FUNCTION WITH SEPARATED ROOTS

A characteristic polynomial with well separated roots will be acquiescent by extension. Our bound will depend on the following quantity of p that characterizes the separateness of the roots.

Definition 6.3 *For a polynomial $h(z)$ of degree n with roots $\lambda_1, \ldots, \lambda_n$ inside unit circle, define the quantity $\Gamma(\cdot)$ of the polynomial h as:*

$$\Gamma(h) := \sum_{j \in [n]} \left| \frac{\lambda_j^n}{\prod_{i \neq j}(\lambda_i - \lambda_j)} \right| .$$

Lemma 6.4 *Suppose $p(z)$ is a polynomial of degree n with distinct roots inside circle with radius α. Let $\Gamma = \Gamma(p)$, then $p(z)$ is α-acquiescent by extension of degree $d = O(\max\{(1 - \alpha)^{-1}\log(\sqrt{n}\Gamma \cdot \|p\|_{\mathcal{H}_2}), 0\})$.*

Our main idea to extend $p(z)$ by multiplying some polynomial u that approximates p^{-1} (in a relatively weak sense) and therefore pu will always take values in the set $\mathcal{C}$. We believe the following lemma should be known though for completeness we provide the proof in Section D.

Lemma 6.5 (Approximation of inverse of a polynomial) *Suppose $p(z)$ is a polynomial of degree n and leading coefficient 1 with distinct roots inside circle with radius α, and $\Gamma = \Gamma(p)$. Then for $d = O(\max\{(\frac{1}{1-\alpha}\log\frac{\Gamma}{(1-\alpha)\zeta}, 0\})$, there exists a polynomial $h(z)$ of degree d and leading coefficient 1 such that for all z on unit circle,*

$$\left| \frac{z^{n+d}}{p(z)} - h(z) \right| \leq \zeta .$$

Proof [Proof of Lemma 6.4] Let $\gamma = 1 - \alpha$. Using Lemma 6.5 with $\zeta = 0.5\|p\|_{\mathcal{H}_\infty}^{-1}$, we have that there exists polynomial u of degree $d = O(\max\{\frac{1}{1-\alpha}\log(\Gamma\|p\|_{\mathcal{H}_\infty}), 0\})$ such that

$$\left| \frac{z^{n+d}}{p(z)} - u(z) \right| \leq \zeta .$$

Then we have that
$$\left| p(z)u(z)/z^{n+d} - 1 \right| \leq \zeta|p(z)| < 0.5 .$$

Therefore $p(z)u(z)/z^{n+d} \in \mathcal{C}_{\tau_0, \tau_1, \tau_2}$ for constant τ_0, τ_1, τ_2. Finally noting that for degree n polynomial we have $\|h\|_{\mathcal{H}_\infty} \leq \sqrt{n} \cdot \|h\|_{\mathcal{H}_2}$, which completes the proof. ∎

6.2.3 EXAMPLE: CHARACTERISTIC POLYNOMIAL WITH RANDOM ROOTS

We consider the following generative model for characteristic polynomial of degree $2n$. We generate n complex numbers $\lambda_1, \ldots, \lambda_n$ uniformly randomly on circle with radius $\alpha < 1$, and take $\lambda_i, \bar{\lambda}_i$ for $i = 1, \ldots, n$ as the roots of $p(z)$. That is, $p(z) = (z - \lambda_1)(z - \bar{\lambda}_1) \ldots (z - \lambda_n)(z - \bar{\lambda}_n)$. We show that with good probability (over the randomness of λ_i's), polynomial $p(z)$ will satisfy the condition in subsection 6.2.2 so that it can be learned efficiently by our improper learning algorithm.

Theorem 6.6 *Suppose $p(z)$ with random roots inside circle of radius α is generated from the process described above. Then with high probability over the choice of p, we have that $\Gamma(p) \le \exp(\widetilde{O}(\sqrt{n}))$ and $\|p\|_{\mathcal{H}_2} \le \exp(\tilde{O}(\sqrt{n}))$. As a corollary, $p(z)$ is α-acquiescent by extension of degree $\widetilde{O}((1-\alpha)^{-1}n)$.*

Towards proving Theorem 6.6, we need the following lemma about the expected distance of two random points with radius ρ and r in log-space.

Lemma 6.7 *Let $x \in \mathbb{C}$ be a fixed point with $|x| = \rho$, and λ uniformly drawn on the circle with radius r. Then $\mathbb{E}\left[\ln |x - \lambda|\right] = \ln \max\{\rho, r\}$.*

Proof When $r \ne \rho$, let N be an integer and $\omega = e^{2i\pi/N}$. Then we have that

$$\mathbb{E}[\ln |x - \lambda| \mid r] = \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} \ln |x - r\omega^k| \tag{6.2}$$

The right hand of equation (6.2) can be computed easily by observing that $\frac{1}{N} \sum_{k=1}^{N} \ln |x - r\omega^k| = \frac{1}{N} \ln \left| \prod_{k=1}^{N}(x - r\omega^k) \right| = \frac{1}{N} \ln |x^N - r^N|$. Therefore, when $\rho > r$, we have $\lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} \ln |x - r\omega^k| = \lim_{N \to \infty} \rho + \frac{1}{N} \ln |(x/\rho)^N - (r/\rho)^N| = \ln \rho$. On the other hand, when $\rho < r$, we have that $\lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} \ln |x - r\omega^k| = \ln r$. Therefore we have that $\mathbb{E}[\ln |x - \lambda| \mid r] = \ln(\max \rho, r)$. For $\rho = r$, similarly proof (with more careful concern of regularity condition) we can show that $\mathbb{E}[\ln |x - \lambda| \mid r] = \ln r$. ∎

Now we are ready to prove Theorem 6.6.

Proof [Proof of Theorem 6.6] Fixing index i, and the choice of λ_i, we consider the random variable $Y_i = \ln(\frac{|\lambda_i|^{2n}}{\prod_{j \ne i} |\lambda_i - \lambda_j| \prod_{j \ne i} |\lambda_i - \bar{\lambda}_j|}) n \ln |\lambda_i| - \sum_{j \ne i} \ln |\lambda_i - \lambda_j|$. By Lemma 6.7, we have that $\mathbb{E}[Y_i] = n \ln |\lambda_i| - \sum_{j \ne i} \mathbb{E}[\ln |\lambda_i - \lambda_j|] = \ln(1 - \delta)$. Let $Z_j = \ln |\lambda_i - \lambda_j|$. Then we have that Z_j are random variable with mean 0 and ψ_1-Orlicz norm bounded by 1 since $\mathbb{E}[e^{\ln |\lambda_i - \lambda_j|} - 1] \le 1$. Therefore by Bernstein inequality for sub-exponential tail random variable (for example, (Ledoux and Talagrand, 2013, Theorem 6.21)), we have that with high probability $(1 - n^{-10})$, it holds that $\left| \sum_{j \ne i} Z_j \right| \le \widetilde{O}(\sqrt{n})$ where $\widetilde{O}$ hides logarithmic factors. Therefore, with high probability, we have $|Y_i| \le \widetilde{O}(\sqrt{n})$.

Finally we take union bound over all $i \in [n]$, and obtain that with high probability, for $\forall i \in [n], |Y_i| \le \widetilde{O}(\sqrt{n})$, which implies that $\sum_{i=1}^{n} \exp(Y_i) \le \exp(\widetilde{O}(\sqrt{n}))$. With similar technique, we can prove that $\|p\|_{\mathcal{H}_2} \le \exp(\tilde{O}(\sqrt{n}))$. ∎

6.2.4 EXAMPLE: PASSIVE SYSTEMS

We will show that with improper learning we can learn almost all passive systems, an important class of stable linear dynamical system as we discussed earlier. We start off with the definition of a strict-input passive system.

Definition 6.8 (Passive System, c.f Kottenstette and Antsaklis (2010)) *A SISO linear system is strict-input passive if and only if for some $\tau_0 > 0$ and any z on unit circle, $\Re(G(z)) \geq \tau_0$.*

In order to learn the passive system, we need to add assumptions in the definition of strict passivity. To make it precise, we define the following subsets of complex plane: For positive constant τ_0, τ_1, τ_2, define

$$\mathcal{C}^+_{\tau_0, \tau_1, \tau_2} = \{z \in \mathbb{C} : |z| \leq \tau_2, \Re(z) \geq \tau_1, \Re(z) \geq \tau_0 |\Im(z)| \}. \tag{6.3}$$

We say a transfer function $G(z) = s(z)/p(z)$ is (τ_0, τ_1, τ_2)-strict input passive if for any z on unit circle we have $G(z) \in \mathcal{C}^+_{\tau_0, \tau_1, \tau_2}$. Note that for small constant τ_0, τ_1 and large constant τ_2, this basically means the system is strict-input passive.

Now we are ready to state our main theorem in this subsection. We will prove that passive systems could be learned improperly with a constant factor more states (dimensions), assuming $s(z)$ has all its roots strictly inside unit circles and $\Gamma(s) \leq \exp(O(n))$.

Theorem 6.9 *Suppose $G(z) = s(z)/p(z)$ is (τ_0, τ_1, τ_2)-strict-input passive. Moreover, suppose the roots of $s(z)$ have magnitudes inside circle with radius α and $\Gamma = \Gamma(s) \leq \exp(O(n))$ and $\|p\|_{\mathcal{H}_2} \leq \exp(O(n))$. Then $p(z)$ is α-acquiescent by extension of degree $d = O_{\tau,\alpha}(n)$, and as a consequence we can learn $G(z)$ with $n + d$ states in polynomial time.*

Moreover, suppose in addition we assume that $G(z) \in \mathcal{C}_{\tau_0, \tau_1, \tau_2}$ for every z on unit circle. Then $p(z)$ is α-acquiescent by extension of degree $d = O_{\tau,\alpha}(n)$.

The proof of Theorem 6.9 is similar in spirit to that of Lemma 6.4, and is deferred to Section D.

6.3 Improper learning using linear regression

In this subsection, we show that under stronger assumption than α-acquiescent by extension, we can improperly learn a linear dynamical system with linear regression, up to some fixed bias.

The basic idea is to fit a linear function that maps $[x_{k-\ell}, \dots, x_k]$ to y_k. This is equivalent to a dynamical system with ℓ hidden states and with the companion matrix A in (1.4) being chosen as $a_\ell = 1$ and $a_{\ell-1} = \cdots = a_1 = 0$. In this case, the hidden states exactly memorize all the previous ℓ inputs, and the output is a linear combination of the hidden states.

Equivalently, in the frequency space, this corresponds to fitting the transfer function $G(z) = s(z)/p(z)$ with a rational function of the form $\frac{c_1 z^{\ell-1} + \cdots + c_1}{z^{\ell-1}} = c_1 z^{-(\ell-1)} + \cdots + c_n$. The following is a sufficient condition on the characteristic polynomial $p(x)$ that guarantees the existence of such fitting,

Definition 6.10 *A polynomial $p(z)$ of degree n is extremely-acquiescent by extension of degree d with bias ε if there exists a polynomial $u(z)$ of degree d and leading coefficient 1 such that for all z on unit circle,*

$$\left| p(z)u(z)/z^{n+d} - 1 \right| \le \varepsilon \tag{6.4}$$

We remark that if $p(z)$ is 1-acquiescent by extension of degree d, then there exists $u(z)$ such that $p(z)u(z)/z^{n+d} \in \mathcal{C}$. Therefore, equation (6.4) above is a much stronger requirement than acquiescence by extension.[7]

When $p(z)$ is extremely-acquiescent, we see that the transfer function $G(z) = s(z)/p(z)$ can be approximated by $s(z)u(z)/z^{n+d}$ up to bias ε. Let $\ell = n+d+1$ and $s(z)u(z) = c_1 z^{\ell-1} + \cdots + c_\ell$. Then we have that $G(z)$ can be approximated by the following dynamical system of ℓ hidden states with ε bias: we choose $A = \mathrm{CC}(a)$ with $a_\ell = 1$ and $a_{\ell-1} = \cdots = a_1 = 0$, and $C = [c_1, \ldots, c_\ell]$. As we have argued previously, such a dynamical system simply memorizes all the previous ℓ inputs, and therefore it is equivalent to linear regression from the feature $[x_{k-\ell}, \ldots, x_k]$ to output y_k.

Proposition 6.11 (Informal) *If the true system $G(z) = s(z)/p(z)$ satisfies that $p(z)$ is extremely-acquiescent by extension of degree d. Then using linear regression we can learn mapping from $[x_{k-\ell}, \ldots, x_k]$ to y_k with bias ε and polynomial sampling complexity.*

We remark that with linear regression the bias ε will only go to zero as we increase the length ℓ of the feature, but not as we increase the number of samples. Moreover, linear regression requires a stronger assumption than the improper learning results in previous subsections do. The latter can be viewed as an interpolation between the proper case and the regime where linear regression works.

7. Learning multi-input multi-output (MIMO) systems

We consider multi-input multi-output systems with the transfer functions that have a common denominator $p(z)$,

$$G(z) = \frac{1}{p(z)} \cdot S(z) \tag{7.1}$$

where $S(z)$ is an $\ell_{\mathrm{in}} \times \ell_{\mathrm{out}}$ matrix with each entry being a polynomial with real coefficients of degree at most n and $p(z) = z^n + a_1 z^{n-1} + \cdots + a_n$. Note that here we use ℓ_{in} to denote the dimension of the inputs of the system and ℓ_{out} the dimension of the outputs.

Although a special case of a general MIMO system, this class of systems still contains many interesting cases, such as the transfer functions studied in Fazel et al. (2001, 2004), where $G(z)$ is assumed to take the form $G(z) = R_0 + \sum_{i=1}^{n} \frac{R_i}{z - \lambda_i}$, for $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ with conjugate symmetry and $R_i \in \mathbb{C}^{\ell_{\mathrm{out}} \times \ell_{\mathrm{in}}}$ satisfies that $R_i = \bar{R}_j$ whenever $\lambda_i = \bar{\lambda}_j$.

In order to learn the system $G(z)$, we parametrize $p(z)$ by its coefficients $a_1, \ldots, a_n$ and $S(z)$ by the coefficients of its entries. Note that each entry of $S(z)$ depends on $n + 1$ real

---

7. We need $(1 - \delta)$-acquiescence by extension in previous subsections for small $\delta > 0$, though this is merely additional technicality needed for the sample complexity. We ignore this difference between $1 - \delta$-acquiescence and 1-acquiescence and for the purpose of this subsection

coefficients and therefore the collection of coefficients forms a third order tensor of dimension $\ell_{\mathrm{out}} \times \ell_{\mathrm{in}} \times (n+1)$. It will be convenient to collect the leading coefficients of the entries of $S(z)$ into a matrix of dimension $\ell_{\mathrm{out}} \times \ell_{\mathrm{in}}$, named D, and the rest of the coefficients into a matrix of dimension $\ell_{\mathrm{out}} \times \ell_{\mathrm{in}}n$, denoted by C. This will be particularly intuitive when a state-space representation is used to learn the system with samples as discussed later. We parameterize the training transfer function $\hat{G}(z)$ by $\hat{a}$, $\hat{C}$ and $\hat{D}$ using the same way.

Let's define the risk function in the frequency domain as,

$$g(\hat{A}, \hat{C}, \hat{D}) = \int_0^{2\pi} \left\| G(e^{i\theta}) - \hat{G}(e^{i\theta}) \right\|_F^2 d\theta . \tag{7.2}$$

The following lemma is an analog of Lemma 3.3 for the MIMO case. Itss proof actually follows from a straightforward extension of the proof of Lemma 3.3 by observing that matrix $S(z)$ (or $\hat{S}(z)$) commute with scalar $p(z)$ and $\hat{p}(z)$, and that $\hat{S}(z), \hat{p}(z)$ are linear in $\hat{a}, \hat{C}$.

Lemma 7.1 *The risk function* $g(\hat{a}, \hat{C})$ *defined in* (7.2) *is* τ-*weakly-quasi-convex in the domain*

$$\mathcal{N}_\tau(a) = \left\{ \hat{a} \in \mathbb{R}^n : \Re\left(\frac{p_a(z)}{p_{\hat{a}}(z)} \right) \geq \tau/2, \forall\ z \in \mathbb{C},\ \mathrm{s.t.}\ |z| = 1 \right\} \otimes \mathbb{R}^{\ell_{\mathrm{in}} \times \ell_{\mathrm{out}} \times n'}$$

Finally, as alluded before, we use a particular state space representation for learning the system in time domain with example sequences. It is known that any transfer function of the form (7.1) can be realized uniquely by the state space system of the following special case of Brunovsky normal form Brunovsky (1970),

$$A = \begin{bmatrix} 0 & I_{\ell_{\mathrm{in}}} & 0 & \cdots & 0 \\ 0 & 0 & I_{\ell_{\mathrm{in}}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & I_{\ell_{\mathrm{in}}} \\ -a_n I_{\ell_{\mathrm{in}}} & -a_{n-1} I_{\ell_{\mathrm{in}}} & -a_{n-2} I_{\ell_{\mathrm{in}}} & \cdots & -a_1 I_{\ell_{\mathrm{in}}} \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ I_{\ell_{\mathrm{in}}} \end{bmatrix}, \tag{7.3}$$

and,

$$C \in \mathbb{R}^{\ell_{\mathrm{out}} \times n\ell_{\mathrm{in}}}, \quad D \in \mathbb{R}^{\ell_{\mathrm{out}} \times \ell_{\mathrm{in}}} .$$

The following Theorem is a straightforward extension of Corollary 5.2 and Theorem 6.2 to the MIMO case.

Theorem 7.2 *Suppose transfer function $G(z)$ of a MIMO system takes form* (7.1), *and has norm $\|G\|_{\mathcal{H}_2} \leq 1$. If the common denominator $p(z)$ is α-acquiescent by extension of degree d then projected stochastic gradient descent over the state space representation* (7.3) *will return $\widehat{\Theta}$ with risk*

$$f(\widehat{\Theta}) \leq \frac{\mathrm{poly}(n + d, \sigma, \tau, (1-\alpha)^{-1})}{TN} .$$

We note that since A and B are simply the tensor product of $I_{\ell_{\mathrm{in}}}$ with $CC(a)$ and e_n, the no blow-up property (Lemma 4.4) for $A^k B$ still remains true. Therefore to prove Theorem 7.2, we essentially only need to run the proof of Lemma 5.5 with matrix notation and matrix norm. We defer the proof to the full version.

8. Simulations

In this section, we provide proof-of-concepts experiments on synthetic data. We will demonstrate that

1) plain SGD tends to blow up even with relatively small learning rate, especially on hard instances

2) SGD with our projection step converges with reasonably large learning rate, and with over-parameterization the final error is competitive

3) SGD with gradient clipping has the strongest performance in terms both of the convergence speed and the final error

Here gradient clipping refers to the technique of using a normalized gradient instead of the true gradient. Specifically, for some positive hyper parameter B, we follow the approximate gradient

$$g_{\mathrm{clip}} = \begin{cases} g & \text{if } \|g\| \leq B \\ Bg/\|g\| & otherwise \end{cases}$$

This method is commonly applied in training recurrent neural networks Pascanu et al. (2013).

Bullet 1) suggests that stability is indeed a real concern. Bullet 2) corroborates our theoretical study. Finding 3) suggests the instability of SGD partly arises from the noise in the batches, and such noise is reduced by the gradient clipping. Our experiments suggest that the landscape of the objective function may be even nicer than what is predicted by our theoretical development. It remains possible that the objective has no non-global local minima, possibly even outside the convex set to which our algorithm projects.

We generate the true system with state dimension $d = 20$ by randomly picking the conjugate pairs of roots of the characteristic polynomial inside the circle with radius $\rho = 0.95$ and randomly generating the vector C from standard normal distribution. The distribution of the norm of the impulse response r (defined in Section 3) of such systems has a heavy-tail. When the norm of r is several magnitudes larger than the median it's difficult to learn the system. Thus we select systems with reasonable $\|r\|$ for experiments, and we observe that the difficulty of learning increases as $\|r\|$ increases. The inputs of the dynamical model are generated from standard normal distribution with length $T = 500$. We note that we generate new fresh inputs and outputs at every iterations and therefore the training loss is equal to the test loss (in expectation.) We use initial learning rate 0.01 in the projected gradient descent and SGD with gradient clipping. We use batch size 100 for all experiments, and decay the learning rate at 200K and 250K iteration by a factor of 10 in all experiments.
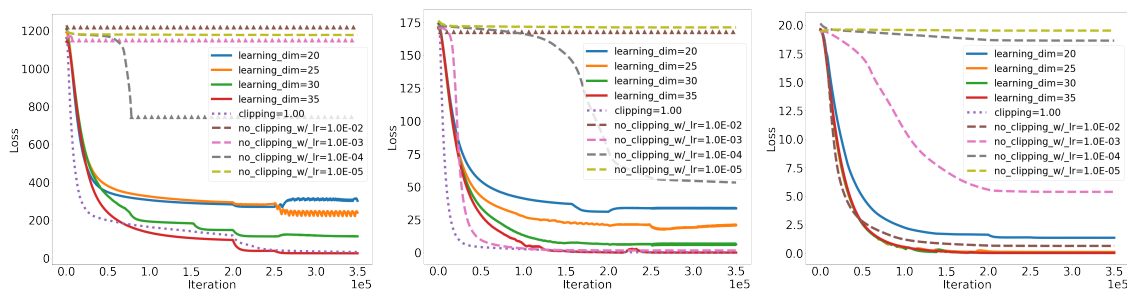
Figure 2: The performance of projected stochastic gradient descent with over-parameterization, vanilla SGD, and SGD with gradient clipping, on three different instance of dynamical systems with true state dimension = 20. The solid lines are from our proposed projected SGD with (over-parameterized) state dimension = 20, 25, 30, 35. The dot line corresponds to SGD with gradient clipped to Frobenius norm 1. The dashed lines correspond vanilla SGD and the triangle marker means the error blows up to infinity. The plot demonstrates the effect of the over-parameterization to our our algorithm. We note that the loss are different scales because the true systems in these three instances have different norms of impulse responses (which is equal to the loss of zero fitting).

References

Alexandre S. Bazanella, Michel Gevers, Ljubisa Miskovic, and Brian D.O. Anderson. Iterative minimization of h_2 control performance criteria. *Automatica*, 44:2549–2559, 2008.

Badri Narayan Bhaskar, Gongguo Tang, and Benjamin Recht. Atomic norm denoising with applications to line spectral estimation. *IEEE Transactions on Signal Processing*, 61(23): 5987–5999, 2013.

Léon Bottou. On-line learning in neural networks. chapter On-line Learning and Stochastic Approximations, pages 9–42. Cambridge University Press, New York, NY, USA, 1998. ISBN 0-521-65263-4. URL http://dl.acm.org/citation.cfm?id=304710.304720.

Pavol Brunovsky. A classification of linear controllable systems. *Kybernetika*, 06(3):(173)– 188, 1970. URL http://eudml.org/doc/28376.

M. C. Campi and Erik Weyer. Finite sample properties of system identification methods. *IEEE Transactions on Automatic Control*, 47(8):1329–1334, 2002.

Hugh Durrant-Whyte and Tim Bailey. Simultaneous localization and mapping: part i. *Robotics & Automation Magazine, IEEE*, 13(2):99–110, 2006.

Diego Eckhard and Alexandre Sanfelice Bazanella. On the global convergence of identification of output error models. In *Proc. 18th IFAC World congress*, 2011.

T. Estermann. *Complex numbers and functions*. Athlone Press, 1962. URL https://books.google.com/books?id=ITbvAAAAMAAJ.

Maryam Fazel, Haitham Hindi, and Stephen P Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proc. American Control Conference*, volume 6, pages 4734–4739. IEEE, 2001.

Maryam Fazel, Haitham Hindi, and S Boyd. Rank minimization and applications in system theory. In *Proc. American Control Conference*, volume 4, pages 3273–3278. IEEE, 2004.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points - online stochastic gradient for tensor decomposition. In *Proc. 28th COLT*, pages 797–842, 2015.

E. Hazan, K. Y. Levy, and S. Shalev-Shwartz. Beyond Convexity: Stochastic Quasi-Convex Optimization. *ArXiv e-prints*, July 2015.

Christiaan Heij, André Ran, and Freek van Schagen. *Introduction to mathematical systems theory : linear systems, identification and control.* Birkhäuser, Basel, Boston, Berlin, 2007. ISBN 3-7643-7548-5. URL `http://opac.inria.fr/record=b1130636`.

Joao P Hespanha. *Linear systems theory.* Princeton university press, 2009.

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

Nicholas Kottenstette and Panos J Antsaklis. Relationships between positive real, passive dissipative, & positive systems. In *American Control Conference (ACC), 2010*, pages 409–416. IEEE, 2010.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 2013.

J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient Descent Converges to Minimizers. *ArXiv e-prints*, February 2016.

Sergey Levine and Vladlen Koltun. Guided policy search. In *Proceedings of The 30th International Conference on Machine Learning*, pages 1–9, 2013.

Lennart Ljung. *System Identification. Theory for the user.* Prentice Hall, Upper Saddle River, NJ, 2nd edition, 1998.

Alexandre Megretski. Convex optimization in robust identification of nonlinear feedback. In *Proceedings of the 47th Conference on Decision and Control*, 2008.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pages 1310–1318, 2013.

Ali Rahimi, Ben Recht, and Trevor Darrell. Learning appearance manifolds from video. In *Proc. IEEE CVPR*, 2005.

A. C. Schaeffer. Inequalities of a. markoff and s. bernstein for polynomials and related functions. *Bull. Amer. Math. Soc.*, 47(8):565–579, 08 1941. URL `http://projecteuclid.org/euclid.bams/1183503783`.

Parikshit Shah, Badri Narayan Bhaskar, Gongguo Tang, and Benjamin Recht. Linear system identification via atomic norm regularization. In *Proceedings of the 51st Conference on Decision and Control*, 2012.

Torsten Söderström and Petre Stoica. Some properties of the output error method. *Automatica*, 18(1):93–99, 1982.

Petre Stoica and Torsten Söderström. Uniqueness of prediction error estimates of multi-variable moving average models. *IFAC Proceedings Volumes*, 15(4):199–204, 1982.

Petre Stoica and Torsten Söderström. Uniqueness of estimated k-step prediction models of arma processes. *Systems & control letters*, 4(6):325–331, 1984.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proc. 27th NIPS*, pages 3104–3112, 2014. URL http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.

M. Vidyasagar and Rajeeva L. Karandikar. A learning theory approach to system identification and stochastic adaptive control. *Journal of Process Control*, 18(3):421–430, 2008.

Erik Weyer and M. C. Campi. Finite sample properties of system identification methods. In *Proceedings of the 38th Conference on Decision and Control*, 1999.

Appendix A. Background on optimization

The proof below uses the standard analysis of gradient descent for non-smooth objectives and demonstrates that the argument still works for weakly-quasi-convex functions.

Proof [Proof of Proposition 2.3] We start by using the weakly-quasi-convex condition and then the rest follows a variant of the standard analysis of non-smooth projected sub-gradient descent[8]. We conditioned on θ_k, and have that

$$
\begin{aligned}
\tau(f(\theta_k) - f(\theta^*)) &\leq \nabla f(\theta_k)^\top (\theta_k - \theta^*) = \mathbb{E}[\mathfrak{r}(\theta_k)^\top (\theta_k - \theta^*) \mid \theta_k] \\
&= \mathbb{E}\left[\frac{1}{\eta}(\theta_k - w_{k+1})(\theta_k - \theta^*) \mid \theta_k\right] \\
&= \frac{1}{\eta}\left(\mathbb{E}\left[\|\theta_k - w_{k+1}\|^2 \mid \theta_k\right] + \|\theta_k - \theta^*\|^2 - \mathbb{E}\left[\|w_{k+1} - \theta^*\|^2 \mid \theta_k\right]\right) \\
&= \eta\, \mathbb{E}\left[\|\mathfrak{r}(\theta_k)\|^2\right] + \frac{1}{\eta}\left(\|\theta_k - \theta^*\|^2 - \mathbb{E}\left[\|w_{k+1} - \theta^*\|^2 \mid \theta_k\right]\right) \qquad \text{(A.1)}
\end{aligned}
$$

where the first inequality uses weakly-quasi-convex and the rest of lines are simply algebraic manipulations. Since θ_{k+1} is the projection of w_{k+1} to $\mathcal{B}$ and θ^* belongs to $\mathcal{B}$, we have $\|w_{k+1} - \theta^*\| \geq \|\theta_{k+1} - \theta^*\|$. Together with (A.1), and

$$
\mathbb{E}\left[\|\mathfrak{r}(\theta_k)\|^2\right] = \|\nabla f(\theta_k)\|^2 + \mathrm{Var}[\mathfrak{r}(\theta_k)] \leq \Gamma(f(\theta_k) - f(\theta^*)) + V,
$$

---

8. Although we used weak smoothness to get a slightly better bound

30

we obtain that

$$\tau(f(\theta_k) - f(\theta^*)) \leq \eta\Gamma(f(\theta_k) - f(\theta^*)) + \eta V + \frac{1}{\eta}\left(\|\theta_k - \theta^*\|^2 - \mathbb{E}\left[\|\theta_{k+1} - \theta^*\|^2 \mid \theta_k\right]\right).$$

Taking expectation over all the randomness and summing over k we obtain that

$$\sum_{k=0}^{K-1} \mathbb{E}\left[f(\theta_k) - f(\theta^*)\right] \leq \frac{1}{\tau - \eta\Gamma}\left(\eta KV + \frac{1}{\eta}\|\theta_0 - \theta^*\|^2\right) \leq \frac{1}{\tau - \eta\Gamma}\left(\eta KV + \frac{1}{\eta}R^2\right).$$

where we use the assumption that $\|\theta_0 - \theta^*\| \leq R$. Suppose $K \geq \frac{4R^2\Gamma^2}{V\tau^2}$, then we take $\eta = \frac{R}{\sqrt{VK}}$. Therefore we have that $\tau - \eta\Gamma \geq \tau/2$ and therefore

$$\sum_{k=0}^{K-1} \mathbb{E}\left[f(\theta_k) - f(\theta^*)\right] \leq \frac{4R\sqrt{V}\sqrt{K}}{\tau}. \tag{A.2}$$

On the other hand, if $K \leq \frac{4R^2\Gamma^2}{V\tau^2}$, we pick $\eta = \frac{\tau}{2\Gamma}$ and obtain that

$$\sum_{k=0}^{K-1} \mathbb{E}\left[f(\theta_k) - f(\theta^*)\right] \leq \frac{2}{\tau}\left(\frac{\tau KV}{2\Gamma} + \frac{2\Gamma R^2}{\tau}\right) \leq \frac{8\Gamma R^2}{\tau^2}. \tag{A.3}$$

Therefore using equation (A.3) and (A.2) we obtain that when choosing η properly according to K as above,

$$\mathbb{E}_{k\in[K]}\left[f(\theta_k) - f(\theta^*)\right] \leq \max\left\{\frac{8\Gamma R^2}{\tau^2 K}, \frac{4R\sqrt{V}}{\tau\sqrt{K}}\right\}.$$

$\blacksquare$

Appendix B. Toolbox

Lemma B.1 *Let $B = e_n \in \mathbb{R}^{n\times 1}$ and $\lambda \in [0, 2\pi]$, $w \in \mathbb{C}$. Suppose A with $\rho(A) \cdot |w| < 1$ has the controllable canonical form $A = \mathrm{CC}(a)$. Then*

$$(I - wA)^{-1}B = \frac{1}{p_a(w^{-1})}\begin{bmatrix} w^{-1} \\ w^{-2} \\ \vdots \\ w^{-n} \end{bmatrix}$$

where $p_a(x) = x^n + a_1 x^{n-1} + \cdots + a_n$ is the characteristic polynomial of A.

Proof let $v = (I - wA)^{-1}B$ then we have $(I - wA)v = B$. Note that $B = e_n$, and $I - wA$ is of the form

$$I - wA = \begin{bmatrix} 1 & -w & 0 & \cdots & 0 \\ 0 & 1 & -w & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -w \\ a_n w & a_{n-1}w & a_{n-2}w & \cdots & 1 + a_1 w \end{bmatrix} \tag{B.1}$$

31

Therefore we obtain that $v_k = wv_{k+1}$ for $1 \leq k \leq n-1$. That is, $v_k = v_0 w^{-k}$ for $v_0 = v_1 w^1$. Using the fact that $((I-wA)v)_n = 1$, we obtain that $v_0 = p_a(w^{-1})^{-1}$ where $p_a(\cdot)$ is the polynomial $p_a(x) = x^n + a_1 x^{n-1} + \cdots + a_n$. Then we have that $u(I-wA)^{-1}B = \frac{u_1 w^{-1} + \cdots + u_n w^{-n}}{p_a(w^{-1})}$ ∎

Lemma B.2 *Suppose* $x_1, \ldots, x_n$ *are independent variables with mean 0 and covariance matrices and* $I, U_1, \ldots, U_d$ *are fixed matrices, then*

$$\mathbb{E}\left[\|\sum_{k=1}^n U_k x_k\|^2\right] = \sum_{k=1}^n \|U_k\|_F^2 .$$

Proof We have that

$$\mathbb{E}\left[\|\sum_{k=1}^n U_k x_k\|_F^2\right] = \mathbb{E}\sum_{k,\ell}^n \mathrm{tr}(U_k x_k x_\ell^\top U_\ell^\top) = \sum_k^n \mathrm{tr}(U_k x_k x_k^\top U_k^\top) = \sum_{k=1}^n \|U_k\|_F^2$$

∎

Appendix C. Missing proofs in Sections 4 and 5

C.1 Monotonicity of acquiescence: Proof of Lemma 4.3

Lemma C.1 (Lemma 4.3 restated) *For any* $0 < \alpha < \beta$, *we have that* $\mathcal{B}_\alpha \subset \mathcal{B}_\beta$.

Proof Let $q_a(z) = 1 + a_1 z + \cdots + a_n z^n$. Note that $q(z^{-1}) = p_a(z)/z^n$. Therefore we note that $B_\alpha = \{a : q_a(z) \in \mathcal{C}, \forall |z| = 1/\alpha\}$. Suppose $a \in \mathcal{B}_\alpha$, then $\Re(q_a(z)) \geq \tau_1$ for any z with $|z| = 1/\alpha$. Since $\Re(q_a(z))$ is the real part of the holomorphic function $q_a(z)$, its a harmonic function. By maximum (minimum) principle of the harmonic functions, we have that for any $|z| \leq 1/\alpha$, $\Re(q_a(z)) \geq \inf_{|z|=1/\alpha} \Re(q_a(z)) \geq \tau_1$. In particular, it holds that for $|z| = 1/\beta < 1/\alpha$, $\Re(q_a(z)) \geq \tau_1$. Similarly we can prove that for z with $|z| = 1/\beta$, $\Re(q_a(z)) \geq (1+\tau_0)\Im(q_a(z))$, and other conditions for a being in $\mathcal{B}_\beta$. ∎

C.2 Proof of Lemma 5.4

Lemma 5.4 follows directly from the following general Lemma which also handles the multi-input multi-output case. It can be seen simply from calculation similar to the proof of Lemma 3.5. We mainly need to control the tail of the series using the no-blow up property (Lemma 4.4) and argue that the wrong value of the initial states h_0 won't cause any trouble to the partial loss function $\ell((x,y), \widehat{\Theta})$ (defined in Algorithm 1). This is simply because after time $T_1 = T/4$, the influence of the initial state is already washed out.

Lemma C.2 *In algorithm 3 the values of* G_A, G_C, G_D *are equal to the gradients of* $g(\hat{A}, \hat{C}) + (\hat{D} - D)^2$ *with respect to* $\hat{A}$, $\hat{C}$ *and* $\hat{D}$ *up to inverse exponentially small error.*

Proof [Proof of Lemma C.2] We first show that the partial empirical loss function $\ell((x,y), \widehat{\Theta})$ has expectation almost equal to the idealized risk (up to the term for $\hat{D}$ and exponential small error),

$$\mathbb{E}[\ell((x,y), \widehat{\Theta})] = g(\hat{A}, \hat{C}) + (\hat{D} - D)^2 \pm \exp(-\Omega((1-\alpha)T)).$$

This can be seen simply from similar calculation to the proof of Lemma 3.5. Note that

$$y_t = Dx_t + \sum_{k=1}^{t-1} CA^{t-k-1}Bx_k + CA^{t-1}h_0 + \xi_t \quad \text{and} \quad \tilde{y}_t = \hat{D}x_t + \sum_{k=1}^{t-1} \hat{C}\hat{A}^{t-k-1}\hat{B}x_k. \quad \text{(C.1)}$$

Therefore noting that when $t \geq T_1 \geq \Omega(T)$, we have that $\|CA^{t-1}h_0\| \leq \exp(-\Omega((1-\alpha)T)$ and therefore the effect of h_0 is negligible. Then we have that

$$\mathbb{E}[\ell((x,y),\widehat{\Theta})] = \frac{1}{T - T_1} \mathbb{E}\left[\sum_{t>T_1}^{T} \|y_t - y_t\|^2\right] \pm \exp(-\Omega((1-\alpha)T))$$

$$= \|\hat{D} - D\|^2 + \frac{1}{T - T_1} \sum_{T \geq t > T_1} \sum_{0 \leq j \leq t-1} \|\hat{C}\hat{A}^j B - CA^j B\|^2 \pm \exp(-\Omega((1-\alpha)T))$$

$$= \|\hat{D} - D\|^2 + \sum_{j=0}^{T_1} \|\hat{C}\hat{A}^j B - CA^j B\|^2 + \sum_{T \geq j \geq T_1} \frac{T - j}{T - T_1} \|\hat{C}\hat{A}^j B - CA^j B\|^2 \pm \exp(-\Omega((1-\alpha)T))$$

$$= \|\hat{D} - D\|^2 + \sum_{j=0}^{\infty} \|\hat{C}\hat{A}^j B - CA^j B\|^2 \pm \exp(-\Omega((1-\alpha)T)).$$

where the first line use the fact that $\|CA^{t-1}h_0\| \leq \exp(-\Omega((1-\alpha)T)$, the second uses equation (3.10) and the last line uses the no-blowing up property of $A^k B$ (Lemma 4.4).

Similarly, we can prove that the gradient of $\mathbb{E}[\ell((x,y),\widehat{\Theta})]$ is also close to the gradient of $g(\hat{A}, \hat{C}) + (\hat{D} - D)^2$ up to inverse exponential error. ∎

C.3 Proof of Lemma 5.5

Proof [Proof of Lemma 5.5] Both G_A and G_C can be written in the form of a quadratic form (with vector coefficients) of $x_1, \ldots, x_T$ and $\xi_1, \ldots, \xi_T$. That is, we will write

$$G_A = \sum_{s,t} x_s x_t u_{st} + \sum_{s,t} x_s \xi_t u'_{st} \quad \text{and} \quad G_C = \sum_{s,t} x_s x_t v_{st} + \sum_{s,t} x_s \xi_t v'_{st}.$$

where u_{st} and v_{st} are vectors that will be calculated later. By Lemma C.3, we have that

$$\text{Var}\left[\sum_{s,t} x_s x_s u_{st} + \sum_{s,t} x_s \xi_t u'_{st}\right] \leq O(1) \sum_{s,t} \|u_{st}\|^2 + O(\sigma^2) \sum_{s,t} \|u'_{st}\|^2. \quad \text{(C.2)}$$

Therefore in order to bound from above $\text{Var}[G_A]$, it suffices to bound $\sum \|u_{st}\|^2$ and $\sum \|u'_{st}\|^2$, and similarly for G_C.

We begin by writing out u_{st} for fixed $s, t \in [T]$ and bounding its norm. We use the same set of notations as int the proof of Lemma 5.4. Recall that we set $r_k = CA^k B$ and $\hat{r}_k = \hat{C}\hat{A}^k B$, and $\Delta r_k = \hat{r}_k - r_k$. Moreover, let $z_k = \hat{A}^k B$. We note that the sums of $\|z_k\|^2$

and r_k^2 can be controlled. By the assumption of the Lemma, we have that

$$\sum_{k=t}^{\infty} \|z_k\|^2 \le 2\pi n \tau_1^{-2}, \quad \|z_k\|^2 \le 2\pi n \alpha^{2k-2n} \tau_1^{-2}. \tag{C.3}$$

$$\sum_{k=t}^{\infty} \Delta r_k^2 \le 4\pi n \tau_1^{-2}, \quad \|\Delta r_k\|^2 \le 4\pi n \alpha^{2k-2n} \tau_1^{-2}. \tag{C.4}$$

which will be used many times in the proof that follows.

We calculate the explicit form of G_A using the explicit back-propagation Algorithm 3. We have that in Algorithm 3,

$$\hat{h}_k = \sum_{j=1}^{k} \hat{A}^{k-j} B x_j = \sum_{j=1}^{k} z_{k-j} x_j \tag{C.5}$$

and

$$\Delta h_k = \sum_{j=k}^{T} (\hat{A}^\top)^{j-k} \hat{C}^\top \Delta y_j = \sum_{j=k}^{T} \alpha^{j-k} (\hat{A}^\top)^{j-k} \hat{C}^\top \mathbf{1}_{j>T_1} \left(\xi_j + \sum_{\ell=1}^{j} \Delta r_{j-\ell} x_\ell \right) \tag{C.6}$$

Then using $G_A = \sum_{k \ge 2} B^\top \Delta h_k h_{k-1}'^\top$ and equation (C.5) and equation (C.6) above, we have that

$$u_{st} = \sum_{k=2}^{T} \left(\sum_{j \ge \max\{k,s,T_1+1\}} \Delta r_{j-s} \hat{C} \hat{A}^{j-k} B \right) \mathbf{1}_{k \ge t+1} \cdot \hat{A}^{k-t-1} B$$

$$= \sum_{k=2}^{T} \left(\sum_{j \ge \max\{k,s,T_1+1\}} \Delta r_{j-s} \hat{r}_{j-k} \right) \mathbf{1}_{k \ge t+1} \cdot z_{k-t-1}. \tag{C.7}$$

and that,

$$u_{st}' = \sum_{k=2}^{T} z_{k-1-s} \cdot \mathbf{1}_{k \ge s+1} \cdot \hat{r}_{t-k}' \cdot \mathbf{1}_{t > \max\{T_1,k\}} = \sum_{s+1 \le k \le t} z_{k-1-s} \cdot \hat{r}_{t-k}' \cdot \mathbf{1}_{t > \max\{T_1\}} \tag{C.8}$$

Towards bounding $\|u_{st}\|$, we consider four different cases. Let $\Lambda = \Omega \left(\{ \max\{n, (1-\alpha)^{-1} \log(\frac{1}{1-\alpha})\} \} \right)$ be a threshold.

Case 1: When $0 \le s - t \le \Lambda$, we rewrite u_{st} by rearranging equation (C.7),

$$u_{st} = \sum_{T \ge k \ge s} z_{k-t-1} \sum_{j \ge \max\{k,T_1+1\}} \Delta r_{j-s} \hat{r}_{j-k} + \sum_{t < k < s} z_{k-t-1} \sum_{j \ge \max\{s,T_1+1\}} \Delta r_{j-s} \hat{r}_{j-k}$$

$$= \sum_{\ell \ge 0, \ell \ge T_1+1-s} \Delta r_\ell \sum_{s \le k \le l+s, k \le T} \hat{r}_{\ell+s-k} z_{k-t-1} + \sum_{\ell \ge 0, \ell \ge T_1+1-s} \Delta r_\ell \sum_{s > k > t} \hat{r}_{\ell+s-k} z_{k-t-1}$$

where at the second line, we did the change of variables $\ell = j - s$. Then by Cauchy-Schartz inequality, we have,

$$
\|u_{st}\|^2 \leq 2 \left(\sum_{\ell \geq 0, \ell \geq T_1 + 1 - s} \Delta r_\ell^2 \right) \underbrace{\left(\sum_{\ell \geq 0, \ell \geq T_1 + 1 - s} \left\| \sum_{s \leq k \leq l+s, k \leq T} \hat{r}_{\ell+s-k} z_{k-t-1} \right\|^2 \right)}_{T_1}
$$

$$
+ 2 \left(\sum_{\ell \geq 0, \ell \geq T_1 + 1 - s} \Delta r_\ell^2 \right) \underbrace{\left(\sum_{\ell \geq 0, \ell \geq T_1 + 1 - s} \left\| \sum_{s > k > t} \hat{r}_{\ell+s-k} z_{k-t-1} \right\|^2 \right)}_{T_2} . \tag{C.9}
$$

We could bound the contribution from Δr_k^2 ssing equation (C.4), and it remains to bound terms T_1 and T_2. Using the tail bounds for $\|z_k\|$ (equation (C.3)) and the fact that $|\hat{r}_k| = |\hat{C} \hat{A}^k B| \leq \|\hat{A}^k B\| = \|z_k\|$, we have that

$$
T_1 = \sum_{\ell \geq 0, \ell \geq T_1 + 1 - s} \left\| \sum_{s \leq k \leq l+s, k \leq T} \hat{r}_{\ell+s-k} z_{k-t-1} \right\|^2 \leq \sum_{\ell \geq 0} \left(\sum_{s \leq k \leq \ell+s} |\hat{r}_{\ell+s-k}| \|z_{k-t-1}\| \right)^2 . \tag{C.10}
$$

We bound the inner sum of RHS of (C.10) using the fact that $\|z_k\|^2 \leq O(n\alpha^{2k-2n}/\tau_1^2)$ and obtain that,

$$
\sum_{s \leq k \leq \ell+s} |\hat{r}_{\ell+s-k}| \|z_{k-t-1}\| \leq \sum_{s \leq k \leq \ell+s} O(n\alpha^{(\ell+s-t-1)-2n}/\tau_1^2)
$$

$$
\leq O(\ell n \alpha^{(\ell+s-t-1)-2n}/\tau_1^2) . \tag{C.11}
$$

Note that equation (C.11) is particular effective when $\ell > \Lambda$. When $\ell \leq \Lambda$, we can refine the bound using equation (C.3) and obtain that

$$
\sum_{s \leq k \leq \ell+s} |\hat{r}_{\ell+s-k}| \|z_{k-t-1}\| \leq \left(\sum_{s \leq k \leq \ell+s} |\hat{r}_{\ell+s-k}|^2 \right)^{1/2} \left(\sum_{s \leq k \leq \ell+s} \|z_{k-t-1}\|^2 \right)^{1/2}
$$

$$
\leq O(\sqrt{n}/\tau_1) \cdot O(\sqrt{n}/\tau_1) = O(n/\tau_1^2) . \tag{C.12}
$$

Plugging equation (C.12) and (C.11) into equation (C.10), we have that

$$
\sum_{\ell \geq 0} \left(\sum_{s \leq k \leq \ell+s} |\hat{r}_{\ell+s-k}| \|z_{k-t-1}\| \right)^2 \leq \sum_{\Lambda \geq \ell \geq 0} O(n^2/\tau_1^4) + \sum_{\ell > \Lambda} O(\ell^2 n^2 \alpha^{2(\ell+s-t-1)-4n}/\tau_1^4)
$$

$$
\leq O(n^2 \Lambda/\tau_1^4) + O(n^2/\tau_1^4) = O(n^2 \Lambda/\tau_1^4) . \tag{C.13}
$$

For the second term in equation (C.9), we bound similarly,

$$
T_2 \leq \sum_{\ell \geq 0, \ell \geq T_1 + 1 - s} \left\| \sum_{s > k > t} \hat{r}_{\ell+s-k} z_{k-t-1} \right\|^2 \leq O(n^2 \Lambda/\tau_1^4) . \tag{C.14}
$$

Therefore using the bounds for T_1 and T_2 we obtain that,

$$\|u_{st}\|^2 \le O(n^3 \Lambda / \tau_1^6) \tag{C.15}$$

Case 2: When $s - t > \Lambda$, we tighten equation (C.13) by observing that,

$$T_1 \le \sum_{\ell \ge 0} \left(\sum_{s \le k \le \ell + s} |\hat{r}_{\ell+s-k}| \|z_{k-t-1}\| \right)^2 \le \alpha^{2(s-t-1)-4n} \sum_{\ell \ge 0} O(\ell^2 n^2 \alpha^{2\ell} / \tau_1^4)$$

$$\le \alpha^{s-t-1} \cdot O(n^2 / (\tau_1^4 (1-\alpha)^3)). \tag{C.16}$$

where we used equation (C.11). Similarly we can prove that

$$T_2 \le \alpha^{s-t-1} \cdot O(n^2 / (\tau_1^4 (1-\alpha)^3)).$$

Therefore, we have when $s - t \ge \Lambda$,

$$\|u_{st}\|^2 \le O(n^3 / ((1-\alpha)^3 \tau_1^6)) \cdot \alpha^{s-t-1}. \tag{C.17}$$

Case 3: When $-\Lambda \le s-t \le 0$, we can rewrite $u_s t$ and use the Cauchy-Schwartz inequality and obtain that

$$u_{st} = \sum_{T \ge k \ge t+1} z_{k-t-1} \sum_{j \ge \max\{k, T_1+1\}} \Delta r_{j-s} \hat{r}_{j-k} = \sum_{\ell \ge 0, \ell \ge T_1+1-s} \Delta r_\ell \sum_{t+1 \le k \le l+s, k \le T} \hat{r}_{\ell+s-k} z_{k-t-1}.$$

and,

$$\|u_{st}\|^2 \le \left(\sum_{\ell \ge 0, \ell \ge T_1+1-s} \Delta r_\ell^2 \right) \left(\sum_{\ell \ge 0, \ell \ge T_1+1-s} \left\| \sum_{t+1 \le k \le l+s, k \le T} \hat{r}_{\ell+s-k} z_{k-t-1} \right\|^2 \right).$$

Using almost the same arguments as in equation (C.11) and (C.12), we that

$$\sum_{t+1 \le k \le \ell+s} |\hat{r}_{\ell+s-k}| \cdot \|z_{k-t-1}\| \le O(\ell n \alpha^{(\ell+s-t-1)-2n} / \tau_1^2)$$

and $\quad \displaystyle\sum_{t+1 \le k \le \ell+s} |\hat{r}_{\ell+s-k}| \cdot \|z_{k-t-1}\| \le O(\sqrt{n}/\tau_1) \cdot O(\sqrt{n}/\tau_1) = O(n/\tau_1^2).$

Then using a same type of argument as equation (C.13), we can have that

$$\sum_{\ell \ge 0, \ell \ge T_1+1-s} \left\| \sum_{t+1 \le k \le l+s, k \le T} \hat{r}'_{\ell+s-k} z'_{k-t-1} \right\|^2 \le O(n^2 \Lambda / \tau_1^4) + O(n^2 / \tau_1^4)$$

$$= O(n^2 \Lambda / \tau_1^4).$$

It follows that in this case $\|u_{st}\|$ can be bounded with the same bound in (C.15).

36

Case 4: When $s - t \leq -\Lambda$, we use a different simplification of u_{st} from above. First of all, it follows (C.7) that

$$\|u_{st}\| \leq \sum_{k=2}^{T} \left(\sum_{j \geq \max\{k,s,T_1+1\}} \|\Delta r_{j-s} \hat{r}'_{j-k} z_{k-t-1}\| \mathbf{1}_{k \geq t+1} \right) \tag{C.18}$$

$$\leq \sum_{k \geq t+1} \|z'_{k-t-1}\| \sum_{j \geq \max\{k,T_1+1\}} |\Delta r_{j-s} \hat{r}'_{j-k}|.$$

Since $j - s \geq k - s > 4n$ and it follows that

$$\sum_{j \geq \max\{k,T_1+1\}} |\Delta r_{j-s} \hat{r}'_{j-k}| \leq \sum_{j \geq \max\{k,T_1+1\}} O(\sqrt{n}/\tau_1 \cdot \alpha^{j-s-n}) \cdot O(\sqrt{n}/\tau_1 \cdot \alpha^{j-k-n})$$

$$\leq O(n/(\tau_1^2(1-\alpha)) \cdot \alpha^{k-s-n})$$

Then we have that

$$\|u_{st}\|^2 \leq \sum_{k \geq t+1} \|z'_{k-t-1}\| \sum_{j \geq \max\{k,T_1+1\}} |\Delta r_{j-s} \hat{r}'_{j-k}|$$

$$\leq \left(\sum_{k \geq t+1} \|z'_{k-t-1}\|^2 \right) \left(\sum_{k \geq t+1} \left(\sum_{j \geq \max\{k,T_1+1\}} |\Delta r_{j-s} \hat{r}'_{j-k}| \right) \right)^2$$

$$\leq O(n/\tau_1^2) \cdot O(n^2/(\tau_1^4(1-\alpha)^3)\alpha^{t-s}) = O(n^3/(\tau_1^6 \delta^3)\alpha^{t-s})$$

Therefore, using the bound for $\|u_{st}\|^2$ obtained in the four cases above, taking sum over s,t, we obtain that

$$\sum_{1 \leq s,t \leq T} \|u_{st}\|^2 \leq \sum_{s,t \in [T]:|s-t| \leq \Lambda} O(n^3 \Lambda/\tau_1^6) + \sum_{s,t:|s-t| \geq \Lambda} O(n^3/(\tau_1^6(1-\alpha)^3)\alpha^{|t-s|-1})$$

$$\leq O(Tn^3 \Lambda^2/\tau_1^6) + O(n^3/\tau_1^6) = O(Tn^3 \Lambda^2/\tau_1^6). \tag{C.19}$$

We finished the bounds for $\|u_{st}\|$ and now we turn to bound $\|u'_{st}\|^2$. Using the formula for u'_{st} (equation C.8), we have that for $t \leq s+1$, $u'_{st} = 0$. For $s + \Lambda \geq t \geq s + 2$, we have that by Cauchy-Schwartz inequality,

$$\|u'_{st}\| \leq \left(\sum_{s+1 \leq k \leq t} \|z_{k-1-s}\|^2 \right)^{1/2} \left(\sum_{s+1 \leq k \leq t} |\hat{r}'_{t-k}|^2 \right)^{1/2} \leq O(n/\tau_1^2) \leq O(n/\tau_1^2).$$

On the other hand, for $t > s + \Lambda$, by the bound that $|\hat{r}'_k|^2 \leq \|z'_k\|^2 \leq O(n\alpha^{2k-2n}/\tau_1^2)$, we have,

$$\|u'_{st}\| \leq \sum_{s+1 \leq k \leq t-1}^{T} \|z_{k-1-s}\| \cdot |\hat{r}'_{t-k}| \leq \sum_{s+1 \leq k \leq t-1}^{T} n\alpha^{t-s-1}/\tau_1^2$$

$$\leq O(n(t-s)\alpha^{t-s-1}/\tau_1^2).$$

37

Therefore taking sum over s, t, similarly to equation (C.19),

$$\sum_{s,t \in [T]} \|u'_{st}\|^2 \leq O(Tn^2\Lambda/\tau_1^4).\tag{C.20}$$

Then using equation (C.2) and equation (C.19) and (C.20), we obtain that

$$\mathrm{Var}[\|G_A\|^2] \leq O\left(Tn^3\Lambda^2/\tau_1^6 + \sigma^2 Tn^2\Lambda/\tau_1^4\right).$$

Hence, it follows that

$$\mathrm{Var}[G_A] \leq \frac{1}{(T-T_1)^2}\mathrm{Var}[G_A] \leq \frac{O\left(n^3\Lambda^2/\tau_1^6 + \sigma^2 n^2\Lambda/\tau_1^4\right)}{T}.$$

We can prove the bound for G_C similarly. ∎

Lemma C.3 *Let $x_1, \ldots, x_T$ be independent random variables with mean 0 and variance 1 and 4-th moment bounded by $O(1)$, and u_{ij} be vectors for $i, j \in [T]$. Moreover, let $\xi_1, \ldots, \xi_T$ be independent random variables with mean 0 and variance σ^2 and u'_{ij} be vectors for $i, j \in [T]$. Then,*

$$\mathrm{Var}\left[\sum_{i,j} x_i x_j u_{ij} + \sum_{i,j} x_i \xi_j u'_{ij}\right] \leq O(1)\sum_{i,j}\|u_{ij}\|^2 + O(\sigma^2)\sum_{i,j}\|u'_{ij}\|^2.$$

Proof

Note that the two sums in the target are independent with mean 0, therefore we only need to bound the variance of both sums individually. The proof follows the linearity of expectation and the independence of x_i's:

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{i,j} x_i x_j u_{ij}\right\|^2\right] &= \sum_{i,j}\sum_{k,\ell}\mathbb{E}\left[x_i x_j x_k x_\ell u_{ij}^\top u_{k\ell}\right]\\
&= \sum_i \mathbb{E}[u_{ii}^\top u_{ii} x_i^4] + \sum_{i \neq j}\mathbb{E}[u_{ii}^\top u_{jj} x_i^2 x_j^2] + \sum_{i,j}\mathbb{E}\left[x_i^2 x_j^2(u_{ij}^\top u_{ij} + u_{ij}^\top u_{ji})\right]\\
&\leq \sum_{i,j} u_{ii}^\top u_{jj} + O(1)\sum_{i,j}\|u_{ij} + u_{ji}\|^2\\
&= \|\textstyle\sum_i u_{ii}\|^2 + O(1)\sum_{i,j}\|u_{ij}\|^2
\end{aligned}$$

where at second line we used the fact that for any monomial x^α with an odd degree on one of the x_i's, $\mathbb{E}[x^\alpha] = 0$. Note that $\mathbb{E}[\sum_{i,j} x_i x_j u_{ij}] = \sum_i u_{ii}$. Therefore,

$$\mathrm{Var}\left[\sum_{i,j} x_i x_j u_{ij}\right] = \mathbb{E}\left[\|\textstyle\sum_{i,j} x_i x_j u_{ij}\|^2\right] - \|\mathbb{E}[\textstyle\sum_{i,j} x_i x_j u_{ij}]\|^2 \leq O(1)\sum_{i,j}\|u_{ij}\|^2 \quad\text{(C.21)}$$

Similarly, we can control $\mathrm{Var}\left[\sum_{i,j} x_i \xi_j u'_{ij}\right]$ by $O(\sigma^2)\sum_{i,j}\|u'_{ij}\|^2$. ∎

Appendix D. Missing proofs in Section 6

D.1 Proof of Lemma 6.5

Towards proving Lemma 6.5, we use the following lemma to express the inverse of a polynomial as a sum of inverses of degree-1 polynomials.

Lemma D.1 *Let $p(z) = (z - \lambda_1) \ldots (z - \lambda_n)$ where λ_j's are distinct. Then we have that*

$$\frac{1}{p(z)} = \sum_{j=1}^{n} \frac{t_j}{z - \lambda_j}, \qquad where \quad t_j = \left(\prod_{i \neq j} (\lambda_j - \lambda_i) \right)^{-1}. \tag{D.1}$$

Proof [Proof of Lemma D.1] By interpolating constant function at points $\lambda_1, \ldots, \lambda_n$ using Lagrange interpolating formula, we have that

$$1 = \sum_{j=1}^{n} \frac{\prod_{i \neq j}(x - \lambda_i)}{\prod_{i \neq j}(\lambda_j - \lambda_i)} \cdot 1 \tag{D.2}$$

Dividing $p(z)$ on both sides we obtain equation (D.1). ∎

The following lemma computes the Fourier transform of function $1/(z - \lambda)$.

Lemma D.2 *Let $m \in \mathbb{Z}$, and $\mathcal{K}$ be the unit circle in complex plane, and $\lambda \in \mathbb{C}$ inside the $\mathcal{K}$. Then we have that*

$$\int_{\mathcal{K}} \frac{z^m}{z - \lambda} \, \mathrm{d}z = \begin{cases} 2\pi i \lambda^m & \text{for } m \geq 0 \\ 0 & \text{o.w.} \end{cases}$$

Proof [Proof of Lemma D.2] For $m \geq 0$, since z^m is a holomorphic function, by Cauchy's integral formula, we have that

$$\int_{\mathcal{K}} \frac{z^m}{z - \lambda} \, \mathrm{d}z = 2\pi i \lambda^m.$$

For $m < 0$, by changing of variable $y = z^{-1}$ we have that

$$\int_{\mathcal{K}} \frac{z^m}{z - \lambda} \, \mathrm{d}z = \int_{\mathcal{K}} \frac{y^{-m-1}}{1 - \lambda y} \, \mathrm{d}y.$$

since $|\lambda y| = |\lambda| < 1$, then we by Taylor expansion we have,

$$\int_{\mathcal{K}} \frac{y^{-m-1}}{1 - \lambda y} \, \mathrm{d}y = \int_{\mathcal{K}} y^{-m-1} \left(\sum_{k=0}^{\infty} (\lambda y)^k \right) \mathrm{d}y.$$

Since the series λy is dominated by $|\lambda|^k$ which converges, we can switch the integral with the sum. Note that y^{-m-1} is holomorphic for $m < 0$, and therefore we conclude that

$$\int_{\mathcal{K}} \frac{y^{-m-1}}{1 - \lambda y} \, \mathrm{d}y = 0.$$

■

Now we are ready to prove Lemma 6.5.

Proof [Proof of Lemma 6.5] Let $m = n + d$. We compute the Fourier transform of $z^m/p(z)$. That is, we write

$$\frac{e^{im\theta}}{p(e^{i\theta})} = \sum_{k=-\infty}^{\infty} \beta_k e^{ik\theta} \,.$$

where

$$\beta_k = \frac{1}{2\pi} \int_0^{2\pi} \frac{e^{i(m-k)\theta}}{p(e^{i\theta})} d\theta = \frac{1}{2\pi i} \int_{\mathcal{K}} \frac{z^{m-k-1}}{p(z)} \, dz$$

By Lemma D.1, we write

$$\frac{1}{p(z)} = \sum_{j=1}^{n} \frac{t_j}{z - \lambda_j} \,.$$

Then it follows that

$$\beta_k = \frac{1}{2\pi i} \sum_{j=1}^{n} t_j \int_{\mathcal{K}} \frac{z^{m-k-1}}{z - \lambda_j} \, dz$$

Using Lemma D.2, we obtain that

$$\beta_k = \begin{cases} \sum_{j=1}^{n} t_j \lambda_j^{m-k-1} & \text{if } -\infty \leq k \leq m-1 \\ 0 & \text{o.w.} \end{cases} \tag{D.3}$$

We claim that

$$\sum_{j=1}^{n} t_j \lambda_j^{n-1} = 1 \,, \qquad \text{and} \qquad \sum_{j=1}^{n} t_j \lambda_j^s = 0 \,, \quad 0 \leq s < n-1 \,.$$

Indeed these can be obtained by writing out the lagrange interpolation for polynomial $f(x) = x^s$ with $s \leq n - 1$ and compare the leading coefficient. Therefore, we further simplify β_k to

$$\beta_k = \begin{cases} \sum_{j=1}^{n} t_j \lambda_j^{m-k-1} & \text{if } -\infty < k < m-n \\ 1 & \text{if } k = m-n \\ 0 & \text{o.w.} \end{cases} \tag{D.4}$$

Let $h(z) = \sum_{k \geq 0} \beta_k z^k$. Then we have that $h(z)$ is a polynomial with degree $d = m - n$ and leading term 1. Moreover, for our choice of d,

$$\left| \frac{z^m}{p(z)} - h(z) \right| = \left| \sum_{k<0} \beta_k z^k \right| \leq \sum_{k<0} |\beta_k| \leq \max_j |t_j|(1 - \lambda_j)^n \sum_{k<0} (1 - \gamma)^{d-k-1}$$

$$\leq \Gamma(1 - \gamma)^d/\gamma < \zeta \,.$$

■

40

D.2 Proof of Theorem 6.9

Theorem 6.9 follows directly from a combination of Lemma D.3 and Lemma D.4 below. Lemma D.3 shows that the denominator of a function (under the stated assumptions) can be extended to a polynomial that takes values in $\mathcal{C}^+$ on unit circle. Lemma D.4 shows that it can be further extended to another polynomial that takes values in $\mathcal{C}$.

Lemma D.3 *Suppose the roots of s are inside circle with radius $\alpha < 1$, and $\Gamma = \Gamma(s)$. If transfer function $G(z) = s(z)/p(z)$ satisfies that $G(z) \in \mathcal{C}_{\tau_0,\tau_1,\tau_2}$ (or $G(z) \in \mathcal{C}^+_{\tau_0,\tau_1,\tau_2}$) for any z on unit circle, then there exists $u(z)$ of degree $d = O_\tau(\max\{(\frac{1}{1-\alpha}\log\frac{\sqrt{n}\Gamma\cdot\|p\|_{\mathcal{H}_2}}{1-\alpha}, 0\})$ such that $p(z)u(z)/z^{n+d} \in \mathcal{C}_{\tau'_0,\tau'_1,\tau'_2}$ (or $p(z)u(z)/z^{n+d} \in \mathcal{C}^+_{\tau'_0,\tau'_1,\tau'_2}$ respectively) for $\tau' = \Theta_\tau(1)$, where $O_\tau(\cdot), \Theta_\tau(\cdot)$ hide the polynomial dependencies on τ_0, τ_1, τ_2.*

Proof [Proof of Lemma D.3] By the fact that $G(z) = s(z)/p(z) \in \mathcal{C}_{\tau_0,\tau_1,\tau_2}$, we have that $p(z)/s(z) \in \mathcal{C}_{\tau'_0,\tau'_1,\tau'_2}$ for some τ' that polynomially depend on τ. Using Lemma 6.5, there exists $u(z)$ of degree d such that

$$\left|\frac{z^{n+d}}{s(z)} - u(z)\right| \leq \zeta.$$

where we set $\zeta \ll \min\{\tau'_0, \tau'_1\}/\tau'_2 \cdot \|p\|_{\mathcal{H}_\infty}^{-1}$. Then we have that

$$\left|p(z)u(z)/z^{n+d} - \frac{p(z)}{s(z)}\right| \leq |p(z)|\zeta \ll \min\{\tau'_0, \tau'_1\}. \tag{D.5}$$

It follows from equation (D.5) implies that that $p(z)u(z)/z^{n+d} \in \mathcal{C}_{\tau''_0,\tau''_1,\tau''_2}$, where τ'' polynomially depends on τ. The same proof still works when we replace $\mathcal{C}$ by $\mathcal{C}^+$. ∎

Lemma D.4 *Suppose $p(z)$ of degree n and leading coefficient 1 satisfies that $p(z) \in \mathcal{C}^+_{\tau_0,\tau_1,\tau_2}$ for any z on unit circle. Then there exists $u(z)$ of degree d such that $p(z)u(z)/z^{n+d} \in \mathcal{C}_{\tau'_0,\tau'_1,\tau'_2}$ for any z on unit circle with $d = O_\tau(n)$ and $\tau'_0, \tau'_1, \tau'_2 = \Theta_\tau(1)$, where $O_\tau(\cdot), \Theta_\tau(\cdot)$ hide the dependencies on τ_0, τ_1, τ_2.*

Proof [Proof of Lemma D.4] We fix z on unit circle first. Let's defined $\sqrt{p(z)/z^n}$ be the square root of $p(z)/z^n$ with principle value. Let's write $p(z)/z^n = \tau_2(1 + (\frac{p(z)}{\tau_2 z^n} - 1))$ and we take Taylor expansion for $\frac{1}{\sqrt{p(z)/z^n}} = \tau_2^{-1/2}(1 + (\frac{p(z)}{\tau_2 z^n} - 1))^{-1/2} = \tau_2^{-1/2}\left(\sum_{k=0}^\infty (\frac{p(z)}{\tau_2 z^n} - 1)^k\right)$. Note that since $\tau_1 < |p(z)| < \tau_2$, we have that $|\frac{p(z)}{\tau_2 z^n} - 1| < 1 - \tau_1/\tau_2$. Therefore truncating the Taylor series at $k = O_\tau(1)$ we obtain a polynomial a rational function $h(z)$ of the form

$$h(z) = \sum_{j\geq 0}^k (\frac{p(z)}{\tau_2 z^n} - 1)^j,$$

which approximates $\frac{1}{\sqrt{p(z)/z^n}}$ with precision $\zeta \ll \min\{\tau_0, \tau_1\}/\tau_2$, that is, $\left|\frac{1}{\sqrt{p(z)/z^n}} - h(z)\right| \leq \zeta$. Therefore, we obtain that $\left|\frac{p(z)h(z)}{z^n} - \sqrt{p(z)/z^n}\right| \leq \zeta|p(z)/z^n| \leq \zeta\tau_2$. Note that since

$p(z)/z^n \in \mathcal{C}_{\tau_0,\tau_1,\tau_2}^+$, we have that $\sqrt{p(z)/z^n} \in \mathcal{C}_{\tau_0',\tau_1',\tau_2'}$ for some constants $\tau_0', \tau_1', \tau_2'$. Therefore $\frac{p(z)h(z)}{z^n} \in \mathcal{C}_{\tau_0',\tau_1',\tau_2'}$. Note that $h(z)$ is not a polynomial yet. Let $u(z) = z^{nk}h(z)$ and then $u(z)$ is polynomial of degree at most nk and $p(z)u(z)/z^{(n+1)k} \in \mathcal{C}_{\tau_0',\tau_1',\tau_2'}$ for any z on unit circle. ∎

Appendix E. Back-propagation implementation

In this section we give a detailed implementation of using back-propagation to compute the gradient of the loss function. The algorithm is for general MIMO case with the parameterization (7.3). To obtain the SISO sub-case, simply take $\ell_{\text{in}} = \ell_{\text{out}} = 1$.

Algorithm 3 Back-propagation

Parameters: $\hat{a} \in \mathbb{R}^n$, $\hat{C} \in \mathbb{R}^{\ell_{\text{in}} \times n\ell_{\text{out}}}$, and $\hat{D} \in \mathbb{R}^{\ell_{\text{in}} \times \ell_{\text{out}}}$. Let $\hat{A} = \text{MCC}(\hat{a}) = \text{CC}(\hat{a}) \otimes I_{\ell_{\text{in}}}$ and $\hat{B} = e_n \otimes I_{\ell_{\text{in}}}$.
Input: samples $((x^{(1)}, y^1), \ldots, x^{(N)}, y^{(N)})$ and projection set $\mathcal{B}_\alpha$.

for each sample $(x^{(j)}, y^j) = ((x_1, \ldots, x_T), (y_1, \ldots, y_T))$ **do**
  Feed-forward pass:
    $h_0 = 0 \in \mathbb{R}^{n\ell_{\text{in}}}$.
    for $k = 1$ to T
      $\hat{h}_k \leftarrow \hat{A}h_{k-1} + \hat{B}x_k$, $\hat{y}_t \leftarrow \hat{C}h_k + \hat{D}x_k$ and $\hat{h}_k \leftarrow \hat{A}h_{k-1} + \hat{B}x_k$.
    end for
  Back-propagation:
    $\Delta h_{T+1} \leftarrow 0$, $G_A \leftarrow 0$, $G_C \leftarrow 0$. $G_D \leftarrow 0$
    $T_1 \leftarrow T/4$
    for $k = T$ to 1
      if $k > T_1$, $\Delta y_k \leftarrow \hat{y}_k - y_k$, o.w. $\Delta y_k \leftarrow 0$. Let $\Delta h_k \leftarrow \hat{C}^\top \Delta y_k + \hat{A}^\top \Delta h_{k+1}$.
      update $G_C \leftarrow G_C + \frac{1}{T-T_1}\Delta y_k \hat{h}_k$, $G_A \leftarrow G_A - \frac{1}{T-T_1}B^\top \Delta h_k \hat{h}_{k-1}^\top$, and $G_D \leftarrow G_D + \frac{1}{T-T_1}\Delta y_k x_k$.
    end for
  Gradient update: $\hat{A} \leftarrow \hat{A} - \eta \cdot G_A$, $\hat{C} \leftarrow \hat{C} - \eta \cdot G_C$, $\hat{D} \leftarrow \hat{D} - \eta \cdot G_D$.
  Projection step: Obtain $\hat{a}$ from $\hat{A}$ and set $\hat{a} \leftarrow \Pi_{\mathcal{B}}(\hat{a})$, and $\hat{A} = \text{MCC}(\hat{a})$
end for

Appendix F. Projection to the set $\mathcal{B}_\alpha$

In order to have a fast projection algorithm to the convex set $\mathcal{B}_\alpha$, we consider a grid $\mathcal{G}_M$ of size M over the circle with radius α. We will show that $M = O_\tau(n)$ will be enough to approximate the set $\mathcal{B}_\alpha$ in the sense that projecting to the approximating set suffices for the convergence.

Let $\mathcal{B}_{\alpha,\tau_0,\tau_1,\tau_2}' = \{a : p_a(z)/z^n \in \mathcal{C}_{\tau_0,\tau_1,\tau_2}, \forall z \in \mathcal{G}_M\}$ and $\mathcal{B}_{\alpha,\tau_0,\tau_1,\tau_2} = \{a : p_a(z)/z^n \in \mathcal{C}_{\tau_0,\tau_1,\tau_2}, \forall |z| = \alpha\}$. Here $\mathcal{C}_{\tau_0,\tau_1,\tau_2}$ is defined the same as before though we used the subscript

to emphasize the dependency on τ_i's,

$$\mathcal{C}_{\tau_0,\tau_1,\tau_2} = \{z \colon \Re z \geq (1+\tau_0)|\Im z|\} \cap \{z \colon \tau_1 < \Re z < \tau_2\}. \tag{F.1}$$

We will first show that with $M = O_\tau(n)$, we can make $\mathcal{B}'_{\alpha,\tau_1,\tau_2,\tau_3}$ to be sandwiched within to two sets $\mathcal{B}_{\alpha,\tau_0,\tau_1,\tau_2}$ and $\mathcal{B}_{\alpha,\tau'_0,\tau'_1,\tau'_2}$.

Lemma F.1 *For any $\tau_0 > \tau'_0, \tau_1 > \tau'_1, \tau_2 < \tau'_2$, we have that for $M = O_\tau(n)$, there exists $\kappa_0, \kappa_1, \kappa_2$ that polynomially depend on τ_i, τ'_i's such that $\mathcal{B}_{\alpha,\tau_0,\tau_1,\tau_2} \subset \mathcal{B}'_{\alpha,\kappa_0,\kappa_1,\kappa_2} \subset \mathcal{B}_{\alpha,\tau'_0,\tau'_1,\tau'_2}$*

Before proving the lemma, we demonstrate how to use the lemma in our algorithm: We will pick $\tau'_0 = \tau_0/2$, $\tau'_1 = \tau_1/2$ and $\tau'_2 = 2\tau_2$, and find κ_i's guaranteed in the lemma above. Then we use $\mathcal{B}'_{\alpha,\kappa_0,\kappa_1,\kappa_2}$ as the projection set in the algorithm (instead of $\mathcal{B}_{\alpha,\tau_0,\tau_1,\tau_2}$)). First of all, the ground-truth solution Θ is in the set $\mathcal{B}'_{\alpha,\kappa_0,\kappa_1,\kappa_2}$. Moreover, since $\mathcal{B}'_{\alpha,\kappa_0,\kappa_1,\kappa_2} \subset \mathcal{B}_{\alpha,\tau'_0,\tau'_1,\tau'_2}$, we will guarantee that the iterates $\widehat{\Theta}$ will remain in the set $\mathcal{B}_{\alpha,\tau'_0,\tau'_1,\tau'_2}$ and therefore the quasi-convexity of the objective function still holds[9].

Note that the set $\mathcal{B}'_{\alpha,\kappa_0,\kappa_1,\kappa_2}$ contains $O(n)$ linear constraints and therefore we can use linear programming to solve the projection problem. Moreover, since the points on the grid forms a Fourier basis and therefore Fast Fourier transform can be potentially used to speed up the projection. Finally, we will prove Lemma F.1. We need S. Bernstein's inequality for polynomials.

Theorem F.2 (Bernstein's inequality, see, for example, Schaeffer (1941)) *Let $p(z)$ be any polynomial of degree n with complex coefficients. Then,*

$$\sup_{|z|\leq 1} |p'(z)| \leq n \sup_{|z|\leq 1} |p(z)|.$$

We will use the following corollary of Bernstein's inequality.

Corollary F.3 *Let $p(z)$ be any polynomial of degree n with complex coefficients. Then, for $m = 20n$,*

$$\sup_{|z|\leq 1} |p'(z)| \leq 2n \sup_{k\in[m]} |p(e^{2ik\pi/m})|.$$

Proof For simplicity let $\tau = \sup_{k\in[m]} |p(e^{2ik\pi/m})|$, and let $\tau' = \sup_{k\in[m]} |p(e^{2ik\pi/m})|$. If $\tau' \leq 2\tau$ then we are done by Bernstein's inequality. Now let's assume that $\tau' > 2\tau$. Suppose $p(z) = \tau'$. Then there exists k such that $|z - e^{2\pi ik/m}| \leq 4/m$ and $|p(e^{2\pi ik/m})| \leq \tau$. Therefore by Cauchy mean-value theorem we have that there exists ξ that lies between z and $e^{2\pi ik/m}$ such that $p'(\xi) \geq m(\tau' - \tau)/4 \geq 1.1n\tau'$, which contradicts Bernstein's inequality. ∎

Lemma F.4 *Suppose a polynomial of degree n satisfies that $|p(w)| \leq \tau$ for every $w = \alpha e^{2i\pi k/m}$ for some $m \geq 20n$. Then for every z with $|z| = \alpha$ there exists $w = \alpha e^{2i\pi k/m}$ such that $|p(z) - p(w)| \leq O(n\alpha\tau/m)$.*

---

9. with a slightly worse parameter up to constant factor since τ_i's are different from τ_i's up to constant factors

Proof Let $g(z) = p(\alpha z)$ by a polynomial of degree at most n. Therefore we have $g'(z) = \alpha p(z)$. Let $w = \alpha e^{2i\pi k/m}$ such that $|z - w| \le O(\alpha/m)$. Then we have

$$|p(z) - p(w)| = |g(z/\alpha) - p(w/\alpha)| \le \sup_{|x| \le 1} |g'(x)| \cdot \frac{1}{\alpha} |z - w|$$

$$\text{(By Cauchy's mean-value Theorem)}$$

$$\le \sup_{|x| \le 1} |p'(x)| \cdot |z - w| \le n\tau |z - w|. \quad \text{(Corallary F.3)}$$

$$\le O(\alpha n\tau/m).$$

∎

Now we are ready to prove Lemma F.1.

Proof [Proof of Lemma F.1] We choose $\kappa_i = \frac{1}{2}(\tau_i + \tau_i')$. The first inequality is trivial. We prove the second one. Consider a such that $a \in \mathcal{B}_{\alpha, \kappa_0, \kappa_1, \kappa_2}$. We wil show that $a \in \mathcal{B}'_{\alpha, \tau_0', \tau_1', \tau_2'}$. Let $q_a(z) = p(z^{-1})z^n$. By Lemma F.4, for every z with $|z| = 1/\alpha$, we have that there exists $w = \alpha^{-1} e^{2\pi ik/M}$ for some integer k such that $|q_a(z) - q_a(w)| \le O(\tau_2 n/(\alpha M))$. Therefore let $M = cn$ for sufficiently large c (which depends on τ_i's), we have that for every z with $|z| = 1/\alpha$, $q_a(z) \in \mathcal{C}_{\tau_0', \tau_1', \tau_2'}$. This completes the proof. ∎