Learning Fractional-Order Dynamics from a Single Trajectory

Xiaole Zhang[1]*, Ziyi Zhang[2], Zehao Zhao[1], Stephen Tu[1],
Guannan Qu[2], Yorie Nakahira[2], Paul Bogdan[1]
[1]Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California
[2]Department of Electrical and Computer Engineering, Carnegie Mellon University

Abstract

Many real-world processes exhibit long-range dependence, where the current state depends on a slowly decaying trace of past states rather than on the most recent state alone. This paper studies system identification for discrete-time fractional-order linear time-invariant systems from a single observed trajectory of length t, a setting that captures such non-Markovian dynamics through the Grünwald–Letnikov difference operator. Unlike Markovian systems, fractional-order systems couple estimation across the entire history, making both statistical analysis and practical identification more challenging. We propose *Fractional-Order Ordinary-Least-Squares Grid-Search (FO-GS)*, a simple two-stage estimator that exploits the diagonal structure of the fractional-difference operator to decouple the identification problem row-wise. Under the stability assumption, we establish high-probability, non-asymptotic error bounds for estimating both the fractional order and the system matrix in the heterogeneous setting, with both estimation errors scaling as $\mathcal{O}(t^{-1/2})$. Through experiments, we show that *FO-GS* outperforms existing baselines in recovering both the fractional order and the underlying system dynamics.

1 Introduction

Many complex natural and technological systems exhibit long-range dependence, with temporal correlations decaying as a power law rather than exponentially (see Figure 1). Long-range dependence is ubiquitous across diverse domains, including brain [21, 22], muscle and blood glucose activity [8, 37], heart-rate variability [10, 18], gene expression [9], climate and hydrology [16, 23], network traffic [20, 36], finance [3, 5], and even machine-learning systems (e.g., large language models [1, 2]).

Classical Markovian models are inherently ill-suited to capture long-range dependence, as their dynamics depend only on the current state and therefore lack a mechanism to encode persistent historical influence and long-range correlations. In contrast, fractional-order systems provide a natural alternative: the Grünwald–Letnikov difference operator replaces the one-step recursion from an integer-order system with a weighted sum over the full history [14, 17, 25, 26]. While this nonlocal mathematical formulation makes fractional-order models particularly well suited for describing history-dependent dynamics, it also introduces substantial challenges for system identification. The main challenge in identifying fractional-order systems is their intrinsic non-Markovian nature: the current state depends nontrivially on a long history of past states, and the unknown fractional order governs this dependence through the coefficients of the Grünwald–Letnikov difference operator. Consequently, jointly estimating the fractional order and the system matrix leads to a nonlinear inference problem with long-range dependence. Recent works have studied learning and sample-complexity questions for discrete-time fractional-order systems [4, 38–40], but statistical guarantees for learning stochastic fractional-order systems from a single trajectory remain underexplored.

*Corresponding to xiaolezh@usc.edu.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

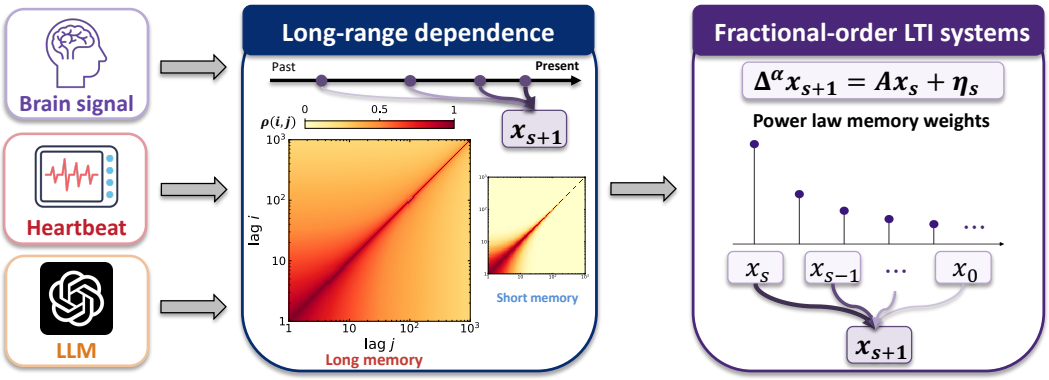

Figure 1: Complex adaptive systems exhibit non-Markovian dynamics, mathematically characterized by long-range dependence. From biological to modern machine learning systems (left panel), the autocorrelation function decays as a power law rather than exponentially (middle panel). Fractional-order operators with their intrinsic power law memory kernel offer compact mathematical strategies to capture this observed long-range dependence dynamics (right panel).

To fill this knowledge gap, we study the fractional-order linear time-invariant (FOLTI) system

$$\Delta^{\boldsymbol{\alpha}} x_{s+1} = Ax_s + \eta_s,$$

where $\Delta^{\boldsymbol{\alpha}}$ denotes the Grünwald–Letnikov difference operator (1), $\boldsymbol{\alpha} \in (0,1)^n$ is the fractional-order vector, $x_s \in \mathbb{R}^n$ is the n-dimensional state at time s, and $\eta_s \sim \mathcal{N}(0, \sigma^2 I_n)$ is the additive Gaussian noise with variance σ^2. We aim to estimate the fractional order $\boldsymbol{\alpha}$ and system matrix $A \in \mathbb{R}^{n \times n}$ given a single observed trajectory $x_0, x_1, \ldots, x_t$. We consider both the heterogeneous setting, in which the coordinates may have different fractional order coefficients in the fractional derivatives, and the homogeneous setting, in which all coordinates share a common fractional order. The central question is whether one can obtain a computationally simple estimator together with non-asymptotic guarantees in this genuinely non-Markovian regime.

This problem has two intertwined challenges: First, the fractional-order difference operator couples each observation to the entire past trajectory, so standard Markovian identification arguments do not apply directly. Second, the dependence on the unknown fractional order $\boldsymbol{\alpha}$ is nonlinear, making joint optimization over $\boldsymbol{\alpha}$ and A difficult even in LTI dynamics. In the single-trajectory setting, these difficulties are compounded by strong temporal dependence and the absence of independent rollouts.

Our contributions. Unlike traditional Ordinary-Least-Squares (OLS)-based algorithms for system identification, we propose the *Fractional-Order OLS Grid-Search (FO-GS)* algorithm, a two-stage identification scheme for a stochastic discrete time FOLTI system from a single trajectory. Using the diagonal structure of the Grünwald–Letnikov difference operator, we identify each row of A together with each coordinate of $\boldsymbol{\alpha}$ separately; along each dimension, *FO-GS* establishes a grid of all potential α_i with a fixed step size ϵ_i and then estimates a_i, the i-th row of system matrix A, conditioned on each candidate α_i. Then, each (α_i, a_i)-pair is evaluated to minimize a profiled loss to select the best candidates. In particular, we show that under the stability assumption, *FO-GS* provably recovers the ground truth parameters α and A with high probability, with both estimation errors scaling as $\mathcal{O}(t^{-1/2})$.

To the best of our knowledge, *FO-GS* offers the first non-asymptotic statistical guarantee for separately identifying $\boldsymbol{\alpha}$ and A for FOLTI systems on a single trajectory. We validate *FO-GS* on both synthetic and real-world datasets. Taken together, *FO-GS* provides a principled framework for learning fractional-order dynamical systems from limited sequential data, thereby substantially broadening the scope of identifiable system classes beyond the standard Markovian setting.

2 Related Work

Our work is rooted in the newly developed branch of fractional-order system identification [13, 39], which borrows inspiration from online learning and identification for ordinary LTI systems.

Learning on a single trajectory for ordinary LTI systems. The identification of the ordinary LTI systems has a long history of research [27, 32, 34, 35, 41–43]. Compared with those works, our algorithm adapts many methodologies and proof techniques and extends the scope to a broader non-Markovian system dynamics by coupling traditional OLS-based identification methods with a grid search scheme on each coordinate of $\boldsymbol{\alpha}$ and row of A, taking advantage of the diagonal structure of the Grünwald–Letnikov difference operator. If $\boldsymbol{\alpha} = \mathbf{1}$ and the fractional-order system simplifies to an ordinary LTI system, *FO-GS* offers a theoretical guarantee comparable to the state-of-the-art guarantee in ordinary LTI systems.

Identification of FOLTI systems. While system identification for ordinary LTI systems is a relatively well understood field, far less attention has been devoted to fractional-order systems [24, 25, 30, 37]. When multiple trajectories can be sampled, some works have been developed in fractional-order system identification. Yaghooti and Sinopoli [38] propose a two-stage identification framework that first estimates the fractional-order parameters from trajectory data generated under a prescribed data-collection procedure and, conditioned on these estimates, reformulates the discrete-time control-affine nonlinear fractional dynamics as a regression problem to infer the unknown system dynamics. Zhang et al. [39, 40] generalize this framework to stochastic settings. Chatterjee and Pequito [4] study single-trajectory identification for fractional-order systems through system truncation and establish a sample-complexity result for the augmented system matrix. By contrast, we study the original single-trajectory identification problem directly, without requiring either data generation or system truncation, and establish the first non-asymptotic guarantees for both the fractional-order parameter $\boldsymbol{\alpha}$ and the system matrix A.

3 Preliminaries and Problem Formulation

3.1 Grünwald–Letnikov Difference Operator

The Grünwald–Letnikov difference operator allows to discretize the fractional-order derivative and represent it as a finite difference of the form as follows:

$$\Delta^{\boldsymbol{\alpha}} x_s := \sum_{j=0}^{s} \Psi(\boldsymbol{\alpha}, j) x_{s-j}, \tag{1}$$

where $x_s \in \mathbb{R}^n$, $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_n]^\top \in (0,1)^n$ represents the fractional order, and $\Psi(\boldsymbol{\alpha}, j) \in \mathbb{R}^{n\times n}$ is a diagonal matrix defined as $\Psi(\boldsymbol{\alpha}, j) := \operatorname{diag}(\psi(\alpha_1, j), \psi(\alpha_2, j), \ldots, \psi(\alpha_n, j))$ with $\psi(\alpha_i, j) := \frac{\Gamma(j-\alpha_i)}{\Gamma(-\alpha_i)\Gamma(j+1)}$ for $i = 1, 2, \ldots, n$. Here $\Gamma(\cdot)$ denotes the gamma function.

3.2 FOLTI System Identification

The state-space representation of the discrete-time FOLTI system reads:

$$\Delta^{\boldsymbol{\alpha}} x_{s+1} = A x_s + \eta_s, \tag{2}$$

where $x_s \in \mathbb{R}^n$ is the state vector, $\eta_s \sim \mathcal{N}(0, \sigma^2 I_n)$ for some $\sigma > 0$ is independent and identically distributed, and $A \in \mathbb{R}^{n\times n}$ is a constant real matrix. Using the Grünwald–Letnikov difference operator (1), we can write the system (2) as follows:

$$x_{s+1} = A x_s - \sum_{j=1}^{s+1} \Psi(\boldsymbol{\alpha}, j) x_{s+1-j} + \eta_s. \tag{3}$$

The solution to the discrete-time FOLTI system (2) is given by [12]:

$$x_s = G_s x_0 + \sum_{j=0}^{s-1} G_{s-1-j} \eta_j,$$

where the matrices G_s are defined recursively by

$$G_s = \begin{cases} I, & s = 0, \\ \sum_{j=0}^{s-1} A_j G_{s-1-j}, & s \geq 1, \end{cases} \qquad A_j = \begin{cases} A + \operatorname{diag}(\alpha_1, \ldots, \alpha_n), & j = 0, \\ -\Psi(\boldsymbol{\alpha}, j+1), & j \geq 1. \end{cases} \tag{4}$$

Problem statement: Given a single observed trajectory $x_0, x_1, \ldots, x_t$, our goal is to identify the fractional order $\boldsymbol{\alpha}$ and the system matrix A, and to establish statistical guarantees for the resulting estimators.

Algorithm 1 Fractional-Order OLS Grid-Search (*FO-GS*)

Require: Trajectory $\{x_s\}_{s=0}^{t}$, row search interval $[\underline{\alpha}_i, \bar{\alpha}_i]$, grid size ϵ_i.
Ensure: Estimates $\hat{\boldsymbol{\alpha}} = [\hat{\alpha}_1, \dots, \hat{\alpha}_n]^\top$ and $\hat{A}$.
Build grid $\mathcal{A}_{\epsilon,i} \subset [\underline{\alpha}_i, \bar{\alpha}_i]$.
Form the data matrix X_t with (9).
for $i = 1, \dots, n$ **do**
  for each $\alpha \in \mathcal{A}_{\epsilon,i}$ **do**
    Compute the fractional-difference row $\Delta^\alpha X_t^{(i)}$ with (5).
    Compute the least-squares row estimator $\hat{a}_i(\alpha)$ with (7).
    Compute the profiled loss $\mathcal{L}^{(i)}(\alpha)$ with (8).
  end for
  Select $\hat{\alpha}_i \leftarrow \arg\min_{\alpha \in \mathcal{A}_{\epsilon,i}} \mathcal{L}^{(i)}(\alpha)$.
  Set $\hat{a}_i \leftarrow \hat{a}_i(\hat{\alpha}_i)$.
end for
Form $\hat{\boldsymbol{\alpha}} \leftarrow [\hat{\alpha}_1, \dots, \hat{\alpha}_n]^\top$, and $\hat{A} \leftarrow [\hat{a}_1^\top, \dots, \hat{a}_n^\top]^\top$.
return $(\hat{\boldsymbol{\alpha}}, \hat{A})$.

4 Main Results

In this section, we introduce the algorithm for identifying the system parameters $\boldsymbol{\alpha}$ and A in Section 4.1, and present the sample complexity results for the resulting estimators in Section 4.2.

4.1 FO-GS

We introduce the algorithm for identifying the system parameters $(\boldsymbol{\alpha}, A)$ of the FOLTI system (2) from a single observed trajectory. The main idea is to isolate each coordinate of $\boldsymbol{\alpha}$ under the diagonal structure of the Grünwald–Letnikov difference operator and, for each candidate, solve an OLS problem to estimate A. To handle the general fractional-order setting, Algorithm 1 performs a grid search over each component of $\boldsymbol{\alpha}$, and then solves a row-wise OLS problem. Specifically, for each $i \in \{1, \dots, n\}$, define

$$\Delta^{\alpha_i} X_t^{(i)} := \left[\Delta^{\alpha_i} x_1^{(i)}, \dots, \Delta^{\alpha_i} x_t^{(i)}\right] \in \mathbb{R}^{1\times t}, \tag{5}$$

where $\Delta^{\alpha_i} x_{s+1}^{(i)} = \sum_{j=0}^{s+1} \psi(\alpha_i, j) x_{s+1-j}^{(i)}$. Denote $A = [a_1^\top, \dots, a_n^\top]^\top$, where $a_i \in \mathbb{R}^{1\times n}$ represents the i-th row of A. We further define the loss function $\mathcal{L}$ as follows:

$$\mathcal{L}(\boldsymbol{\alpha}, A) = \sum_{s=0}^{t-1} \left\|\Delta^{\boldsymbol{\alpha}} x_{s+1} - A x_s\right\|_2^2 = \sum_{i=1}^{n} \sum_{s=0}^{t-1} \left(\Delta^{\alpha_i} x_{s+1}^{(i)} - a_i x_s\right)^2. \tag{6}$$

For each row i, let $\mathcal{A}_i = [\underline{\alpha}_i, \bar{\alpha}_i] \subset (0, 1]$ be a compact search interval. Fixing a step size $\epsilon_i > 0$, we construct a uniform grid $\mathcal{A}_{\epsilon,i} = \left\{\alpha_{i,k} := \underline{\alpha}_i + k\epsilon_i \,|\, k \in \{1, \dots, M_i\}\right\} \subset \mathcal{A}_i$. For each candidate $\alpha_{i,k}$, we solve the row-wise OLS problem and obtain:

$$\hat{a}_i(\alpha_{i,k}) := \arg\min_{a_i} \mathcal{L}^{(i)}(\alpha_{i,k}, a_i) = (\Delta^{\alpha_{i,k}} X_t^{(i)}) X_t^\top (X_t X_t^\top)^{-1}. \tag{7}$$

We then minimize the corresponding profiled loss and obtain the estimated α_i:

$$\hat{\alpha}_i = \arg\min_{\alpha \in \mathcal{A}_{\epsilon,i}} \mathcal{L}^{(i)}(\alpha) = \arg\min_{\alpha \in \mathcal{A}_{\epsilon,i}} \sum_{s=0}^{t-1} \left(\Delta^{\alpha} x_{s+1}^{(i)} - \hat{a}_i(\alpha) x_s\right)^2. \tag{8}$$

Finally, by stacking the row estimators $\hat{a}_i(\alpha_i)$, we obtain the system matrix estimator

$$\hat{A}(\boldsymbol{\alpha}) = (\Delta^{\boldsymbol{\alpha}} X_t) X_t^\top (X_t X_t^\top)^{-1},$$

where

$$\Delta^{\boldsymbol{\alpha}} X_t = [\Delta^{\boldsymbol{\alpha}} x_1, \dots, \Delta^{\boldsymbol{\alpha}} x_t],\ X_t = [x_0, \dots, x_{t-1}]. \tag{9}$$

The commensurate setting is a direct specialization of the above procedure. When all coordinates share a common fractional order, i.e., $\boldsymbol{\alpha} = \alpha \mathbf{1}$, the row-wise searches collapse to a single one-dimensional search over α.

4.2 Theoretical Guarantees

In this section, we present complexity guarantees for the proposed algorithm. We begin by introducing the assumptions required for our main results. For the rest of the paper, we use $\boldsymbol{\alpha}_\star = [\alpha_{1,\star}, ...\alpha_{n,\star}]^\top$ to denote the true fractional-order vector and $A_\star$ to denote the true system matrix. We assume $x_0 = 0$ for simplicity. We first introduce a stability assumption that is standard in the LTI system identification literature.

Assumption 1. The system parameters $(A_\star, \boldsymbol{\alpha}_\star)$ are stable in the sense that for all $|z| \leq 1$,

$$\det\Big(\mathrm{diag}\big((1-z)^{\alpha_{1,\star}}, \ldots, (1-z)^{\alpha_{n,\star}}\big) - zA_\star\Big) \neq 0.$$

Assumption 1 can be interpreted as the fractional-order version of the stability assumption common in control literature [19, 28, 29, 31, 33]. It is slightly stronger than minimal stability, since imposing the condition at $z = 1$ implies that $A_\star$ is invertible. This property is used in our analysis to bound the matrices G_s in (4). It ensures that the state x_t does not blow-up with time. If the system is unstable, then any error at the early time period would be exponentially amplified by the unstable system dynamics in a phenomenon known as *stochastic coupling* [41]. We leave this as a future direction of this paper.

We are now ready to introduce the main theorems. First, we introduce the error bound on $\boldsymbol{\alpha}_\star$:

Theorem 1. *Suppose Assumption 1 holds and the population separation gap γ in* (64) *is positive. Let $\epsilon_{\max} := \max_{1\leq i\leq n} \epsilon_i$. Fix $\delta \in (0, 1/2)$. If $t \gtrsim \frac{1}{\min_i \underline{\alpha}_i^4}\left(n + \log\frac{\sum_{i=1}^n M_i}{\delta}\right)$, and the excitation* (65) *and localization* (66) *conditions are satisfied, then with probability at least $1-\delta$,*

$$\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_\star\|_\infty^2 \lesssim \mathrm{poly}\left(n, \frac{1}{\delta}\right)\left[\epsilon_{\max}^2 + \frac{1}{t}\sum_{i=1}^n \log\frac{M_i n}{\delta} + \frac{1}{t}\log\frac{n}{\delta}\right], \tag{10}$$

where $\lesssim$ hides system-dependent constants independent of t, n, δ, $\epsilon_{\max}$, and M_i. Consequently, if $\epsilon_{\max} = \mathcal{O}\left(t^{-1/2}\right)$, then $\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_\star\|_\infty = \mathcal{O}\left(t^{-1/2}\right)$.

We defer the proof of Theorem 1 to Appendix B. The condition for t requires the trajectory to be long enough for both the global and local lower isometry bounds to hold uniformly over the finite search grid. In addition to the excitation and localization conditions, the trajectory length scales as $t \gtrsim \frac{1}{\min_i \underline{\alpha}_i^4}\left(n + \log\frac{\sum_{i=1}^n M_i}{\delta}\right)$, thus smaller fractional orders require longer trajectories, reflecting the stronger long-memory dependence in this regime. Theorem 1 makes explicit the tradeoff between statistical error and grid discretization error. Up to logarithmic factors in the grid size M_i, the statistical term decays as $t^{-1/2}$, whereas the discretization term decays as $\epsilon_{\max}$. Accordingly, choosing $\epsilon_{\max} = \mathcal{O}(t^{-1/2})$ makes the two contributions comparable. For a uniform grid over a bounded interval, this corresponds to $M_i = \mathcal{O}(t^{1/2})$, which is sufficient to match the statistical precision. To the best of our knowledge, this is the first high-probability single trajectory error bound for estimation of the fractional order $\boldsymbol{\alpha}_\star$ in the FOLTI setting. Our result is complementary to prior work based on truncated, bisection-like identification schemes [4], and to more recent analyses developed under different data-generation frameworks [38–40]. We then discuss the error complexity of estimating $A_\star$ in the following theorem:

Theorem 2. *Under the same condition as in Theorem 1, fix $\delta \in (0, \frac{1}{2})$ and consider the system* (3)*. Let $\Gamma_s = \sum_{m=0}^{s-1} G_m G_m^\top$ and $\Xi_t(\delta, k) := n\log\frac{9n}{\delta} + \log\det\left(\Gamma_t\Gamma_k^{-1}\right)$. Then there exist universal constants $c, C > 0$ such that, for any integer k satisfying $\frac{t}{k} \geq c\Xi_t(\delta, k)$, the following holds with probability at least $1-\delta$:*

$$\left\|\hat{A}(\hat{\boldsymbol{\alpha}}) - A_\star\right\|_{\mathrm{op}} \leq C\left(\sqrt{\frac{\Xi_t(\delta,k)}{t\,\lambda_{\min}(\Gamma_k)}} + S_1\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_\star\|_\infty\sqrt{\frac{n\tilde{C}_G^2}{\delta\lambda_{\min}(\Gamma_{\lfloor k/2\rfloor})}}\right),$$

where $\tilde{C}_G$ is a constant depending on $(A_\star, \boldsymbol{\alpha}_\star)$, and S_1 is a constant depending on the grid search interval.

We defer the proof of Theorem 2 to Appendix C. Theorem 2 shows that the estimation error for $A_\star$ decomposes into two parts. The first term is a standard oracle least-squares error, which is the error

that would arise even if the true fractional-order vector were known. The second term quantifies the propagation of the fractional-order estimation error into the estimation of $A_\star$. Consequently, when $\hat{\boldsymbol{\alpha}}$ is obtained from Theorem 1, the bound for $A_\star$ inherits the same statistical-discretization tradeoff as the bound for $\boldsymbol{\alpha}_\star$. In particular, if $\epsilon_{\max} = \mathcal{O}(t^{-1/2})$, then the propagated term is of order $t^{-1/2}$, while the oracle term is also of order $t^{-1/2}$. Therefore, the overall estimation error for $A_\star$ achieves the $t^{-1/2}$ rate. Chatterjee and Pequito [4] analyzes a truncated system identification scheme, but does not explicitly quantify the estimation error for the original system matrix $A_\star$ or how fractional-order estimation error propagates to the estimation of $A_\star$. We also note that in the case when $\boldsymbol{\alpha}_\star = \mathbf{1}$ is known a priori, there is no need to estimate the fractional-order parameter. So Theorem 2 is equivalent to the current state-of-the-art bound for estimating $A_\star$ for ordinary LTI systems [34].

5 Proof Outline

The proof in this paper is split into two steps, bounding the estimation error for $\boldsymbol{\alpha}_\star$ and then for $A_\star$. In this section, we provide an outline of the proof for each, and defer the details to the appendix.

5.1 Proof of Theorem 1: bounding the error of $\boldsymbol{\alpha}_\star$

The estimation of $\boldsymbol{\alpha}_\star$ can be analyzed in two steps: controlling the in-sample error and establishing a lower isometry bound. The first step gives an absorbable upper bound on the in-sample prediction error, while the second localizes the estimator and converts the same error into a quadratic lower bound on the fractional-order estimation error.

Controlling the in-sample error. For each row i, define $b_s^{(i)}(\alpha_i) := (\Delta^{\alpha_i} - \Delta^{\alpha_{i,\star}})\, x_{s+1}^{(i)}$ and $y_s^{(i)}(\alpha_i) := b_s^{(i)}(\alpha_i) + 2\eta_s^{(i)}$, and let α_i° denote the grid point closest to $\alpha_{i,\star}$. We study the unnormalized in-sample error

$$\mathcal{E}_t := \sum_{s=0}^{t-1} \left\| \left(\Delta^{\hat{\boldsymbol{\alpha}}} - \Delta^{\boldsymbol{\alpha}_\star}\right) x_{s+1} - (\hat{A} - A_\star) x_s \right\|_2^2 .$$

By the optimality of the row-wise estimator $L^{(i)}(\hat{\alpha}_i, \hat{a}_i) \le L^{(i)}(\alpha_i^\circ, \hat{a}_i(\alpha_i^\circ)) \le L^{(i)}(\alpha_i^\circ, a_{i,\star})$ and a quadratic maximization over the system matrix perturbation, we obtain

$$\mathcal{E}_t \le \sum_{i=1}^{n} \max_{\alpha_i \in \mathcal{A}_{\epsilon,i}} \left\{ \underbrace{-4\sum_{s=0}^{t-1} \eta_s^{(i)} b_s^{(i)}(\alpha_i) - \sum_{s=0}^{t-1} |b_s^{(i)}(\alpha_i)|^2}_{U_{t,i}(\alpha_i)} + \underbrace{\left\| \left(\sum_{s=0}^{t-1} y_s^{(i)}(\alpha_i) x_s^\top \right) (X_t X_t^\top)^{-\frac{1}{2}} \right\|_2^2}_{V_{t,i}(\alpha_i)} \right\} + \underbrace{\sum_{i=1}^{n} \left[4\sum_{s=0}^{t-1} \eta_s^{(i)} b_s^{(i)}(\alpha_i^\circ) + 2\sum_{s=0}^{t-1} |b_s^{(i)}(\alpha_i^\circ)|^2 \right]}_{\Gamma_t^{\mathrm{grid}}} . \tag{11}$$

Exact grid case. If the true parameter lies exactly on the search grid, i.e., $\alpha_{i,\star} \in \mathcal{A}_{\epsilon,i}$ for every i, then we may take $\alpha_i^\circ = \alpha_{i,\star}$. In this case $b_s^{(i)}(\alpha_i^\circ) = 0$ for every i and s, and hence $\Gamma_t^{\mathrm{grid}} = 0$. Thus, the continuous empirical risk minimizer offset inequality is recovered as a special case.

Row-wise complexity decomposition. Since the fractional-order operator is diagonal across state coordinates and the estimator searches for each α_i separately, the offset complexity decomposes as

$$\sum_{i=1}^{n} \max_{\alpha_i \in \mathcal{A}_{\epsilon,i}} \left\{ U_{t,i}(\alpha_i) + V_{t,i}(\alpha_i) \right\},$$

rather than requiring a supremum over the full Cartesian grid $\mathcal{A}_{\epsilon,1} \times \cdots \times \mathcal{A}_{\epsilon,n}$. This row-wise decomposition is important for obtaining the sharp complexity dependence of the grid-search estimator.

The offset martingale complexity argument [44] preserves the negative quadratic term in $U_{t,i}$, yielding a term proportional to $\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_\star\|_\infty^2$ with a tunable coefficient, which can later be absorbed by the

lower isometry bound. The additional term Γ_t^{grid} accounts for the finite grid approximation and vanishes when the true fractional orders lie on the search grid. Bounding $U_{t,i}$, $V_{t,i}$, and Γ_t^{grid} separately then yields the following high probability control of $\mathcal{E}_t$.

Lemma 1. *For any fixed $\tau \in (0,1]$ and $\rho > 0$, if $t/k \geq c\Xi_t(\delta, k)$, then with probability at least $1-\delta$,*

$$\mathcal{E}_t \leq (\tau+\rho)S_1^2 \left\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_\star\right\|_\infty^2 \operatorname{tr}(X_t X_t^\top) + 3S_1^2\epsilon_{\max}^2 \operatorname{tr}(X_t X_t^\top) + \frac{4(1+\rho^{-1})C^2 \operatorname{tr}(X_t X_t^\top)}{t\lambda_{\min}(\Gamma_k)}\Xi_t(\delta,k) + \frac{8\sigma^2}{\tau}\sum_i \log\frac{3M_i n}{\delta} + 8\sigma^2 \log\frac{3}{\delta}.$$

The coefficient $\tau+\rho$ is tunable and can be absorbed by lower isometry.

We defer the proof of Lemma 1 to Appendix B.1.

Lower isometry. For each row i, define the unprofiled and profiled noiseless errors $\mathcal{Q}_{t,i}(\alpha_i, v) := \sum_{s=0}^{t-1}\left|b_s^{(i)}(\alpha_i) - vx_s\right|^2$ and $\underline{\mathcal{Q}}_{t,i}(\alpha_i) := \inf_{v\in\mathbb{R}^{1\times n}} \mathcal{Q}_{t,i}(\alpha_i, v)$, and the row-wise population risk $R_t^{(i)}(\alpha_i) := \frac{1}{t}\inf_{a_i\in\mathbb{R}^{1\times n}} \sum_{s=0}^{t-1}\mathbb{E}\left[\left(\Delta^{\alpha_i}x_{s+1}^{(i)} - a_i x_s\right)^2\right]$. A global profiled lower isometry bound shows that, uniformly over grid points outside the local neighborhood of $\alpha_{i,\star}$,

$$\underline{\mathcal{Q}}_{t,i}(\alpha_i) \geq \frac{t}{2}\left[R_t^{(i)}(\alpha_i) - R_t^{(i)}(\alpha_{i,\star})\right].$$

Together with the upper bound on $\mathcal{E}_t$ in Lemma 1, this excludes grid points outside the local neighborhood and localizes $\hat{\alpha}_i$ to a set $\mathcal{G}_i$ in (34) around $\alpha_{i,\star}$, where the first-order expansion of $b_s^{(i)}(\alpha_i)$ has controlled remainder and the corresponding population derivative curvature $\mu_{t,i}$ is nondegenerate. Within $\mathcal{G}_i$, this yields the quadratic lower isometry bound $\underline{\mathcal{Q}}_{t,i}(\alpha_i) \geq \frac{t\mu_{t,i}}{8}|\alpha_i - \alpha_{i,\star}|^2$. The following lemma formalizes this local lower isometry property.

Lemma 2. *Under Assumption 1, if $t \gtrsim \frac{1}{\min_i \underline{\alpha}_i^4}\left(n + \log\frac{\sum_{i=1}^n M_i}{\delta}\right)$, then with probability at least $1-\delta$, simultaneously for all $i\in[n]$ and all $\alpha_i \in \mathcal{G}_i$, $\underline{\mathcal{Q}}_{t,i}(\alpha_i) \geq \frac{t\mu_{t,i}}{8}|\alpha_i - \alpha_{i,\star}|^2$. Consequently, on any event for which $\hat{\alpha}_i \in \mathcal{G}_i$ for every $i\in[n]$, $\mathcal{Q}_{t,i}(\hat{\alpha}_i, \hat{a}_i - a_{i,\star}) \geq \frac{t\mu_{t,i}}{8}|\hat{\alpha}_i - \alpha_{i,\star}|^2$.*

We defer the proof of Lemma 2 to Appendix B.2. Combining the lower isometry bound in Lemma 2 with the in-sample upper bound in Lemma 1 and choosing the tunable coefficient sufficiently small gives

$$\left\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_\star\right\|_\infty^2 \lesssim \epsilon_{\max}^2 + \mathcal{O}(t^{-1}).$$

Hence, choosing $\epsilon_{\max} = O(t^{-1/2})$ yields the $t^{-1/2}$ rate in Theorem 1.

5.2 Proof of Theorem 2: bounding the error of $A_\star$

We decompose the identification error of A given $\boldsymbol{\alpha}$. For any $\boldsymbol{\alpha}$, we have

$$\hat{A}(\boldsymbol{\alpha}) - A_* = (B_t(\boldsymbol{\alpha}) + W_t)X_t^\top(X_t X_t^\top)^{-1},$$

where $W_t := [\eta_0, \ldots, \eta_{t-1}] \in \mathbb{R}^{n\times t}$, $B_t(\boldsymbol{\alpha}) := [\, b_0(\boldsymbol{\alpha}), \ldots, b_{t-1}(\boldsymbol{\alpha})\,] \in \mathbb{R}^{n\times t}$, and $b_s(\boldsymbol{\alpha}) = \sum_{j=0}^s \big(\Psi(\boldsymbol{\alpha}, j) - \Psi(\boldsymbol{\alpha}_\star, j)\big)\, x_{s-j} = [b_s^{(1)}(\boldsymbol{\alpha}), \ldots, b_s^{(n)}(\boldsymbol{\alpha})]^\top$. Thus

$$\left\|\hat{A}(\boldsymbol{\alpha}) - A_\star\right\|_{\mathrm{op}} \leq \underbrace{\left\|W_t X_t^\top(X_t X_t^\top)^{-1}\right\|_{\mathrm{op}}}_{\text{noise term}} + \underbrace{\left\|B_t(\boldsymbol{\alpha})X_t^\top(X_t X_t^\top)^{-1}\right\|_{\mathrm{op}}}_{\text{bias term}}. \tag{12}$$

We therefore decompose the identification error in (12) into noise and bias terms, which we bound separately in Lemmas 3 and 4. To bound the noise term, we adapt the technique from Sarkar and Rakhlin [32], Simchowitz et al. [34] and get the following lemma:

Lemma 3. *Fix $\delta \in (0, \frac{1}{2})$ and consider the system (3) under Assumption 1. Then there exist universal constants $c, C > 0$ such that*

$$\mathbb{P}\left[\left\|W_t X_t^\top(X_t X_t^\top)^{-1}\right\|_{\mathrm{op}} > \frac{C}{\sqrt{t\,\lambda_{\min}(\Gamma_k)}}\sqrt{n\log\frac{n}{\delta} + \log\det(\Gamma_t\Gamma_k^{-1})}\right] \leq \delta,$$

for any k such that $\frac{t}{k} \geq c\big(n\log(n/\delta) + \log\det(\Gamma_t\Gamma_k^{-1})\big)$ holds.

We defer the proof of Lemma 3 to Appendix C.1. We now offer the bound for the bias term in (12):

Lemma 4. *With probability at least* $1 - 2\delta$*, the following holds*

$$\left\|B_t(\boldsymbol{\alpha})X_t^\top (X_t X_t^\top)^{-1}\right\|_{\mathrm{op}} \le S_1 \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_\star\|_\infty \sqrt{\frac{320 n \tilde{C}_G^2}{9\delta p^2 \lambda_{\min}(\Gamma_{\lfloor k/2 \rfloor})}},$$

where $\alpha_{\min} = \min_{1 \le i \le n} \alpha_{i,\star}$*, and* $p = \frac{3}{20}$.

We defer the proof of Lemma 4 to Appendix C.2.

6 Experiments

We evaluate the proposed method *FO-GS* through two sets of experiments. First, we use synthetic data to validate the theoretical guarantee and compare *FO-GS* with existing fractional-order identification algorithms [4, 7]. *FO-WT* estimates the fractional order $\boldsymbol{\alpha}_\star$ using a wavelet-based technique [7] and then applies OLS to estimate the system matrix $A_\star$. *FO-BS* estimates $\boldsymbol{\alpha}_\star$ via binary search and then identifies $A_\star$ by applying OLS to an augmented state representation obtained through system truncation [4]. Second, as fractional-order systems have been used in analyzing electroencephalogram (EEG) data, we further demonstrate that *FO-GS* excels in minimizing the one-step normalized mean squared error (NMSE) on both training and testing datasets in comparison to both existing methods (*FO-BS* and *FO-WT*). Both the synthetic and real-world experiments demonstrate that *FO-GS* outperforms the baselines. We provide additional experimental details in Appendix A.

6.1 Performance Evaluation on Synthetic Data

In this section, we compare *FO-GS* with two existing baselines on synthetic data. The trajectories are generated according to the fractional-order dynamics in (2). The ground truth parameters $\boldsymbol{\alpha}_\star$ and $A_\star$ are randomly sampled, and the reported mean squared error (MSE) is averaged over five randomly generated system instances, with 20 independent rollouts for each instance.

Varying trajectory horizons. We evaluate the MSE of the fractional order $\boldsymbol{\alpha}_\star$ and the system matrix $A_\star$ as functions of the trajectory horizon t by fixing the noise scale σ. As shown in Figure 2, *FO-GS* outperforms baselines in estimating both $\boldsymbol{\alpha}_\star$ and $A_\star$ across all trajectory lengths. The improvement is particularly pronounced for shorter horizons, where accurate identification is most challenging. A plausible reason is that our method directly fits the original fractional-order model and exploits its structural decomposition, whereas the baselines rely either on a wavelet-based proxy for estimating the fractional order $\boldsymbol{\alpha}_\star$ or on a truncated lifted state approximation for identifying the system matrix $A_\star$. These additional approximation steps can introduce non-negligible finite-sample error, especially when the available trajectory is short. We also observe that the estimation error generally decreases as t increases, which is consistent with the theoretical predictions in Theorems 1 and 2.

Varying noise scales. We examines how the MSE of the fractional order $\boldsymbol{\alpha}_\star$ and the system matrix $A_\star$ varies with the noise scale σ given the same horizon t. Figure 2 shows that *FO-GS* consistently outperforms the two baselines across all noise levels. As expected, the MSE of all methods increases as the noise level grows, but the *FO-GS* remains the most robust, likely because it estimates $\boldsymbol{\alpha}_\star$ and $A_\star$ directly from the original fractional-order model, whereas the baselines incur additional approximation error through wavelet-based estimation or system truncation.

Varying grid sizes. We study how the number of grids M affects the performance of *FO-GS* (Figure 2). Consistent with Theorems 1 and 2, the MSE for both the fractional order $\boldsymbol{\alpha}$ and the system matrix A decreases as the number of grids increases. Notably, *FO-GS* outperforms both *FO-BS* and *FO-WT* without requiring a large number of grids: it surpasses the baselines in estimating $\boldsymbol{\alpha}_\star$ with roughly ten grids and in estimating $A_\star$ with roughly five grids. This demonstrates that *FO-GS* is not only accurate but also computationally efficient.

Small $\boldsymbol{\alpha}_\star$ regime. We further examine the small fractional-order regime by uniformly sampling $\boldsymbol{\alpha}_\star$ from the interval $[0.01, 0.2]$ for systems with $n = 10$ and $n = 20$, and extending the trajectory horizon up to $t = 25{,}600$. As shown in Table 1, over the shorter horizons $t \in \{100, 200, 300, 400\}$, the empirical convergence is slower than the predicted $t^{-1/2}$ rate, indicating stronger finite-sample effects when the fractional orders are small. As the trajectory length increases, however, the fitted

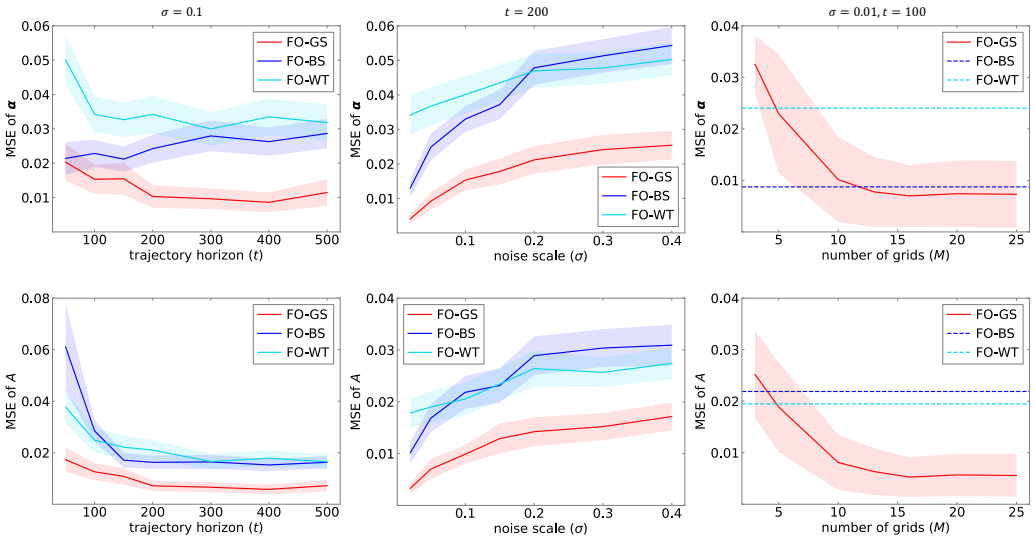

Figure 2: Comparison of the MSE on synthetic FOLTI system identification. Shaded regions indicate 95% confidence intervals (CI).

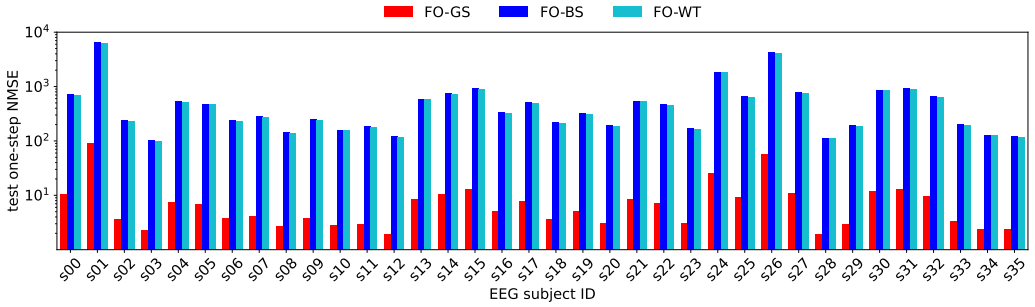

Figure 3: Subject-level average test one-step NMSE on the EEG mental-arithmetic dataset. Each point corresponds to one subject and is obtained by averaging the window-level test one-step NMSE over all non-overlapping windows from that subject.

rates become progressively faster. Specifically, Table 2 shows that over the full extended horizon, the rate scales approximately as $t^{-0.36}$ for $n = 10$ and $t^{-0.38}$ for $n = 20$, while fitting only the larger horizon regime $t \geq 3200$ yields rates of approximately $t^{-0.43}$ and $t^{-0.42}$, respectively. These results show a clear trend toward the $t^{-1/2}$ rate predicted by Theorem 1 as the trajectory becomes longer. We emphasize that the sample-size requirement in Theorem 1, including its explicit $\alpha_{\min}^{-4}$ dependence, is a sufficient condition for entering the fast rate regime rather than a necessary or optimal threshold. Thus, the observed finite-sample behavior may be better than what is implied by the conservative sufficient condition.

6.2 Performance Evaluation on Real-World Data

We evaluate *FO-GS* on an EEG mental-arithmetic dataset [45] comprising artifact-free recordings from 36 subjects, sampled at 500 Hz with a Neurocom 23-channel system and 19 electrodes placed obeying the International $10/20$ scheme. We use the first minute of the serial-subtraction task, treating each subject's recording as a 19-dimensional time series. The data are segmented into non-overlapping windows of length $W = 150$, with 70% of samples used for training and 30% for testing. We compare *FO-GS* against *FO-BS* and *FO-WT* using one-step NMSE. Figure 3 reports subject-level average test NMSE. Training errors are similar across methods as shown in Figure 4, but test performance differs markedly: *FO-GS* consistently achieves the lowest NMSE for all subjects, while *FO-BS* and *FO-WT* incur higher errors. This improvement stems from the fractional-order

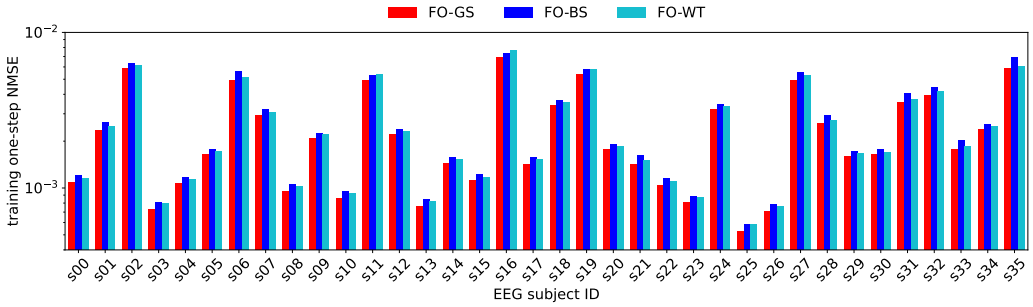

Figure 4: Subject-level average training one-step NMSE on the EEG mental-arithmetic dataset. Each point corresponds to one subject and is obtained by averaging the window-level training one-step NMSE over all non-overlapping windows from that subject.

Table 1: Short horizon convergence rates from log–log fits. We report a 95% bootstrap confidence interval for each MSE exponent.

$\boldsymbol{\alpha}_\star$ regime	MSE exponent (95% CI)	R^2	RMSE rate
Small $\boldsymbol{\alpha}_\star$ ($n = 10$)	$-0.505^{+0.096}_{-0.091}$	0.998	$t^{-0.253}$
Small $\boldsymbol{\alpha}_\star$ ($n = 20$)	$-0.571^{+0.101}_{-0.093}$	0.999	$t^{-0.286}$
Large $\boldsymbol{\alpha}_\star$ ($n = 2$)	$-1.129^{+0.348}_{-0.362}$	0.973	$t^{-0.564}$

Table 2: Long horizon convergence rates from log–log fits. Results are reported for the small $\boldsymbol{\alpha}_\star$ regime. "All horizons" refers to fits using all time points.

Dimension	Fitting window	MSE exponent (95% CI)	RMSE rate
$n = 10$	All horizons	$-0.723^{+0.023}_{-0.024}$	$t^{-0.362}$
	$t \geq 3200$	$-0.868^{+0.089}_{-0.085}$	$t^{-0.434}$
$n = 20$	All horizons	$-0.758^{+0.013}_{-0.013}$	$t^{-0.379}$
	$t \geq 3200$	$-0.848^{+0.056}_{-0.058}$	$t^{-0.424}$

identification in *FO-GS*, where accurate estimation of $\boldsymbol{\alpha}$ yields better history weighting and prediction. In contrast, *FO-BS* uses a fixed finite-memory approximation and *FO-WT* estimates the order via a separate wavelet-based step. Overall, *FO-GS* delivers superior predictive performance.

Learned-order interpretation. We also examine the learned fractional orders on the EEG dataset and find clear evidence of non-integer, channel-dependent memory. For *FO-GS*, the median window-averaged order is 0.855; 74.6% of windows have a cross-channel order range greater than 0.5, and 72.8% contain at least one channel with $\alpha < 0.1$. These results suggest substantial heterogeneity in long-term memory across EEG channels.

7 Conclusion

We study the identification of FOLTI systems from a single observed trajectory and propose *FO-GS*, a simple two-stage estimator that exploits the diagonal structure of the Grünwald–Letnikov difference operator to decouple the estimation of the fractional order $\boldsymbol{\alpha}_\star$ and the system matrix $A_\star$ row-wise. Under the stability assumption, we show that *FO-GS* admits high-probability non-asymptotic error guarantees for recovering both $\boldsymbol{\alpha}_\star$ and $A_\star$. *FO-GS* outperforms existing baselines on both synthetic and EEG data. These results indicate that direct single-trajectory identification of FOLTI systems is both statistically analyzable and practically effective despite the non-Markovian system dynamics.

Limitations and future work. The convergence rate for $\boldsymbol{\alpha}_\star$ may be suboptimal. Our finite-sample condition is sufficient for achieving the fast $t^{-1/2}$ rate, but grid search introduces an additional $\log t$ factor in the bound. Our EEG evaluation includes only two FOLTI-based baselines, and appropriately tuned classical long-memory *ARFIMA* models [11] can outperform *FO-GS*. Future work should relax the stability assumption and sharpen the $\boldsymbol{\alpha}_\star$ dependence of the convergence bound.

Acknowledgments and Disclosure of Funding

Stephen Tu acknowledges support from the National Science Foundation (NSF) under NSF CPS 2434460. Guannan Qu was supported by NSF Grants 2339112, 2512805, AFOSR Grant FA9550261B281, Jane Street, Pennsylvania Infrastructure Technology Alliance, CMU Manufacturing Futures Institute. Yorie Nakahira was supported by NSF Grant No. ECCS-2442948 (CAREER) and Japan Science and Technology Agency (JST): Grant No. JPMJPR2516 (PRESTO). Paul Bogdan was supported by the National Science Foundation under NSF Award 2243104 under the Center for Complex Particle Systems (COMPASS), CAREER Award CPS-1453860, NSF Mid-Career Advancement Award BCS-2527046, U.S. Army Research Office (ARO) under Grant No. W911NF-23-1-0111, Defense Advanced Research Projects Agency (DARPA) Young Faculty Award and DARPA Director Fellowship Award under Grant Number N66001-17-1-4044, DARPA MATHBAC grant HR0011263E153, Intel Faculty Awards, and a Northrop Grumman grant. P.B. is also grateful to the National Institutes of Health (NIH) for Grants R01 AG079957 and RF1 AG082201. The views, opinions, and/or findings in this article are those of the authors and should not be interpreted as official views or policies of the Department of Defense, the National Institutes of Health, or the National Science Foundation.

References

[1] Ibrahim Alabdulmohsin and Andreas Peter Steiner. A tale of two structures: Do LLMs capture the fractal complexity of language? In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=p2smPMRQae.

[2] Ibrahim Alabdulmohsin, Vinh Q. Tran, and Mostafa Dehghani. Fractal patterns may illuminate the success of next-token prediction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=clAFYReaYE.

[3] Torben G. Andersen, Tim Bollerslev, Francis X. Diebold, and Paul Labys. Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–625, 2003. doi: https://doi.org/10.1111/1468-0262.00418. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00418.

[4] Sarthak Chatterjee and Sérgio Pequito. On learning discrete-time fractional-order dynamical systems. In *2022 American Control Conference (ACC)*, pages 4335–4340. IEEE, 2022.

[5] Zhuanxin Ding, Clive W.J. Granger, and Robert F. Engle. A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, 1(1):83–106, 1993. ISSN 0927-5398. doi: https://doi.org/10.1016/0927-5398(93)90006-D. URL https://www.sciencedirect.com/science/article/pii/092753989390006D.

[6] Philippe Flajolet and Andrew M. Odlyzko. Singularity analysis of generating functions. *SIAM Journal on Discrete Mathematics*, 3(2):216–240, 1990. doi: 10.1137/0403019. URL https://epubs.siam.org/doi/10.1137/0403019.

[7] Patrick Flandrin. Wavelet analysis and synthesis of fractional brownian motion. *IEEE Transactions on information theory*, 38(2):910–917, 2002.

[8] Mahboobeh Ghorbani and Paul Bogdan. A cyber-physical system approach to artificial pancreas design. In *2013 International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS)*, pages 1–10, 2013. doi: 10.1109/CODES-ISSS.2013.6659004.

[9] Mahboobeh Ghorbani, Edmond A. Jonckheere, and Paul Bogdan. Gene expression is not random: Scaling, long-range cross-dependence, and fractal characteristics of gene regulatory networks. *Frontiers in Physiology*, Volume 9 - 2018, 2018. ISSN 1664-042X. doi: 10.3389/fphys.2018.01446. URL https://www.frontiersin.org/journals/physiology/articles/10.3389/fphys.2018.01446.

[10] Ary L. Goldberger, Luis A. N. Amaral, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, C.-K. Peng, and H. Eugene Stanley. Fractal dynamics in physiology: Alterations with disease and aging. *Proceedings of the National Academy of Sciences*, 99(suppl_1):2466–2472, 2002. doi: 10.1073/pnas.012579499. URL https://www.pnas.org/doi/abs/10.1073/pnas.012579499.

[11] C. W. J. Granger and Roselyne Joyeux. An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis*, 1(1):15–29, 1980. doi: https://doi.org/10.1111/j.1467-9892.1980.tb00297.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9892.1980.tb00297.x.

[12] Saïd Guermah, Saïd Djennoune, and Maâmar Bettayeb. Discrete-time fractional-order systems: Modeling and stability issues. *Advances in Discrete Time Systems*, pages 183–212, 2012.

[13] Gaurav Gupta, Sérgio Pequito, and Paul Bogdan. Learning latent fractional dynamics with unknown unknowns. In *2019 American Control Conference (ACC)*, pages 217–222. IEEE, 2019.

[14] Rudolf Hilfer. *Applications of fractional calculus in physics*. World scientific, 2000.

[15] Daniel Hsu, Sham M Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

[16] H. E. Hurst. Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, 116(1):770–799, 1951. doi: 10.1061/TACEAT.0006518. URL https://ascelibrary.org/doi/abs/10.1061/TACEAT.0006518.

[17] C Ionescu, A Lopes, Dana Copot, JA Tenreiro Machado, and Jason HT Bates. The role of fractional calculus in modeling biological phenomena: A review. *Communications in Nonlinear Science and Numerical Simulation*, 51:141–159, 2017.

[18] Plamen Ch Ivanov, Luis A Nunes Amaral, Ary L Goldberger, Shlomo Havlin, Michael G Rosenblum, Zbigniew R Struzik, and H Eugene Stanley. Multifractality in human heartbeat dynamics. *Nature*, 399(6735):461–465, 1999.

[19] Yassir Jedra and Alexandre Proutiere. Finite-time identification of stable linear systems: Optimality of the least-squares estimator, 2020. URL https://arxiv.org/abs/2003.07937.

[20] Will E. Leland, Walter Willinger, Murad S. Taqqu, and Daniel V. Wilson. On the self-similar nature of ethernet traffic. *Comput. Commun. Rev.*, 25:202–213, 1993. URL https://api.semanticscholar.org/CorpusID:6011907.

[21] Klaus Linkenkaer-Hansen, Vadim Nikulin, J. Matias Palva, and Risto Ilmoniemi. Long-range temporal correlations and scaling behavior in human brain oscillations. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 21:1370–7, 03 2001. doi: 10.1523/JNEUROSCI.21-04-01370.2001.

[22] Brian N Lundstrom, Matthew H Higgs, William J Spain, and Adrienne L Fairhall. Fractional differentiation by neocortical pyramidal neurons. *Nature neuroscience*, 11(11):1335–1342, 2008.

[23] Benoit B. Mandelbrot and James R. Wallis. Some long-run properties of geophysical records. *Water Resources Research*, 5(2):321–340, 1969. doi: https://doi.org/10.1029/WR005i002p00321. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/WR005i002p00321.

[24] Kenneth S. Miller and Bertram Ross. *An introduction to the fractional calculus and fractional differential equations*. A Wiley-Interscience publication. Wiley, 1993. URL https://cir.nii.ac.jp/crid/1971712334804565506.

[25] Concepción A Monje, YangQuan Chen, Blas M Vinagre, Dingyu Xue, and Vicente Feliu-Batlle. *Fractional-order systems and controls: fundamentals and applications*. Springer Science & Business Media, 2010.

[26] Keith Oldham and Jerome Spanier. *The fractional calculus theory and applications of differentiation and integration to arbitrary order*. Elsevier, 1974.

[27] Samet Oymak and Necmiye Ozay. Non-asymptotic identification of lti systems from a single trajectory. *2019 American Control Conference (ACC)*, pages 5655–5661, 2018.

[28] Samet Oymak and Necmiye Ozay. Non-asymptotic identification of lti systems from a single trajectory, 2019. URL https://arxiv.org/abs/1806.05722.

[29] Ivo Petráš. Stability of fractional-order systems. In *Fractional-Order Nonlinear Systems: Modeling, Analysis and Simulation*, pages 55–101. Springer, 2021.

[30] Igor Podlubny. *Fractional differential equations: an introduction to fractional derivatives, fractional differential equations, to methods of their solution and some of their applications*, volume 198. elsevier, 1998.

[31] Margarita Rivero, Sergei V Rogosin, Jose A Tenreiro Machado, and Juan J Trujillo. Stability of fractional order systems. *Mathematical Problems in Engineering*, 2013(1):356215, 2013.

[32] Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, 2018.

[33] Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. Finite time lti system identification. *Journal of Machine Learning Research*, 22(26):1–61, 2021.

[34] Max Simchowitz, Horia Mania, Stephen Tu, Michael I. Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 439–473. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/simchowitz18a.html.

[35] Yue Sun, Samet Oymak, and Maryam Fazel. Finite sample system identification: Optimal rates and the role of regularization. In Alexandre M. Bayen, Ali Jadbabaie, George Pappas, Pablo A. Parrilo, Benjamin Recht, Claire Tomlin, and Melanie Zeilinger, editors, *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 16–25. PMLR, 10–11 Jun 2020.

[36] Walter Willinger, Vern Paxson, Rolf H Riedi, and Murad S Taqqu. Long-range dependence and data network traffic. *Theory and applications of long-range dependence*, pages 373–407, 2003.

[37] Yuankun Xue, Saul Rodriguez, and Paul Bogdan. A spatio-temporal fractal model for a cps approach to brain-machine-body interfaces. In *2016 Design, Automation Test in Europe Conference Exhibition (DATE)*, pages 642–647, 2016.

[38] Bahram Yaghooti and Bruno Sinopoli. Inferring dynamics of discrete-time, fractional-order control-affine nonlinear systems. In *2023 American Control Conference (ACC)*, pages 935–940. IEEE, 2023.

[39] Xiaole Zhang, Vijay Gupta, and Paul Bogdan. A sampling complexity-aware framework for discrete-time fractional-order dynamical system identification. In *2025 American Control Conference (ACC)*, pages 5093–5098, 2025. doi: 10.23919/ACC63710.2025.11107451.

[40] Xiaole Zhang, Peiyu Zhang, Xiongye Xiao, Shixuan Li, Vasileios Tzoumas, Vijay Gupta, and Paul Bogdan. End-to-end learning framework for solving non-markovian optimal control. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=1k4dKH1XOz.

[41] Ziyi Zhang, Yorie Nakahira, and Guannan Qu. Learning to stabilize unknown LTI systems on a single trajectory under stochastic noise. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025. URL https://openreview.net/forum?id=duNaSFJ1sF.

[42] Ziyi Zhang, Yorie Nakahira, and Guannan Qu. Stabilizing LTI systems under partial observability: Sample complexity and fundamental limits. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=KwHsZJatB8.

[43] Yang Zheng and Na Li. Non-asymptotic identification of linear dynamical systems using multiple trajectories. *IEEE Control Systems Letters*, PP:1–1, 12 2020. doi: 10.1109/LCSYS.2020.3042924.

[44] Ingvar M Ziemann, Henrik Sandberg, and Nikolai Matni. Single trajectory nonparametric learning of nonlinear dynamics. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 3333–3364. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/ziemann22a.html.

[45] Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1), 2019. ISSN 2306-5729. doi: 10.3390/data4010014. URL https://www.mdpi.com/2306-5729/4/1/14.

Content

A Experiment Details **15**

A.1 Synthetic Experiments 15

A.2 Real-World Experiments 15

B Proof of Theorem 1 **16**

B.1 Controlling In-Sample Error via Martingale Offset Complexity 16

B.2 Lower Isometry for the Row-Wise Grid-Search Estimator 21

C Proof of Theorem 2 **36**

C.1 Proof of Lemma 3 36

C.2 Proof of Lemma 4 37

D Auxiliary Lemmas **39**

A Experiment Details

The code to reproduce our experiments is available at: `https://github.com/Eric0312/FO-GS`.

A.1 Synthetic Experiments

We generate trajectories from two-dimensional fractional-order LTI systems. For each synthetic system, $\alpha_{i,\star}$ are sampled independently and uniformly from $[0.1, 0.5]$, and $A_\star$ is generated with eigenvalues sampled uniformly from $[-0.5, 0.5]$. We compare *FO-GS* with two baselines. *FO-BS* uses a truncated lifted state representation with memory length $p = 40$, binary search tolerance 10^{-2}, search interval $[0.05, 0.55]$, and ridge parameter 10^{-6}. *FO-WT* uses Haar wavelets with minimum level 2, maximum level chosen automatically, linear detrending, and ridge parameter 10^{-6} for the subsequent OLS step. Unless otherwise specified, *FO-GS* searches over $[0.05, 0.55]$ using 20 equally spaced grid points per coordinate and ridge parameter 10^{-6}. We report the MSE of both $\boldsymbol{\alpha}_\star$ and $A_\star$, averaged over 5 matched systems and 20 Monte Carlo trials.

Varying trajectory horizons. To study the effect of trajectory length, we vary the horizon over $t \in \{50, 100, 150, 200, 300, 400, 500\}$ while fixing the noise scale to $\sigma = 0.1$. The initial state is sampled from a zero mean Gaussian distribution with standard deviation 4.0.

Varying noise scales. To evaluate robustness to process noise, we fix the trajectory horizon at $t = 200$ and vary the noise scale over $\sigma \in \{0.02, 0.05, 0.10, 0.15, 0.20, 0.30, 0.40\}$. The initial state is sampled from a zero mean Gaussian distribution with standard deviation 4.0.

Varying grid sizes. To examine the effect of grid resolution in *FO-GS*, we fix $t = 100$ and $\sigma = 0.01$ and vary the number of grid points over $M \in \{3, 5, 10, 13, 16, 20, 25\}$. The initial state is sampled from a zero mean Gaussian distribution with standard deviation 2.0.

Small $\boldsymbol{\alpha}_\star$ regime. To examine whether the slower small $\boldsymbol{\alpha}_\star$ rates are caused by finite sample effects, we further extend the horizon to $t \in \{100, 200, 400, 800, 1600, 3200, 6400, 12800, 25600\}$. Small and large regimes in Table 1 correspond to $\boldsymbol{\alpha}_\star \sim \mathcal{U}[0.01, 0.20]$ and $\boldsymbol{\alpha}_\star \sim \mathcal{U}[0.50, 0.99]$, respectively.

A.2 Real-World Experiments

EEG preprocessing. We evaluate all methods on a multi-subject EEG dataset with 36 subjects and $n = 19$ channels. Each subject is treated as a multivariate time series. We split each subject trajectory into non-overlapping windows of length $W = 150$ with stride $S = 150$. For each window, the first 70% of samples are used for training and the remaining 30% for testing, giving 105 training samples

and 45 test samples per window. The experiments are run on the raw EEG data without additional normalization.

Hyperparameters. We fix the search interval $[0.05, 0.95]$. *FO-GS* uses 50 equally spaced grid points for each row-wise search. *FO-BS* uses binary search with tolerance 10^{-2} and a truncated lifted state representation with memory length $p = 40$. *FO-WT* uses Haar wavelets with linear detrending and estimates the fractional order from a weighted log-variance regression over wavelet levels $2, 3, 4$.

Evaluation metric. We evaluate methods by one-step prediction. We report training and test NMSE, where $\mathrm{NMSE} = \frac{\sum_t \|\hat{x}_t - x_t\|_2^2}{\sum_t \|x_t\|_2^2}$.

B Proof of Theorem 1

B.1 Controlling In-Sample Error via Martingale Offset Complexity

We now adapt the offset martingale complexity argument [44] to the row-wise grid-search estimator used in Algorithm 1. The main difference from the idealized continuous empirical risk minimizer (ERM) is that the true parameter $\alpha_{i,\star}$ need not belong to the finite search grid. Consequently, an additional discretization term appears in the basic inequality [44].

For each coordinate $i \in [n]$, define the row-wise empirical loss

$$L^{(i)}(\alpha_i, a_i) := \sum_{s=0}^{t-1} \left| \Delta^{\alpha_i} x_{s+1}^{(i)} - a_i x_s \right|^2,$$

where $a_i \in \mathbb{R}^{1\times n}$ denotes the i-th row of A. For any candidate α_i, let

$$\hat{a}_i(\alpha_i) \in \arg\min_{a_i \in \mathbb{R}^{1\times n}} L^{(i)}(\alpha_i, a_i).$$

The row-search estimator is

$$\hat{\alpha}_i \in \arg\min_{\alpha_i \in \mathcal{A}_{\epsilon,i}} L^{(i)}\big(\alpha_i, \hat{a}_i(\alpha_i)\big), \qquad \hat{a}_i := \hat{a}_i(\hat{\alpha}_i).$$

Let α_i° denote the grid point closest to the true parameter:

$$\alpha_i^\circ \in \arg\min_{\alpha_i \in \mathcal{A}_{\epsilon,i}} |\alpha_i - \alpha_{i,\star}|. \tag{13}$$

In particular, if the grid resolution is ϵ_i, then

$$|\alpha_i^\circ - \alpha_{i,\star}| \le \epsilon_i.$$

Define

$$b_s^{(i)}(\alpha_i) := (\Delta^{\alpha_i} - \Delta^{\alpha_{i,\star}})\, x_{s+1}^{(i)}.$$

Since the true dynamics satisfy

$$\Delta^{\alpha_{i,\star}} x_{s+1}^{(i)} = a_{i,\star} x_s + \eta_s^{(i)}, \tag{14}$$

we have

$$\Delta^{\alpha_i} x_{s+1}^{(i)} - a_i x_s = \eta_s^{(i)} + b_s^{(i)}(\alpha_i) - (a_i - a_{i,\star}) x_s. \tag{15}$$

For convenience, define

$$r_s^{(i)}(\alpha_i, \Delta a_i) := b_s^{(i)}(\alpha_i) - \Delta a_i x_s, \qquad \Delta a_i := a_i - a_{i,\star},$$

and

$$\hat{r}_s^{(i)} := r_s^{(i)}\left(\hat{\alpha}_i, \hat{a}_i - a_{i,\star}\right).$$

Lemma 5 (Row-wise basic inequality)**.** *For every $i \in [n]$,*

$$\begin{aligned}\sum_{s=0}^{t-1} \left|\hat{r}_s^{(i)}\right|^2 \le & 4 \sum_{s=0}^{t-1} \left\langle -\eta_s^{(i)}, \hat{r}_s^{(i)} \right\rangle - \sum_{s=0}^{t-1} \left|\hat{r}_s^{(i)}\right|^2 \\ & + 4 \sum_{s=0}^{t-1} \left\langle \eta_s^{(i)}, b_s^{(i)}(\alpha_i^\circ) \right\rangle + 2 \sum_{s=0}^{t-1} \left| b_s^{(i)}(\alpha_i^\circ) \right|^2. \end{aligned} \tag{16}$$

Proof. By optimality of the row-wise grid-search estimator,

$$L^{(i)}\left(\hat{\alpha}_i, \hat{a}_i\right) \leq L^{(i)}\left(\alpha_i^{\circ}, \hat{a}_i(\alpha_i^{\circ})\right) \leq L^{(i)}\left(\alpha_i^{\circ}, a_{i,\star}\right).$$

Using (15), this gives

$$\sum_{s=0}^{t-1} \left|\eta_s^{(i)} + \hat{r}_s^{(i)}\right|^2 \leq \sum_{s=0}^{t-1} \left|\eta_s^{(i)} + b_s^{(i)}(\alpha_i^{\circ})\right|^2.$$

Expanding both sides and cancelling $\sum_{s=0}^{t-1} |\eta_s^{(i)}|^2$ yields

$$\begin{aligned}\sum_{s=0}^{t-1} \left|\hat{r}_s^{(i)}\right|^2 \leq & -2\sum_{s=0}^{t-1} \left\langle \eta_s^{(i)}, \hat{r}_s^{(i)} \right\rangle \\ & + 2\sum_{s=0}^{t-1} \left\langle \eta_s^{(i)}, b_s^{(i)}(\alpha_i^{\circ}) \right\rangle + \sum_{s=0}^{t-1} \left|b_s^{(i)}(\alpha_i^{\circ})\right|^2. \end{aligned} \tag{17}$$

Multiplying (17) by two and subtracting $\sum_s |\hat{r}_s^{(i)}|^2$ from both sides gives (16). □

The first line of (16) can now be controlled by an offset martingale complexity argument. In particular, since $(\hat{\alpha}_i, \hat{a}_i - a_{i,\star})$ is an admissible choice,

$$\begin{aligned}\sum_{s=0}^{t-1} |\hat{r}_s^{(i)}|^2 \leq \max_{\alpha_i \in \mathcal{A}_{\epsilon,i}} \sup_{\Delta a_i \in \mathbb{R}^{1\times n}} \Bigg\{ & 4\sum_{s=0}^{t-1} \left\langle -\eta_s^{(i)}, b_s^{(i)}(\alpha_i) - \Delta a_i x_s \right\rangle \\ & - \sum_{s=0}^{t-1} \left|b_s^{(i)}(\alpha_i) - \Delta a_i x_s\right|^2 \Bigg\} + \Gamma_{t,i}^{\mathrm{grid}}, \end{aligned}$$

where

$$\Gamma_{t,i}^{\mathrm{grid}} := 4\sum_{s=0}^{t-1} \left\langle \eta_s^{(i)}, b_s^{(i)}(\alpha_i^{\circ}) \right\rangle + 2\sum_{s=0}^{t-1} \left|b_s^{(i)}(\alpha_i^{\circ})\right|^2$$

is the additional error induced by the finite grid.

We next explicitly perform the maximization over Δa_i. Recall from (9) that

$$X_t = [x_0, \ldots, x_{t-1}], \qquad X_t X_t^{\top} = \sum_{s=0}^{t-1} x_s x_s^{\top}.$$

For a fixed α_i, let

$$y_s^{(i)}(\alpha_i) := b_s^{(i)}(\alpha_i) + 2\eta_s^{(i)}.$$

Then

$$\begin{aligned} & 4\sum_{s=0}^{t-1} \left\langle -\eta_s^{(i)}, b_s^{(i)}(\alpha_i) - \Delta a_i x_s \right\rangle - \sum_{s=0}^{t-1} \left|b_s^{(i)}(\alpha_i) - \Delta a_i x_s\right|^2 \\ & = -4\sum_{s=0}^{t-1} \left\langle \eta_s^{(i)}, b_s^{(i)}(\alpha_i) \right\rangle - \sum_{s=0}^{t-1} \left|b_s^{(i)}(\alpha_i)\right|^2 + 2\left\langle \sum_{s=0}^{t-1} y_s^{(i)}(\alpha_i) x_s^{\top}, \Delta a_i \right\rangle - \Delta a_i X_t X_t^{\top} \Delta a_i^{\top}. \end{aligned}$$

Assuming $X_t X_t^{\top} \succ 0$, the maximizer is

$$\Delta a_i^*(\alpha_i) = \left(\sum_{s=0}^{t-1} y_s^{(i)}(\alpha_i) x_s^{\top}\right) (X_t X_t^{\top})^{-1}.$$

Therefore,

$$\begin{aligned} & \sup_{\Delta a_i} \left\{ 4\sum_{s=0}^{t-1} \left\langle -\eta_s^{(i)}, b_s^{(i)}(\alpha_i) - \Delta a_i x_s \right\rangle - \sum_{s=0}^{t-1} \left|b_s^{(i)}(\alpha_i) - \Delta a_i x_s\right|^2 \right\} \\ & = U_{t,i}(\alpha_i) + V_{t,i}(\alpha_i), \end{aligned}$$

where

$$U_{t,i}(\alpha_i) := 4\sum_{s=0}^{t-1}\left\langle -\eta_s^{(i)}, b_s^{(i)}(\alpha_i)\right\rangle - \sum_{s=0}^{t-1}\left|b_s^{(i)}(\alpha_i)\right|^2,$$

and

$$V_{t,i}(\alpha_i) := \left\|\left(\sum_{s=0}^{t-1} y_s^{(i)}(\alpha_i)x_s^\top\right)(X_tX_t^\top)^{-1/2}\right\|_2^2.$$

Combining everything together gives the following row-wise in-sample error bound:

$$\sum_{s=0}^{t-1}\left|b_s^{(i)}(\hat{\alpha}_i) - (\hat{a}_i - a_{i,\star})x_s\right|^2 \le \max_{\alpha_i\in\mathcal{A}_{\epsilon,i}}\{U_{t,i}(\alpha_i) + V_{t,i}(\alpha_i)\} + \Gamma_{t,i}^{\mathrm{grid}}.$$

Finally, summing over all coordinates $i \in [n]$ gives

$$\sum_{s=0}^{t-1}\left\|\left(\Delta^{\hat{\boldsymbol{\alpha}}} - \Delta^{\boldsymbol{\alpha}^\star}\right)x_{s+1} - (\hat{A} - A_\star)x_s\right\|_2^2 \le \sum_{i=1}^{n}\max_{\alpha_i\in\mathcal{A}_{\epsilon,i}}\{U_{t,i}(\alpha_i) + V_{t,i}(\alpha_i)\} + \Gamma_t^{\mathrm{grid}},$$

where

$$\Gamma_t^{\mathrm{grid}} := \sum_{i=1}^{n}\Gamma_{t,i}^{\mathrm{grid}}.$$

Equivalently, letting

$$\boldsymbol{\alpha}^\circ := (\alpha_1^\circ, \ldots, \alpha_n^\circ),$$

and defining the matrices

$$B_t(\boldsymbol{\alpha}^\circ) := [b_0(\boldsymbol{\alpha}^\circ), \quad \cdots, \quad b_{t-1}(\boldsymbol{\alpha}^\circ)], \qquad W_t := [\eta_0, \quad \cdots, \quad \eta_{t-1}],$$

the grid-discretization contribution can be written compactly as

$$\Gamma_t^{\mathrm{grid}} = 4\,\langle W_t, B_t(\boldsymbol{\alpha}^\circ)\rangle_F + 2\,\|B_t(\boldsymbol{\alpha}^\circ)\|_F^2.$$

Remark 1. If the true parameter lies exactly on the search grid, i.e., $\alpha_{i,\star} \in \mathcal{A}_{\epsilon,i}$ for every i, then we may take $\alpha_i^\circ = \alpha_{i,\star}$. In this case

$$b_s^{(i)}(\alpha_i^\circ) = 0$$

for every i and s, and hence

$$\Gamma_t^{\mathrm{grid}} = 0.$$

Thus, the continuous ERM offset inequality is recovered as a special case. For a finite grid, however, the additional term Γ_t^{grid} must be retained.

Remark 2. Since the fractional-order operator is diagonal across state coordinates and the estimator searches for each α_i separately, the relevant offset complexity decomposes as

$$\sum_{i=1}^{n}\max_{\alpha_i\in\mathcal{A}_{\epsilon,i}}\{U_{t,i}(\alpha_i) + V_{t,i}(\alpha_i)\},$$

rather than requiring a supremum over the full Cartesian grid $\mathcal{A}_{\epsilon,1}\times\cdots\times\mathcal{A}_{\epsilon,n}$. This row-wise decomposition is important for obtaining the sharp complexity dependence of the grid-search estimator.

Lemma 6 (Parameterized bound for the offset term)**.** *Fix $\boldsymbol{\alpha}$ and assume $X_tX_t^\top \succ 0$. For an arbitrary auxiliary matrix perturbation $\Delta A \in \mathbb{R}^{n\times n}$, where $\Delta A := A - A_\star$, consider*

$$\Phi_{\boldsymbol{\alpha}}(\Delta A) = 4\sum_{s=0}^{t-1}\langle -\eta_s, b_s(\boldsymbol{\alpha}) - \Delta A x_s\rangle - \sum_{s=0}^{t-1}\|b_s(\boldsymbol{\alpha}) - \Delta A x_s\|_2^2. \tag{18}$$

For any $\tau \in (0,1]$, define

$$U_{t,\tau}(\boldsymbol{\alpha}) := -4\langle W_t, B_t(\boldsymbol{\alpha})\rangle_F - \tau\|B_t(\boldsymbol{\alpha})\|_F^2. \tag{19}$$

Then

$$U_t(\boldsymbol{\alpha}) = U_{t,\tau}(\boldsymbol{\alpha}) - (1-\tau)\|B_t(\boldsymbol{\alpha})\|_F^2. \tag{20}$$

Moreover, for every fixed deterministic $\boldsymbol{\alpha}$ and $\delta \in (0,1)$,

$$\Pr\left(U_{t,\tau}(\boldsymbol{\alpha}) \le \frac{8\sigma^2}{\tau}\log\frac{1}{\delta}\right) \ge 1-\delta. \tag{21}$$

For the row-wise grid search, define

$$U_{t,\tau,i}(\alpha_i) := -4\sum_{s=0}^{t-1}\eta_s^{(i)}b_s^{(i)}(\alpha_i) - \tau\sum_{s=0}^{t-1}|b_s^{(i)}(\alpha_i)|^2. \tag{22}$$

Then, with probability at least $1-\delta$, simultaneously for all $i \in [n]$,

$$\max_{\alpha_i\in\mathcal{A}_{\epsilon,i}} U_{t,\tau,i}(\alpha_i) \le \frac{8\sigma^2}{\tau}\log\left(\frac{M_i n}{\delta}\right). \tag{23}$$

Proof. Expanding (18) gives

$$\begin{aligned}\Phi_{\boldsymbol{\alpha}}(\Delta A) = &-4\langle W_t, B_t(\boldsymbol{\alpha})\rangle_F - \|B_t(\boldsymbol{\alpha})\|_F^2 \\ &+ 2\left\langle \big(B_t(\boldsymbol{\alpha})+2W_t\big)X_t^\top, \Delta A\right\rangle_F - \operatorname{tr}\big(\Delta A X_t X_t^\top \Delta A^\top\big).\end{aligned} \tag{24}$$

The maximizer of this auxiliary offset objective is

$$\Delta A_{\mathrm{opt}} = \big(B_t(\boldsymbol{\alpha})+2W_t\big)X_t^\top(X_tX_t^\top)^{-1}.$$

Substitution gives

$$\sup_{\Delta A}\Phi_{\boldsymbol{\alpha}}(\Delta A) = U_t(\boldsymbol{\alpha}) + V_t(\boldsymbol{\alpha}),$$

where

$$U_t(\boldsymbol{\alpha}) = -4\langle W_t, B_t(\boldsymbol{\alpha})\rangle_F - \|B_t(\boldsymbol{\alpha})\|_F^2,$$

and

$$V_t(\boldsymbol{\alpha}) = \left\|\big(B_t(\boldsymbol{\alpha})+2W_t\big)X_t^\top(X_tX_t^\top)^{-1/2}\right\|_F^2.$$

Then (20) follows directly from the definitions.

For fixed deterministic $\boldsymbol{\alpha}$, $b_s(\boldsymbol{\alpha})$ is $\mathcal{F}_s$-measurable. Hence, for $\lambda > 0$,

$$\mathbb{E}\left[\exp\left(\lambda\left[-4\langle\eta_s, b_s(\boldsymbol{\alpha})\rangle - \tau\|b_s(\boldsymbol{\alpha})\|_2^2\right]\right)\middle|\mathcal{F}_s\right] = \exp\left((8\sigma^2\lambda^2-\tau\lambda)\|b_s(\boldsymbol{\alpha})\|_2^2\right).$$

The right-hand side is at most one whenever $0 < \lambda \le \tau/(8\sigma^2)$. Iterating conditional expectations, taking $\lambda = \tau/(8\sigma^2)$, and applying Markov's inequality proves (21). The row-wise statement follows by taking failure probability $\delta/(nM_i)$ for each fixed grid point and applying a union bound over all rows and grid points. □

Lemma 7 (High-probability bound for $V_t(\boldsymbol{\alpha})$)**.** *Define*

$$\mathfrak{r}_t(\delta,k) := \frac{C}{\sqrt{t\lambda_{\min}(\Gamma_k)}}\left(n\log\frac{n}{\delta} + \log\det(\Gamma_t\Gamma_k^{-1})\right)^{1/2},$$

where $C > 0$ is the universal constant in Lemma 3. Suppose that

$$\frac{t}{k} \ge c\left(n\log\frac{n}{\delta} + \log\det(\Gamma_t\Gamma_k^{-1})\right).$$

For the row-wise search intervals $\mathcal{A}_i = [\underline{\alpha}_i, \overline{\alpha}_i]$, define

$$S_{1,i} := \sup_{u\in\mathcal{A}_i}\sum_{j\ge1}|\partial_u\psi(u,j)|, \qquad S_1 := \max_{1\le i\le n} S_{1,i}. \tag{25}$$

Then, with probability at least $1-\delta$, simultaneously for all $\boldsymbol{\alpha}$ in the search set,

$$V_t(\boldsymbol{\alpha}) \le \operatorname{tr}(X_tX_t^\top)\left(S_1\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_\star\|_\infty + 2\mathfrak{r}_t(\delta,k)\right)^2. \tag{26}$$

In particular,

$$\begin{aligned}V_t(\boldsymbol{\alpha}) \le\ & 2S_1^2\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_\star\|_\infty^2\operatorname{tr}(X_tX_t^\top) \\ &+ \frac{8C^2\operatorname{tr}(X_tX_t^\top)}{t\lambda_{\min}(\Gamma_k)}\left(n\log\frac{n}{\delta} + \log\det(\Gamma_t\Gamma_k^{-1})\right).\end{aligned}$$

Proof. Since

$$\sum_{s=0}^{t-1}\big(b_s(\boldsymbol{\alpha})+2\eta_s\big)x_s^\top = \big(B_t(\boldsymbol{\alpha})+2W_t\big)X_t^\top,$$

we have

$$V_t(\boldsymbol{\alpha})^{1/2} \le \left\|B_t(\boldsymbol{\alpha})X_t^\top(X_tX_t^\top)^{-1/2}\right\|_F + 2\left\|W_tX_t^\top(X_tX_t^\top)^{-1/2}\right\|_F. \tag{27}$$

Let

$$P_X := X_t^\top(X_tX_t^\top)^{-1}X_t.$$

Then P_X is an orthogonal projection, so

$$\left\|B_t(\boldsymbol{\alpha})X_t^\top(X_tX_t^\top)^{-1/2}\right\|_F^2 = \operatorname{tr}\big(B_t(\boldsymbol{\alpha})P_XB_t(\boldsymbol{\alpha})^\top\big) \le \|B_t(\boldsymbol{\alpha})\|_F^2.$$

The convolution bound established in the proof of Lemma 4 gives

$$\|B_t(\boldsymbol{\alpha})\|_F^2 \le S_1^2\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_\star\|_\infty^2\operatorname{tr}(X_tX_t^\top).$$

Hence

$$\left\|B_t(\boldsymbol{\alpha})X_t^\top(X_tX_t^\top)^{-1/2}\right\|_F \le S_1\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_\star\|_\infty\sqrt{\operatorname{tr}(X_tX_t^\top)}. \tag{28}$$

Moreover,

$$W_tX_t^\top(X_tX_t^\top)^{-1/2} = \big(W_tX_t^\top(X_tX_t^\top)^{-1}\big)\,(X_tX_t^\top)^{1/2},$$

and therefore Lemma 3 implies

$$\left\|W_tX_t^\top(X_tX_t^\top)^{-1/2}\right\|_F \le \mathfrak{r}_t(\delta,k)\sqrt{\operatorname{tr}(X_tX_t^\top)} \tag{29}$$

with probability at least $1-\delta$. Substituting (28) and (29) into (27) proves (26); the expanded form follows from $(a+b)^2 \le 2a^2+2b^2$. □

Lemma 8 (Absorbable high-probability in-sample error bound)**.** *For $\tau \in (0,1]$ and $\rho > 0$, suppose that*

$$\frac{t}{k} \ge c\left(n\log\frac{3n}{\delta} + \log\det(\Gamma_t\Gamma_k^{-1})\right).$$

Then, with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{s=0}^{t-1}\left\|\left(\Delta^{\hat{\boldsymbol{\alpha}}}-\Delta^{\boldsymbol{\alpha}_\star}\right)x_{s+1} - (\hat{A}-A_\star)x_s\right\|_2^2 \\
&\quad\le (\tau+\rho)S_1^2\|\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_\star\|_\infty^2\operatorname{tr}(X_tX_t^\top) + 3S_1^2\epsilon_{\max}^2\operatorname{tr}(X_tX_t^\top) \\
&\qquad + \frac{4(1+\rho^{-1})C^2\operatorname{tr}(X_tX_t^\top)}{t\lambda_{\min}(\Gamma_k)}\left(n\log\frac{3n}{\delta} + \log\det(\Gamma_t\Gamma_k^{-1})\right) \\
&\qquad + \frac{8\sigma^2}{\tau}\sum_{i=1}^{n}\log\left(\frac{3M_in}{\delta}\right) + 8\sigma^2\log\left(\frac{3}{\delta}\right).
\end{aligned} \tag{30}$$

Proof. For each row, recall

$$\Delta a_i := \hat{a}_i - a_{i,\star}, \qquad \hat{r}_s^{(i)} := b_s^{(i)}(\hat{\alpha}_i) - \Delta a_i x_s.$$

The row-wise basic inequality in Lemma 5 and the quadratic maximization in Lemma 6 give

$$\sum_{s=0}^{t-1}|\hat{r}_s^{(i)}|^2 \le U_{t,i}(\hat{\alpha}_i) + V_{t,i}(\hat{\alpha}_i) + \Gamma_{t,i}^{\mathrm{grid}}.$$

Summing over rows,

$$\mathcal{E}_t := \sum_{s=0}^{t-1}\left\|\left(\Delta^{\hat{\boldsymbol{\alpha}}}-\Delta^{\boldsymbol{\alpha}_\star}\right)x_{s+1} - (\hat{A}-A_\star)x_s\right\|_2^2 \le U_t(\hat{\boldsymbol{\alpha}}) + V_t(\hat{\boldsymbol{\alpha}}) + \Gamma_t^{\mathrm{grid}}. \tag{31}$$

By (20),

$$U_t(\hat{\boldsymbol{\alpha}}) = U_{t,\tau}(\hat{\boldsymbol{\alpha}}) - (1-\tau)\|B_t(\hat{\boldsymbol{\alpha}})\|_F^2.$$

Young's inequality and the projection inequality give

$$U_t(\hat{\boldsymbol{\alpha}}) + V_t(\hat{\boldsymbol{\alpha}}) \le U_{t,\tau}(\hat{\boldsymbol{\alpha}}) + (\tau+\rho)\|B_t(\hat{\boldsymbol{\alpha}})\|_F^2 + 4(1+\rho^{-1})\|W_t X_t^\top (X_t X_t^\top)^{-1/2}\|_F^2.$$

Lemma 6, with failure probability $\delta/3$, gives

$$U_{t,\tau}(\hat{\boldsymbol{\alpha}}) \le \frac{8\sigma^2}{\tau}\sum_{i=1}^{n}\log\left(\frac{3M_i n}{\delta}\right).$$

Also, the convolution bound established in the proof of Lemma 4 gives

$$\|B_t(\hat{\boldsymbol{\alpha}})\|_F^2 \le S_1^2\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_\star\|_\infty^2 \operatorname{tr}(X_t X_t^\top).$$

Next, retain the full good event used in the proof of Lemma 3. With failure probability $\delta/3$, this event simultaneously gives

$$\left\|W_t X_t^\top (X_t X_t^\top)^{-1}\right\|_{\mathrm{op}} \le \mathfrak{r}_t\left(\frac{\delta}{3}, k\right)$$

and the complement of the event $\mathcal{E}_3$ from that proof, namely

$$X_t X_t^\top \preceq t\,\overline{\Gamma}_t\left(\frac{\delta}{3}\right), \qquad \overline{\Gamma}_t(\delta_0) := \frac{\sigma^2 n}{\delta_0}\Gamma_t. \tag{32}$$

Consequently,

$$\|W_t X_t^\top (X_t X_t^\top)^{-1/2}\|_F^2 \le \mathfrak{r}_t^2\left(\frac{\delta}{3}, k\right)\operatorname{tr}(X_t X_t^\top).$$

Thus the upper empirical state energy control in (32) does not require a fourth event or an additional allocation of the failure probability.

Finally, let $B_t^\circ := B_t(\boldsymbol{\alpha}^\circ)$. Then

$$\Gamma_t^{\mathrm{grid}} = \left(4\langle W_t, B_t^\circ\rangle_F - \|B_t^\circ\|_F^2\right) + 3\|B_t^\circ\|_F^2.$$

The same conditional moment generating function argument in Lemma 6, with failure probability $\delta/3$, yields

$$4\langle W_t, B_t^\circ\rangle_F - \|B_t^\circ\|_F^2 \le 8\sigma^2\log\left(\frac{3}{\delta}\right),$$

while the convolution bound gives

$$\|B_t^\circ\|_F^2 \le S_1^2\epsilon_{\max}^2\operatorname{tr}(X_t X_t^\top).$$

Substituting these three bounds into (31) and applying a union bound proves (30) and (32). □

B.2 Lower Isometry for the Row-Wise Grid-Search Estimator

The row-wise structure of *FO-GS* allows the lower isometry analysis to be carried out coordinate-wise. For each row i, we profile out the corresponding row of A and study the resulting noiseless prediction error as a function of α_i. This quantity captures the curvature in the fractional-order parameter and will be used to relate the in-sample prediction error to the estimation error in α_i.

For each $i \in [n]$, recall

$$b_s^{(i)}(\alpha_i) := (\Delta^{\alpha_i} - \Delta^{\alpha_{i,\star}})\,x_{s+1}^{(i)}, \qquad \Delta a_i := a_i - a_{i,\star},$$

and define

$$Q_{t,i}(\alpha_i, \Delta a_i) := \sum_{s=0}^{t-1}\left|b_s^{(i)}(\alpha_i) - \Delta a_i x_s\right|^2.$$

Since *FO-GS* profiles out a_i for every candidate α_i, it is natural to introduce the profiled noiseless error

$$\underline{Q}_{t,i}(\alpha_i) := \inf_{\Delta a_i \in \mathbb{R}^{1\times n}} Q_{t,i}(\alpha_i, \Delta a_i).$$

In particular,

$$Q_{t,i}\left(\hat{\alpha}_i, \hat{a}_i - a_{i,\star}\right) \ge \underline{Q}_{t,i}(\hat{\alpha}_i).$$

Lemma 9. *Suppose Assumption 1 holds. For each $i \in [n]$, define*

$$\begin{aligned}\underline{\alpha}_{i,\mathrm{loc}} &:= \frac{\alpha_{i,\star}}{2},\\ \mathfrak{Z}_i &:= \zeta\Big(1+\frac{\alpha_{i,\star}}{2}\Big) - 2\zeta'\Big(1+\frac{\alpha_{i,\star}}{2}\Big) + \zeta''\Big(1+\frac{\alpha_{i,\star}}{2}\Big),\\ K_i &:= \frac{9e}{4}\sigma\widetilde{C}_G\sqrt{\zeta(2)}\,\mathfrak{Z}_i.\end{aligned}$$

Then, for every $\alpha_i \in [\underline{\alpha}_{i,\mathrm{loc}}, 1]$,

$$\frac{1}{t}\sum_{s=0}^{t-1}\mathbb{E}\left|b_s^{(i)}(\alpha_i) - (\alpha_i-\alpha_{i,\star})\sum_{j\ge 1}\partial_\alpha\psi(\alpha_{i,\star},j)x_{s+1-j}^{(i)}\right|^2 \le \frac{K_i^2}{4}|\alpha_i-\alpha_{i,\star}|^4.$$

Proof. For row i, let

$$S_2 := \sup_{u\in[a,1]}\sum_{j=1}^{\infty}|\partial_u^2\psi(u,j)|, \qquad a>0$$

and

$$\rho_s^{(i)}(\alpha) := b_s^{(i)}(\alpha) - (\alpha-\alpha_{i,\star})\sum_{j=1}^{s+1}\partial_\alpha\psi(\alpha_{i,\star},j)x_{s+1-j}^{(i)}.$$

Taylor's theorem in integral form gives

$$\psi(\alpha,j)-\psi(\alpha_{i,\star},j)-(\alpha-\alpha_{i,\star})\,\partial_\alpha\psi(\alpha_{i,\star},j) = (\alpha-\alpha_{i,\star})^2\int_0^1(1-\tau)\partial_\alpha^2\psi(\alpha_{i,\star}+\tau(\alpha-\alpha_{i,\star}),j)\,d\tau.$$

Therefore,

$$\rho_s^{(i)}(\alpha) = (\alpha-\alpha_{i,\star})^2\int_0^1(1-\tau)\sum_{j=1}^{s+1}\partial_\alpha^2\psi(\alpha_{i,\star}+\tau(\alpha-\alpha_{i,\star}),j)x_{s+1-j}^{(i)}\,d\tau.$$

For $j\ge 2$, let $\psi(u,j) = -\frac{u(1-u)}{j}Q_j(u)$, and $Q_j(u) := \prod_{m=2}^{j-1}\left(1-\frac{u}{m}\right)$.

Define

$$H_{1,j}(u) = \sum_{m=2}^{j-1}\frac{1}{m-u}, \qquad H_{2,j}(u) = \sum_{m=2}^{j-1}\frac{1}{(m-u)^2}.$$

Then

$$Q_j'(u) = -Q_j(u)H_{1,j}(u),$$

and

$$\left|Q_j''(u)\right| = Q_j(u)\left(H_{1,j}(u)^2 - H_{2,j}(u)\right) \le Q_j(u)H_{1,j}(u)^2.$$

For $u\in[a,1]$, where $a>0$, $Q_j(u)\le e\,j^{-u}\le e\,j^{-a}$, and $H_{1,j}(u)\le 1+\log j$,

Consequently,

$$\begin{aligned}\left|\partial_u^2\psi(u,j)\right| &\le \frac{Q_j(u)}{j}\Big[2+2H_{1,j}(u)+\frac{1}{4}H_{1,j}(u)^2\Big]\\ &\le \frac{e}{j^{1+a}}\left[2+2(1+\log j)+\frac{1}{4}(1+\log j)^2\right].\end{aligned}$$

By simplification, we get

$$\left|\partial_u^2\psi(u,j)\right| \le \frac{9}{4}e(1+\log j)^2 j^{-(1+a)}$$

for every $u\in[a,1]$. Hence

$$S_2(a) \le \frac{9}{4}e\left[\zeta(1+a) - 2\zeta'(1+a) + \zeta''(1+a)\right]. \tag{33}$$

For coordinate i, simply choose $\underline{\alpha}_{i,\mathrm{loc}} = \frac{\alpha_{i,\star}}{2}$. Then the result holds for all $\alpha \in [\underline{\alpha}_{i,\mathrm{loc}}, 1]$. Using $\mathbb{E}|x_s^{(i)}|^2 \le \sigma^2 \sum_{m=0}^{\infty} \|e_i^\top G_m\|_2^2 \le \sigma^2 \tilde{C}_G^2 \zeta(2)$ and Minkowski's inequality

$$\begin{aligned}\left\|\rho_s^{(i)}(\alpha)\right\|_{L^2} &\le (\alpha-\alpha_{i,\star})^2 \int_0^1 (1-\tau) \sum_{j=1}^{s+1} \left|\partial_\alpha^2 \psi(\alpha_{i,*}+\tau(\alpha-\alpha_{i,\star}), j)\right| \left\|x_{s+1-j}^{(i)}\right\|_{L^2} d\tau \\ &\le \frac{(\alpha-\alpha_{i,\star})^2}{2} S_2(a) \sqrt{\sigma^2 \tilde{C}_G^2 \zeta(2)}.\end{aligned}$$

Squaring, we have

$$\mathbb{E}|\rho_s^{(i)}(\alpha)|^2 \le \frac{(\alpha-\alpha_{i,\star})^4}{4} S_2(\tfrac{\alpha_{i,\star}}{2})^2 \sigma^2 \tilde{C}_G^2 \zeta(2)$$

Hence, we have

$$\frac{1}{t}\sum_{s=0}^{t-1} \mathbb{E}\left| b_s^{(i)}(\alpha) - (\alpha-\alpha_{i,\star}) \sum_{j\ge 1} \partial_\alpha \psi(\alpha_{i,\star}, j) x_{s+1-j}^{(i)} \right|^2 \le \frac{K_i^2}{4}(\alpha-\alpha_{i,\star})^4$$

with $K_i = \frac{9}{4} e \sigma \tilde{C}_G \sqrt{\zeta(2)} \left[\zeta(1+\frac{\alpha_{i,\star}}{2}) - 2\zeta'(1+\frac{\alpha_{i,\star}}{2}) + \zeta''(1+\frac{\alpha_{i,\star}}{2})\right]$ □

Recall the row-wise population risk

$$R_t^{(i)}(\alpha_i) := \frac{1}{t} \inf_{a_i} \sum_{s=0}^{t-1} \mathbb{E}\left[\left(\Delta^{\alpha_i} x_{s+1}^{(i)} - a_i x_s\right)^2\right].$$

Using the true dynamics,

$$\Delta^{\alpha_i} x_{s+1}^{(i)} - a_i x_s = \eta_s^{(i)} + b_s^{(i)}(\alpha_i) - \Delta a_i x_s.$$

Since $b_s^{(i)}(\alpha_i)$ and x_s are $\mathcal{F}_s$-measurable and $\mathbb{E}[\eta_s^{(i)} \mid \mathcal{F}_s] = 0$,

$$R_t^{(i)}(\alpha_i) - R_t^{(i)}(\alpha_{i,\star}) = \frac{1}{t} \inf_{\Delta a_i} \sum_{s=0}^{t-1} \mathbb{E}\left[\left| b_s^{(i)}(\alpha_i) - \Delta a_i x_s \right|^2\right].$$

Then we use the lower isometry argument in two stages. First, a global profiled lower isometry bound in Lemma 12 shows that, with high probability, simultaneously over all rows and all grid points outside the separation neighborhood of $\alpha_{i,\star}$,

$$\underline{\mathcal{Q}}_{t,i}(\alpha_i) \ge \frac{t}{2}\left(R_t^{(i)}(\alpha_i) - R_t^{(i)}(\alpha_{i,\star})\right).$$

Since each $\mathcal{A}_{\epsilon,i}$ is finite, the uniform statement is obtained by establishing the bound for a fixed α_i and taking a union bound over the grid points and rows. Combined with the in-sample upper bound and the population separation gap γ, this global bound rules out grid points outside the separation neighborhood and localizes the estimator to the set

$$\mathcal{G}_i := \left\{\alpha_i \in \mathcal{A}_{\epsilon,i} \cap [\alpha_{i,\mathrm{loc}}, 1] : |\alpha_i - \alpha_{i,\star}| \le \frac{\sqrt{\mu_{t,i}}}{2K_i}\right\}. \tag{34}$$

Within $\mathcal{G}_i$, the population risk has a quadratic local curvature. In particular, by Lemma 9 and the reverse triangle inequality,

$$\sqrt{R_t^{(i)}(\alpha_i) - R_t^{(i)}(\alpha_{i,\star})} \ge |\alpha_i - \alpha_{i,\star}|\sqrt{\mu_{t,i}} - \frac{K_i}{2}|\alpha_i - \alpha_{i,\star}|^2.$$

Hence, whenever $|\alpha_i - \alpha_{i,\star}| \le \sqrt{\mu_{t,i}}/(2K_i)$,

$$R_t^{(i)}(\alpha_i) - R_t^{(i)}(\alpha_{i,\star}) \ge \frac{\mu_{t,i}}{2}|\alpha_i - \alpha_{i,\star}|^2.$$

The corresponding local lower isometry argument in Lemma 2 transfers this curvature to the empirical profiled error and gives, uniformly over $\alpha_i \in \mathcal{G}_i$,

$$\underline{\mathcal{Q}}_{t,i}(\alpha_i) \ge \frac{t\mu_{t,i}}{8}|\alpha_i - \alpha_{i,\star}|^2.$$

Therefore, once $\hat{\alpha}_i \in \mathcal{G}_i$ for every i,

$$\mathcal{Q}_{t,i}\left(\hat{\alpha}_i, \hat{a}_i - a_{i,\star}\right) \geq \underline{\mathcal{Q}}_{t,i}(\hat{\alpha}_i) \geq \frac{t\mu_{t,i}}{8}|\hat{\alpha}_i - \alpha_{i,\star}|^2.$$

Summing over the rows yields

$$\mathcal{E}_t \geq \frac{t}{8}\sum_{i=1}^{n}\mu_{t,i}|\hat{\alpha}_i - \alpha_{i,\star}|^2 \geq \frac{t\mu_{\min}}{8}\left\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\star\right\|_\infty^2.$$

Combining this lower bound with the in-sample upper bound and choosing the tunable coefficient sufficiently small allows the quadratic estimation-error term on the upper-bound side to be absorbed. This gives

$$\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_\star\|_\infty^2 \lesssim \epsilon_{\max}^2 + \mathcal{O}(t^{-1}).$$

Common notation for Lemmas 10–13. For each row $i \in [n]$ and candidate α_i, let

$$h_i(\alpha_i) := \alpha_i - \alpha_{i,\star}, \tag{35}$$

$$b_s^{(i)}(\alpha_i) := \left(\Delta^{\alpha_i} - \Delta^{\alpha_{i,\star}}\right)x_{s+1}^{(i)}, \tag{36}$$

$$g_s^{(i)} := \sum_{j\geq 1} d_{i,j}x_{s+1-j}^{(i)}, \qquad d_{i,j} := \partial_\alpha\psi(\alpha_{i,\star}, j), \tag{37}$$

$$\rho_s^{(i)}(\alpha_i) := b_s^{(i)}(\alpha_i) - h_i(\alpha_i)g_s^{(i)}. \tag{38}$$

Recall the unprofiled and profiled noiseless errors

$$\mathcal{Q}_{t,i}(\alpha_i, v) := \sum_{s=0}^{t-1}\left|b_s^{(i)}(\alpha_i) - vx_s\right|^2, \tag{39}$$

$$\underline{\mathcal{Q}}_{t,i}(\alpha_i) := \inf_{v\in\mathbb{R}^{1\times n}}\mathcal{Q}_{t,i}(\alpha_i, v). \tag{40}$$

Define the empirical and population profiled derivative curvatures using the same profiling operation:

$$\hat{\mu}_{t,i} := \inf_{v\in\mathbb{R}^{1\times n}}\frac{1}{t}\sum_{s=0}^{t-1}\left(g_s^{(i)} - vx_s\right)^2, \tag{41}$$

$$\mu_{t,i} := \inf_{v\in\mathbb{R}^{1\times n}}\frac{1}{t}\sum_{s=0}^{t-1}\mathbb{E}\left(g_s^{(i)} - vx_s\right)^2. \tag{42}$$

Define

$$r_{i,\mathrm{sep}} := \frac{\sqrt{\mu_{i,\mathrm{lb}}}}{2K_i}. \tag{43}$$

Lemma 10 (Row-wise lower isometry)**.** *Suppose that the following two events hold simultaneously:*

$$\hat{\mu}_{t,i} \geq \frac{1}{2}\mu_{t,i}, \qquad i \in [n], \tag{E1}$$

$$\frac{1}{t}\sum_{s=0}^{t-1}|\rho_s^{(i)}(\alpha_i)|^2 \leq \frac{K_i^2}{2}|\alpha_i - \alpha_{i,\star}|^4, \qquad i \in [n], \quad \alpha_i \in \mathcal{A}_{\epsilon,i} \cap [\underline{\alpha}_{i,\mathrm{loc}}, 1]. \tag{E2}$$

Then, simultaneously for every $i \in [n]$ and every $\alpha_i \in \mathcal{G}_i$,

$$\underline{\mathcal{Q}}_{t,i}(\alpha_i) \geq \frac{t\mu_{t,i}}{8}|\alpha_i - \alpha_{i,\star}|^2. \tag{44}$$

Consequently, on any event on which $\hat{\alpha}_i \in \mathcal{G}_i$,

$$\mathcal{Q}_{t,i}\left(\hat{\alpha}_i, \hat{a}_i - a_{i,\star}\right) \geq \frac{t\mu_{t,i}}{8}|\hat{\alpha}_i - \alpha_{i,\star}|^2. \tag{45}$$

Proof. Fix i and $\alpha_i \in \mathcal{G}_i$, and write $h_i := \alpha_i - \alpha_{i,\star}$. Let

$$X_t = [x_0, \ldots, x_{t-1}], \qquad P_X := X_t^\top(X_tX_t^\top)^\dagger X_t,$$

where † denotes the Moore–Penrose pseudoinverse. Define the corresponding time-stacked vectors

$$b_i(\alpha_i) := \big(b_0^{(i)}(\alpha_i), \dots, b_{t-1}^{(i)}(\alpha_i)\big)^\top,$$
$$g_i := \big(g_0^{(i)}, \dots, g_{t-1}^{(i)}\big)^\top,$$
$$\rho_i(\alpha_i) := \big(\rho_0^{(i)}(\alpha_i), \dots, \rho_{t-1}^{(i)}(\alpha_i)\big)^\top.$$

Then P_X is the orthogonal projection onto the row space of X_t, and profiling gives

$$\underline{\mathcal{Q}}_{t,i}(\alpha_i) = \|(I - P_X)b_i(\alpha_i)\|_2^2.$$

Since $b_i(\alpha_i) = h_i g_i + \rho_i(\alpha_i)$, the reverse triangle inequality and $\|I - P_X\|_{op} \le 1$ imply

$$\sqrt{\frac{\underline{\mathcal{Q}}_{t,i}(\alpha_i)}{t}} \ge |h_i|\sqrt{\hat{\mu}_{t,i}} - \left(\frac{1}{t}\sum_{s=0}^{t-1}|\rho_s^{(i)}(\alpha_i)|^2\right)^{1/2}.$$

On (E1)–(E2),

$$\sqrt{\frac{\underline{\mathcal{Q}}_{t,i}(\alpha_i)}{t}} \ge |h_i|\sqrt{\frac{\mu_{t,i}}{2}} - \frac{K_i}{\sqrt{2}}|h_i|^2.$$

Because $\alpha_i \in \mathcal{G}_i$, $K_i|h_i| \le \sqrt{\mu_{t,i}}/2$, and hence

$$\sqrt{\frac{\underline{\mathcal{Q}}_{t,i}(\alpha_i)}{t}} \ge \frac{\sqrt{\mu_{t,i}}}{2\sqrt{2}}|h_i|.$$

Squaring proves (44). Finally,

$$\mathcal{Q}_{t,i}\big(\hat{\alpha}_i, \hat{a}_i - a_{i,\star}\big) \ge \underline{\mathcal{Q}}_{t,i}(\hat{\alpha}_i),$$

which proves (45) whenever $\hat{\alpha}_i \in \mathcal{G}_i$. □

Lemma 11 (Relative concentration of the empirical derivative curvature)**.** *Suppose Assumption 1 holds. Let*

$$D_i := \sum_{j\ge1}|d_{i,j}|, \qquad D_{2,i} := \left(\sum_{j\ge1} d_{i,j}^2\right)^{1/2}, \qquad \mathcal{L}_x := \widetilde{C}_x\left(1 + \frac{2}{\alpha_{\min}}\right).$$

Define

$$z_s^{(i)} := \begin{bmatrix} x_s \\ g_s^{(i)} \end{bmatrix}, \qquad \hat{\Sigma}_{z,i} := \frac{1}{t}\sum_{s=0}^{t-1} z_s^{(i)} z_s^{(i)\top}, \qquad \Sigma_{z,i} := \mathbb{E}\hat{\Sigma}_{z,i}.$$

Set

$$\kappa_{i,\star} := d_{i,2}^2 + d_{i,3}^2,$$
$$C_{\mathrm{der}} := \max\Big\{2\|G_1\|_{\mathrm{op}}^2 + 3\big(\|G_2\|_{\mathrm{op}} + \|G_1\|_{\mathrm{op}}^2\big)^2,\ 2 + 3\|G_1\|_{\mathrm{op}}^2,\ 3\Big\},$$

and

$$\mu_{i,\mathrm{lb}} := \frac{\sigma^2\kappa_{i,\star}}{2C_{\mathrm{der}}}.$$

For each i, define

$$\overline{\mathfrak{R}}_i := \mathcal{L}_x\left[\frac{\sqrt{2}}{\sigma} + \frac{D_i + \frac{\sqrt{2\mathcal{L}_x}}{\sigma}D_{2,i}}{\sqrt{\mu_{i,\mathrm{lb}}}}\right]^2,$$

and let

$$q_1(\delta) := (n+1)\log 9 + \log\frac{4n}{\delta}.$$

Then, for $t \ge 6$, $\Sigma_{z,i} \succ 0$ and $\mu_{t,i} \ge \mu_{i,\mathrm{lb}} > 0$. Moreover, with probability at least $1 - \delta/2$, simultaneously for all $i \in [n]$,

$$\left\|\Sigma_{z,i}^{-1/2}\big(\hat{\Sigma}_{z,i} - \Sigma_{z,i}\big)\Sigma_{z,i}^{-1/2}\right\|_{\mathrm{op}} \le 4\left(\sqrt{\frac{\overline{\mathfrak{R}}_i q_1(\delta)}{t}} + \frac{\overline{\mathfrak{R}}_i q_1(\delta)}{t}\right).$$

Consequently, if

$$t \geq t_{\mathrm{E1}}(\delta) := 256 q_1(\delta) \max_{i \in [n]} \overline{\mathfrak{R}}_i,$$

then, with probability at least $1 - \delta/2$*,*

$$\hat{\Sigma}_{z,i} \succeq \frac{1}{2} \Sigma_{z,i}, \qquad i \in [n],$$

and hence

$$\hat{\mu}_{t,i} \geq \frac{1}{2} \mu_{t,i}, \qquad i \in [n].$$

Proof. For

$$h_{i,\ell}^\top := \sum_{j=1}^{\ell} d_{i,j} e_i^\top G_{\ell - j},$$

independence of the innovations gives

$$\frac{1}{t} \sum_{s=0}^{t-1} \mathbb{E}\big(g_s^{(i)} - v x_s\big)^2 = \frac{\sigma^2}{t} \sum_{\ell=1}^{t-1} (t - \ell) \| h_{i,\ell}^\top - v G_{\ell-1} \|_2^2.$$

For $t \geq 6$, retaining only $\ell = 1, 2, 3$ yields

$$\mu_{t,i} \geq \frac{\sigma^2}{2} \inf_v \sum_{\ell=1}^{3} \| h_{i,\ell}^\top - v G_{\ell-1} \|_2^2.$$

If

$$r = d_{i,1} e_i^\top - v,$$

$$u_2 = r G_1 + d_{i,2} e_i^\top, \qquad u_3 = r G_2 + d_{i,2} e_i^\top G_1 + d_{i,3} e_i^\top,$$

then

$$\kappa_{i,\star} \leq C_{\mathrm{der}} \left(\|r\|_2^2 + \|u_2\|_2^2 + \|u_3\|_2^2 \right).$$

Therefore,

$$\mu_{t,i} \geq \frac{\sigma^2 \kappa_{i,\star}}{2 C_{\mathrm{der}}} = \mu_{i,\mathrm{lb}}.$$

Next write

$$\Sigma_{z,i} = \begin{bmatrix} \Sigma_{x,t} & c_{t,i} \\ c_{t,i}^\top & q_{t,i} \end{bmatrix},$$

where

$$\Sigma_{x,t} := \frac{1}{t} \sum_{s=0}^{t-1} \mathbb{E}[x_s x_s^\top], \qquad q_{t,i} := \frac{1}{t} \sum_{s=0}^{t-1} \mathbb{E}\big[(g_s^{(i)})^2\big].$$

Since x_s contains the fresh innovation η_{s-1} for $s \geq 1$,

$$\Sigma_{x,t} \succeq \frac{\sigma^2}{2} I_n.$$

The Schur complement of $\Sigma_{x,t}$ in $\Sigma_{z,i}$ is $\mu_{t,i} > 0$. Hence $\Sigma_{z,i} \succ 0$.

We now derive a deterministic bound on the temporal dependence of the whitened process. Let

$$a_{t,i} := \Sigma_{x,t}^{-1} c_{t,i}.$$

The Schur-complement identity gives

$$q_{t,i} = \mu_{t,i} + a_{t,i}^\top \Sigma_{x,t} a_{t,i}.$$

In particular,

$$a_{t,i}^\top \Sigma_{x,t} a_{t,i} \leq q_{t,i}.$$

We first bound $q_{t,i}$. The covariance decay in Lemma 14 gives, for every s,

$$\mathbb{E}\big[(g_s^{(i)})^2\big] \le \widetilde{C}_x \sum_{j,k\ge 1} |d_{i,j}||d_{i,k}|\big(|j-k|+1\big)^{-(1+\alpha_{\min})}.$$

Let

$$\gamma_k := (|k|+1)^{-(1+\alpha_{\min})}, \qquad k \in \mathbb{Z}.$$

Young's convolution inequality gives

$$\sum_{j,k\ge 1} |d_{i,j}||d_{i,k}|\gamma_{j-k} \le \|\gamma\|_{\ell_1(\mathbb{Z})} \sum_{j\ge 1} d_{i,j}^2.$$

Moreover,

$$\|\gamma\|_{\ell_1(\mathbb{Z})} \le 1 + \frac{2}{\alpha_{\min}}.$$

Consequently,

$$q_{t,i} \le \mathcal{L}_x D_{2,i}^2.$$

Since $\Sigma_{x,t} \succeq \sigma^2 I_n/2$, it follows that

$$\|a_{t,i}\|_2^2 \le \frac{2q_{t,i}}{\sigma^2} \le \frac{2\mathcal{L}_x D_{2,i}^2}{\sigma^2},$$

and therefore

$$\|a_{t,i}\|_2 \le \frac{\sqrt{2\mathcal{L}_x}}{\sigma} D_{2,i}.$$

Fix i and a unit vector $u \in \mathbb{R}^{n+1}$, and write

$$w := \Sigma_{z,i}^{-1/2} u = \begin{bmatrix} \beta \\ b \end{bmatrix}.$$

Since $w^\top \Sigma_{z,i} w = 1$, the factorization

$$\Sigma_{z,i} = \begin{bmatrix} I & a_{t,i} \\ 0 & 1 \end{bmatrix}^\top \begin{bmatrix} \Sigma_{x,t} & 0 \\ 0 & \mu_{t,i} \end{bmatrix} \begin{bmatrix} I & a_{t,i} \\ 0 & 1 \end{bmatrix}$$

implies

$$1 = (\beta + b a_{t,i})^\top \Sigma_{x,t} (\beta + b a_{t,i}) + \mu_{t,i} b^2.$$

Therefore,

$$\|\beta + b a_{t,i}\|_2 \le \frac{\sqrt{2}}{\sigma}, \qquad |b| \le \frac{1}{\sqrt{\mu_{i,\mathrm{lb}}}}.$$

It follows that

$$\begin{aligned} \|\beta\|_2 + D_i|b| &\le \|\beta + b a_{t,i}\|_2 + \big(\|a_{t,i}\|_2 + D_i\big)|b| \\ &\le \frac{\sqrt{2}}{\sigma} + \frac{D_i + \frac{\sqrt{2\mathcal{L}_x}}{\sigma} D_{2,i}}{\sqrt{\mu_{i,\mathrm{lb}}}}. \end{aligned}$$

Now define the scalar whitened process

$$r_s^{(i,u)} := u^\top \Sigma_{z,i}^{-1/2} z_s^{(i)} = \beta^\top x_s + b g_s^{(i)}.$$

It is a linear filter of x_s whose coefficient ℓ_1 norm is at most

$$F_i := \frac{\sqrt{2}}{\sigma} + \frac{D_i + \frac{\sqrt{2\mathcal{L}_x}}{\sigma} D_{2,i}}{\sqrt{\mu_{i,\mathrm{lb}}}}.$$

The covariance decay convolution bound in Lemma 14 therefore gives

$$\sup_p \sum_{k\in\mathbb{Z}} \Big|\mathrm{Cov}\big(r_p^{(i,u)}, r_{p+k}^{(i,u)}\big)\Big| \le F_i^2 \mathcal{L}_x = \overline{\mathfrak{R}}_i.$$

If

$$\mathcal{T}_{i,u} := \operatorname{Cov}\left(r_0^{(i,u)}, \dots, r_{t-1}^{(i,u)}\right),$$

then

$$\|\mathcal{T}_{i,u}\|_{\mathrm{op}} \le \overline{\mathfrak{R}}_i.$$

Furthermore,

$$\begin{aligned}\operatorname{tr}(\mathcal{T}_{i,u}) &= \sum_{s=0}^{t-1} \mathbb{E}\left[\left(r_s^{(i,u)}\right)^2\right] \\ &= t\, u^\top \Sigma_{z,i}^{-1/2} \Sigma_{z,i} \Sigma_{z,i}^{-1/2} u \\ &= t.\end{aligned}$$

Since $\mathcal{T}_{i,u} \succeq 0$,

$$\|\mathcal{T}_{i,u}\|_F^2 \le \|\mathcal{T}_{i,u}\|_{\mathrm{op}} \operatorname{tr}(\mathcal{T}_{i,u}) \le t\overline{\mathfrak{R}}_i.$$

Let

$$r^{(i,u)} := \left(r_0^{(i,u)}, \dots, r_{t-1}^{(i,u)}\right)^\top.$$

The Gaussian quadratic-form inequality [15] gives, for every $x > 0$,

$$\left|\frac{1}{t}\|r^{(i,u)}\|_2^2 - 1\right| \le 2\sqrt{\frac{\overline{\mathfrak{R}}_i x}{t}} + 2\frac{\overline{\mathfrak{R}}_i x}{t}$$

with probability at least $1 - 2e^{-x}$.

Equivalently,

$$\left|u^\top \Sigma_{z,i}^{-1/2}\big(\hat{\Sigma}_{z,i} - \Sigma_{z,i}\big)\Sigma_{z,i}^{-1/2} u\right| \le 2\sqrt{\frac{\overline{\mathfrak{R}}_i x}{t}} + 2\frac{\overline{\mathfrak{R}}_i x}{t}.$$

Apply this inequality to a $1/4$-net of the unit sphere in $\mathbb{R}^{n+1}$, whose cardinality is at most 9^{n+1}, and take a union bound over $i \in [n]$. With

$$x = q_1(\delta) = (n+1)\log 9 + \log\frac{4n}{\delta},$$

the total failure probability is at most

$$2n9^{n+1}e^{-q_1(\delta)} = \frac{\delta}{2}.$$

The standard $1/4$-net bound for symmetric matrices then yields

$$\left\|\Sigma_{z,i}^{-1/2}\big(\hat{\Sigma}_{z,i} - \Sigma_{z,i}\big)\Sigma_{z,i}^{-1/2}\right\|_{\mathrm{op}} \le 4\left(\sqrt{\frac{\overline{\mathfrak{R}}_i q_1(\delta)}{t}} + \frac{\overline{\mathfrak{R}}_i q_1(\delta)}{t}\right)$$

simultaneously for all i.

If

$$t \ge 256 q_1(\delta) \max_i \overline{\mathfrak{R}}_i,$$

then

$$\sqrt{\frac{\overline{\mathfrak{R}}_i q_1(\delta)}{t}} \le \frac{1}{16}, \qquad \frac{\overline{\mathfrak{R}}_i q_1(\delta)}{t} \le \frac{1}{256},$$

so the preceding right-hand side is at most

$$4\left(\frac{1}{16} + \frac{1}{256}\right) = \frac{17}{64} < \frac{1}{2}.$$

Hence

$$\Sigma_{z,i}^{-1/2}\hat{\Sigma}_{z,i}\Sigma_{z,i}^{-1/2} \succeq \frac{1}{2}I,$$

and therefore

$$\hat{\Sigma}_{z,i} \succeq \frac{1}{2}\Sigma_{z,i}.$$

Finally, for

$$w(v) := (-v^\top, 1)^\top,$$

we have

$$\hat{\mu}_{t,i} = \inf_v w(v)^\top \hat{\Sigma}_{z,i} w(v)$$

and

$$\mu_{t,i} = \inf_v w(v)^\top \Sigma_{z,i} w(v).$$

Thus

$$\hat{\mu}_{t,i} \geq \frac{1}{2}\mu_{t,i},$$

which proves event (E1). □

Lemma 12 (Global profiled lower isometry)**.** *Suppose Assumption 1 holds. Define*

$$\gamma := \min_{i\in[n]} \inf_{\substack{\alpha_i \in \mathcal{A}_{\epsilon,i} \\ |h_i(\alpha_i)|>r_{i,\mathrm{sep}}}} \left\{ R_t^{(i)}(\alpha_i) - R_t^{(i)}(\alpha_i^\circ) \right\},$$

and suppose that $\gamma > 0$. *Let* $F_{\mathrm{G}} := \frac{\sqrt{2}}{\sigma} + \frac{S_1}{\sqrt{\gamma}}\left(1 + \frac{\sqrt{2L_x}}{\sigma}\right)$, *and*

$$q_{\mathrm{G}}(\delta) := (n+1)\log 9 + \log \frac{2\sum_{i=1}^n M_i}{\delta}.$$

If $t \geq \max\left\{2,\, 256 L_x F_{\mathrm{G}}^2 q_{\mathrm{G}}(\delta)\right\}$, *then with probability at least* $1-\delta$, *simultaneously for every* $i \in [n]$ *and every* $\alpha_i \in \mathcal{A}_{\epsilon,i}$ *satisfying* $|h_i(\alpha_i)| > r_{i,\mathrm{sep}}$,

$$\underline{\mathcal{Q}}_{t,i}(\alpha_i) \geq \frac{t}{2}\left[R_t^{(i)}(\alpha_i) - R_t^{(i)}(\alpha_{i,\star})\right] \geq \frac{t\gamma}{2}.$$

Proof. Fix $i \in [n]$ and $\alpha_i \in \mathcal{A}_{\epsilon,i}$ satisfying $|h_i(\alpha_i)| > r_{i,\mathrm{sep}}$. Define $z_s^{(i,\alpha_i)} := \begin{bmatrix} x_s \\ b_s^{(i)}(\alpha_i) \end{bmatrix}$, $\hat{\Sigma}_{i,\alpha_i} := \frac{1}{t}\sum_{s=0}^{t-1} z_s^{(i,\alpha_i)} z_s^{(i,\alpha_i)\top}$, and $\Sigma_{i,\alpha_i} := \mathbb{E}\hat{\Sigma}_{i,\alpha_i}$. For $w(v) := \begin{bmatrix} -v^\top \\ 1 \end{bmatrix}$, we have

$$w(v)^\top z_s^{(i,\alpha_i)} = b_s^{(i)}(\alpha_i) - v x_s.$$

Hence, by the definition of the profiled noiseless error,

$$\frac{1}{t}\underline{\mathcal{Q}}_{t,i}(\alpha_i) = \inf_v w(v)^\top \hat{\Sigma}_{i,\alpha_i} w(v). \tag{46}$$

Write $\Sigma_{i,\alpha_i} = \begin{bmatrix} \Sigma_{x,t} & c_{i,\alpha_i} \\ c_{i,\alpha_i}^\top & q_{i,\alpha_i} \end{bmatrix}$, where $c_{i,\alpha_i} = \frac{1}{t}\sum_{s=0}^{t-1}\mathbb{E}\left[x_s b_s^{(i)}(\alpha_i)\right]$, and $q_{i,\alpha_i} = \frac{1}{t}\sum_{s=0}^{t-1}\mathbb{E}\left[|b_s^{(i)}(\alpha_i)|^2\right]$. The population-risk identity gives

$$R_t^{(i)}(\alpha_i) - R_t^{(i)}(\alpha_{i,\star}) = \inf_v \frac{1}{t}\sum_{s=0}^{t-1}\mathbb{E}\left[|b_s^{(i)}(\alpha_i) - v x_s|^2\right].$$

Expanding the square,

$$\frac{1}{t}\sum_{s=0}^{t-1}\mathbb{E}|b_s^{(i)}(\alpha_i) - v x_s|^2 = q_{i,\alpha_i} - 2 v c_{i,\alpha_i} + v\Sigma_{x,t}v^\top.$$

For $t \geq 2$, the innovation gives

$$\Sigma_{x,t} \succeq \frac{\sigma^2}{2} I_n. \tag{47}$$

Thus $\Sigma_{x,t} \succ 0$, and minimizing the preceding quadratic function over v yields $v^\star = c_{i,\alpha_i}^\top \Sigma_{x,t}^{-1}$. Consequently,

$$R_t^{(i)}(\alpha_i) - R_t^{(i)}(\alpha_{i,\star}) = q_{i,\alpha_i} - c_{i,\alpha_i}^\top \Sigma_{x,t}^{-1} c_{i,\alpha_i}. \tag{48}$$

Equivalently,

$$R_t^{(i)}(\alpha_i) - R_t^{(i)}(\alpha_{i,\star}) = \inf_v w(v)^\top \Sigma_{i,\alpha_i} w(v). \tag{49}$$

Since

$$R_t^{(i)}(\alpha_i^\circ) - R_t^{(i)}(\alpha_{i,\star}) \geq 0,$$

the definition of γ implies

$$R_t^{(i)}(\alpha_i) - R_t^{(i)}(\alpha_{i,\star}) \geq \gamma. \tag{50}$$

Hence the Schur complement in (48) is strictly positive and $\Sigma_{i,\alpha_i} \succ 0$.

Let

$$\Delta\psi_j := \psi(\alpha_i, j) - \psi(\alpha_{i,\star}, j).$$

Then

$$b_s^{(i)}(\alpha_i) = \sum_{j \geq 1} \Delta\psi_j x_{s+1-j}^{(i)}.$$

By the fundamental theorem of calculus and the definition of S_1,

$$\sum_{j \geq 1} |\Delta\psi_j| \leq S_1 |h_i(\alpha_i)|. \tag{51}$$

Therefore,

$$\left(\sum_{j \geq 1} |\Delta\psi_j|^2 \right)^{1/2} \leq S_1 |h_i(\alpha_i)|. \tag{52}$$

Let $\gamma_k := (|k|+1)^{-(1+\alpha_{\min})}$, where $k \in \mathbb{Z}$. The covariance-decay bound in Lemma 14 gives

$$\left| \mathrm{Cov}\left(x_p^{(i)}, x_q^{(i)} \right) \right| \leq \widetilde{C}_x \gamma_{p-q}.$$

Hence, for every s,

$$\mathbb{E}|b_s^{(i)}(\alpha_i)|^2 \leq \widetilde{C}_x \sum_{j,k \geq 1} |\Delta\psi_j||\Delta\psi_k|\gamma_{j-k} \leq \widetilde{C}_x \|\gamma\|_{\ell_1(\mathbb{Z})} \sum_{j \geq 1} |\Delta\psi_j|^2,$$

where the second inequality follows from Young's convolution inequality. Since

$$\|\gamma\|_{\ell_1(\mathbb{Z})} \leq 1 + \frac{2}{\alpha_{\min}},$$

the definition of L_x and (52) imply

$$q_{i,\alpha_i} \leq L_x S_1^2 |h_i(\alpha_i)|^2. \tag{53}$$

Let $m_{i,\alpha_i} := \Sigma_{x,t}^{-1} c_{i,\alpha_i}$. By (48),

$$q_{i,\alpha_i} = \left[R_t^{(i)}(\alpha_i) - R_t^{(i)}(\alpha_{i,\star}) \right] + m_{i,\alpha_i}^\top \Sigma_{x,t} m_{i,\alpha_i}.$$

Thus

$$m_{i,\alpha_i}^\top \Sigma_{x,t} m_{i,\alpha_i} \leq q_{i,\alpha_i}.$$

Combining this with (47) and (53) yields

$$\|m_{i,\alpha_i}\|_2 \leq \frac{\sqrt{2L_x}}{\sigma} S_1 |h_i(\alpha_i)|. \tag{54}$$

Fix a unit vector $u \in \mathbb{R}^{n+1}$ and write $\Sigma_{i,\alpha_i}^{-1/2} u = \begin{bmatrix} \beta \\ b \end{bmatrix}$. The block factorization

$$\Sigma_{i,\alpha_i} = \begin{bmatrix} I & m_{i,\alpha_i} \\ 0 & 1 \end{bmatrix}^\top \begin{bmatrix} \Sigma_{x,t} & 0 \\ 0 & R_t^{(i)}(\alpha_i) - R_t^{(i)}(\alpha_{i,\star}) \end{bmatrix} \begin{bmatrix} I & m_{i,\alpha_i} \\ 0 & 1 \end{bmatrix}$$

and $u^\top u = 1$ give

$$1 = (\beta + b m_{i,\alpha_i})^\top \Sigma_{x,t} (\beta + b m_{i,\alpha_i}) + \left[R_t^{(i)}(\alpha_i) - R_t^{(i)}(\alpha_{i,\star}) \right] b^2.$$

Hence, by (47) and (50),

$$\|\beta + b m_{i,\alpha_i}\|_2 \le \frac{\sqrt{2}}{\sigma}, \qquad |b| \le \frac{1}{\sqrt{\gamma}}. \tag{55}$$

Define the scalar whitened process

$$r_s := u^\top \Sigma_{i,\alpha_i}^{-1/2} z_s^{(i,\alpha_i)} = \beta^\top x_s + b\, b_s^{(i)}(\alpha_i).$$

It is a linear filter of the state process. By (51), (54), and (55), the sum of the Euclidean norms of its filter coefficients is at most

$$\begin{aligned} \|\beta\|_2 + |b| \sum_{j \ge 1} |\Delta\psi_j| &\le \|\beta + b m_{i,\alpha_i}\|_2 + |b| \left(\|m_{i,\alpha_i}\|_2 + \sum_{j \ge 1} |\Delta\psi_j| \right) \\ &\le \frac{\sqrt{2}}{\sigma} + \frac{S_1 |h_i(\alpha_i)|}{\sqrt{\gamma}} \left(1 + \frac{\sqrt{2L_x}}{\sigma} \right). \end{aligned}$$

Since $\alpha_i, \alpha_{i,\star} \in (0,1]$, $|h_i(\alpha_i)| \le 1$, and therefore this coefficient norm is at most F_{G}.

Writing the filter representation as

$$r_s = \sum_{\ell \ge 0} f_\ell^\top x_{s-\ell}, \qquad \sum_{\ell \ge 0} \|f_\ell\|_2 \le F_{\mathrm{G}},$$

we obtain

$$\sum_{k \in \mathbb{Z}} |\operatorname{Cov}(r_p, r_{p+k})| \le \widetilde{C}_x \sum_{\ell,m \ge 0} \|f_\ell\|_2 \|f_m\|_2 \sum_{k \in \mathbb{Z}} \gamma_{k+\ell-m} \le L_x F_{\mathrm{G}}^2. \tag{56}$$

Let $T_{i,\alpha_i,u}$ denote the covariance matrix of $(r_0, \dots, r_{t-1})^\top$. The preceding covariance-sum bound (56) yields

$$\|T_{i,\alpha_i,u}\|_{\mathrm{op}} \le L_x F_{\mathrm{G}}^2. \tag{57}$$

Moreover,

$$\operatorname{tr}(T_{i,\alpha_i,u}) = \sum_{s=0}^{t-1} \mathbb{E}[r_s^2] = t\, u^\top \Sigma_{i,\alpha_i}^{-1/2} \Sigma_{i,\alpha_i} \Sigma_{i,\alpha_i}^{-1/2} u = t.$$

Since $T_{i,\alpha_i,u} \succeq 0$,

$$\|T_{i,\alpha_i,u}\|_F^2 \le \|T_{i,\alpha_i,u}\|_{\mathrm{op}} \operatorname{tr}(T_{i,\alpha_i,u}) \le t L_x F_{\mathrm{G}}^2. \tag{58}$$

The vector $(r_0, \dots, r_{t-1})^\top$ is jointly Gaussian. Thus, for every $x > 0$, the Gaussian quadratic-form inequality with (57) and (58) gives

$$\left| \frac{1}{t} \sum_{s=0}^{t-1} r_s^2 - 1 \right| \le 2\sqrt{\frac{L_x F_{\mathrm{G}}^2 x}{t}} + 2\frac{L_x F_{\mathrm{G}}^2 x}{t}$$

with probability at least $1 - 2e^{-x}$. Equivalently,

$$\left| u^\top \Sigma_{i,\alpha_i}^{-1/2} \left(\hat{\Sigma}_{i,\alpha_i} - \Sigma_{i,\alpha_i} \right) \Sigma_{i,\alpha_i}^{-1/2} u \right| \le 2\sqrt{\frac{L_x F_{\mathrm{G}}^2 x}{t}} + 2\frac{L_x F_{\mathrm{G}}^2 x}{t}.$$

Apply this inequality to a $1/4$-net of the unit sphere in $\mathbb{R}^{n+1}$, whose cardinality is at most 9^{n+1}, and take a union bound over all grid points. Their total number is at most $\sum_{i=1}^n M_i$. Taking $x = q_{\mathrm{G}}(\delta)$ gives, with probability at least $1 - \delta$, simultaneously for all i and $\alpha_i \in \mathcal{A}_{\epsilon,i}$ satisfying $|h_i(\alpha_i)| > r_{\mathrm{i,sep}}$,

$$\left\| \Sigma_{i,\alpha_i}^{-1/2} \left(\hat{\Sigma}_{i,\alpha_i} - \Sigma_{i,\alpha_i} \right) \Sigma_{i,\alpha_i}^{-1/2} \right\|_{\mathrm{op}} \le 4 \left(\sqrt{\frac{L_x F_{\mathrm{G}}^2 q_{\mathrm{G}}(\delta)}{t}} + \frac{L_x F_{\mathrm{G}}^2 q_{\mathrm{G}}(\delta)}{t} \right). \tag{59}$$

If

$$t \geq 256 L_x F_{\mathrm{G}}^2 q_{\mathrm{G}}(\delta),$$

then the right-hand side of (59) is at most

$$4\left(\frac{1}{16}+\frac{1}{256}\right) = \frac{17}{64} < \frac{1}{2}.$$

Therefore

$$\hat{\Sigma}_{i,\alpha_i} \succeq \frac{1}{2}\Sigma_{i,\alpha_i}. \tag{60}$$

Finally, (60) implies, for every v,

$$w(v)^\top \hat{\Sigma}_{i,\alpha_i} w(v) \geq \frac{1}{2} w(v)^\top \Sigma_{i,\alpha_i} w(v).$$

Taking the infimum over v and using (46) and (49),

$$\frac{1}{t}\underline{\mathcal{Q}}_{t,i}(\alpha_i) \geq \frac{1}{2}\left[R_t^{(i)}(\alpha_i) - R_t^{(i)}(\alpha_{i,\star})\right].$$

Thus

$$\underline{\mathcal{Q}}_{t,i}(\alpha_i) \geq \frac{t}{2}\left[R_t^{(i)}(\alpha_i) - R_t^{(i)}(\alpha_{i,\star})\right].$$

Finally, (50) gives

$$\underline{\mathcal{Q}}_{t,i}(\alpha_i) \geq \frac{t\gamma}{2}.$$

□

Lemma 13 (Concentration of the Taylor remainder)**.** *Suppose Assumption 1 holds. Let*

$$\mathcal{L}_{\rho,i} := \frac{1}{4}\left(\frac{9e}{4}\,\mathfrak{Z}_i\right)^2 \mathcal{L}_x.$$

Define

$$M_{\mathrm{loc}} := \sum_{i=1}^{n} \left|\mathcal{A}_{\epsilon,i} \cap [\underline{\alpha}_{i,\mathrm{loc}}, 1]\right|, \qquad q_2(\delta) := \log\frac{2M_{\mathrm{loc}}}{\delta}.$$

Then, with probability at least $1-\delta/2$, simultaneously for all $i \in [n]$ and all $\alpha_i \in \mathcal{A}_{\epsilon,i} \cap [\underline{\alpha}_{i,\mathrm{loc}}, 1]$,

$$\frac{1}{t}\sum_{s=0}^{t-1} |\rho_s^{(i)}(\alpha_i)|^2 \leq |h_i(\alpha_i)|^4 \left[\frac{K_i^2}{4} + 2\mathcal{L}_{\rho,i}\left(\sqrt{\frac{q_2(\delta)}{t}} + \frac{q_2(\delta)}{t}\right)\right]. \tag{61}$$

Consequently, if

$$t \geq t_{\mathrm{E2}}(\delta) := q_2(\delta)\max\left\{\frac{16\mathcal{L}_x^2}{\sigma^4 \widetilde{C}_G^4 \zeta(2)^2}, \frac{4\mathcal{L}_x}{\sigma^2 \widetilde{C}_G^2 \zeta(2)}\right\}, \tag{62}$$

then event (E2) *holds simultaneously over all local grid points with probability at least $1-\delta/2$.*

Proof. Fix i and $\alpha_i \in \mathcal{A}_{\epsilon,i} \cap [\underline{\alpha}_{i,\mathrm{loc}}, 1]$. For

$$q_{i,j}(\alpha_i) := \psi(\alpha_i, j) - \psi(\alpha_{i,\star}, j) - h_i(\alpha_i)\partial_\alpha \psi(\alpha_{i,\star}, j),$$

the integral Taylor formula gives

$$q_{i,j}(\alpha_i) = \int_{\alpha_{i,\star}}^{\alpha_i} (\alpha_i - u)\partial_u^2 \psi(u, j)\,\mathrm{d}u, \qquad \rho_s^{(i)}(\alpha_i) = \sum_{j\geq 1} q_{i,j}(\alpha_i) x_{s+1-j}^{(i)}.$$

Therefore, by (33),

$$\sum_{j\geq 1} |q_{i,j}(\alpha_i)| \leq \frac{9e}{8}\,\mathfrak{Z}_i |h_i(\alpha_i)|^2. \tag{63}$$

Using (63) and the same convolution calculation as in Lemma 11,

$$\sup_p \sum_{k\in\mathbb{Z}} |\operatorname{Cov}(\rho_p^{(i)}(\alpha_i), \rho_{p+k}^{(i)}(\alpha_i))| \le \mathcal{L}_{\rho,i}|h_i(\alpha_i)|^4.$$

Hence, for $\rho_i(\alpha_i) = (\rho_0^{(i)}(\alpha_i), \dots, \rho_{t-1}^{(i)}(\alpha_i))^\top \sim \mathcal{N}(0, \mathcal{R}_{i,\alpha_i})$,

$$\|\mathcal{R}_{i,\alpha_i}\|_{\mathrm{op}} \le \mathcal{L}_{\rho,i}|h_i(\alpha_i)|^4, \qquad \|\mathcal{R}_{i,\alpha_i}\|_F \le \sqrt{t}\,\mathcal{L}_{\rho,i}|h_i(\alpha_i)|^4.$$

The one-sided Gaussian quadratic-form inequality gives, with probability at least $1 - e^{-x}$,

$$\frac{1}{t}\sum_{s=0}^{t-1}|\rho_s^{(i)}(\alpha_i)|^2 \le \frac{1}{t}\sum_{s=0}^{t-1}\mathbb{E}|\rho_s^{(i)}(\alpha_i)|^2 + 2\mathcal{L}_{\rho,i}|h_i(\alpha_i)|^4\left(\sqrt{\frac{x}{t}} + \frac{x}{t}\right).$$

Lemma 9 bounds the expectation term by $K_i^2|h_i(\alpha_i)|^4/4$. Taking $x = q_2(\delta)$ and applying a union bound proves (61). Finally, (62) implies

$$2\mathcal{L}_{\rho,i}\left(\sqrt{\frac{q_2(\delta)}{t}} + \frac{q_2(\delta)}{t}\right) \le \frac{K_i^2}{4},$$

which proves event (E2). □

We now have all the tools needed to prove Theorem 1. We begin by restating Theorem 1 in its full finite sample form for the estimation error of $\boldsymbol{\alpha}_\star$.

Theorem 3 (Full error bound for $\boldsymbol{\alpha}_\star$)**.** *Under Assumption 1, define* $\epsilon_{\max} := \max_{1\le i\le n}\epsilon_i$, *and* $\mu_{\min} := \min_{1\le i\le n}\mu_{t,i}$. *Let*

$$\gamma := \min_{1\le i\le n}\ \inf_{\alpha\in\mathcal{A}_{\epsilon,i},\,|\alpha_i-\alpha_{i,\star}|>\frac{\sqrt{\mu_{i,\mathrm{lb}}}}{2K_i}} \left\{R_t^{(i)}(\alpha_i) - R_t^{(i)}(\alpha_i^\circ)\right\} \tag{64}$$

and suppose $\gamma > 0$. *Fix* $\delta \in (0, 1/2)$ *and choose* $0 < \theta \le \min\left\{\frac{1}{2},\ \frac{\delta\mu_{\min}}{576\sigma^2 nS_1^2\operatorname{tr}(\Gamma_t)}\right\}$. *Define*

$$\mathfrak{B}(\theta,\delta,k) := \frac{18\theta\sigma^2 nS_1^2 t\operatorname{tr}(\Gamma_t)}{\delta} + \frac{27\sigma^2 nS_1^2 t\operatorname{tr}(\Gamma_t)}{\delta}\epsilon_{\max}^2 + \frac{8\sigma^2}{\theta}\sum_{i=1}^n \log\frac{9M_i n}{\delta} + 8\sigma^2\log\frac{9}{\delta} + \frac{36(1+\theta^{-1})C^2\sigma^2 n\operatorname{tr}(\Gamma_t)}{\delta\lambda_{\min}(\Gamma_k)}\Xi_t(\delta,k),$$

where $\Xi_t(\delta,k) := n\log\frac{9n}{\delta} + \log\det\left(\Gamma_t\Gamma_k^{-1}\right)$. *If* $t \gtrsim \frac{1}{\alpha_{\min}^4}\left(n + \log\frac{\sum_{i=1}^n M_i}{\delta}\right)$, *and that, for some integer* k,

$$\frac{t}{k} \ge c\,\Xi_t(\delta,k). \tag{65}$$

Suppose that

$$\mathfrak{B}(\theta,\delta,k) < \frac{t\gamma}{2}. \tag{66}$$

Then, with probability at least $1-\delta$,

$$\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_\star\|_\infty^2 \le \frac{288\sigma^2 nS_1^2\operatorname{tr}(\Gamma_t)}{\delta\mu_{\min}}\epsilon_{\max}^2 + \frac{256}{3t\mu_{\min}\theta}\left[\sigma^2\left(\sum_{i=1}^n\log\frac{9M_i n}{\delta} + \log\frac{9}{\delta}\right) + \frac{9C^2\sigma^2 n\operatorname{tr}(\Gamma_t)}{\delta\lambda_{\min}(\Gamma_k)}\Xi_t(\delta,k)\right]. \tag{67}$$

Consequently, if $\epsilon_{\max} = \mathcal{O}\left(t^{-1/2}\right)$, *then* $\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_\star\|_\infty = \mathcal{O}\left(t^{-1/2}\right)$.

Proof. We intersect three events: the global profiled lower isometry event from Lemma 12, the local lower isometry event from Lemma 10, and the in-sample event from Lemma 8, with failure probability $\delta/3$ allocated to each. By $t \gtrsim \frac{1}{\alpha_{\min}^4}\left(n + \log\frac{\sum_{i=1}^n M_i}{\delta}\right)$, their intersection has probability at least $1-\delta$. We work on this intersection throughout.

Step 1: Global localization. Suppose, toward a contradiction, that for some $i \in [n]$,

$$|\hat{\alpha}_i - \alpha_{i,\star}| > r_{i,\mathrm{sep}}.$$

Since the global profiled lower-isometry event holds simultaneously over all grid points outside the separation neighborhood,

$$\underline{\mathcal{Q}}_{t,i}(\hat{\alpha}_i) \geq \frac{t}{2}\left[R_t^{(i)}(\hat{\alpha}_i) - R_t^{(i)}(\alpha_{i,\star})\right] \geq \frac{t\gamma}{2}.$$

Recall

$$\mathcal{E}_t = \sum_{s=0}^{t-1}\left\|\left(\Delta^{\hat{\boldsymbol{\alpha}}} - \Delta^{\boldsymbol{\alpha}^\star}\right)x_{s+1} - (\hat{A} - A_\star)x_s\right\|_2^2 = \sum_{j=1}^n \mathcal{Q}_{t,j}\big(\hat{\alpha}_j, \hat{a}_j - a_{j,\star}\big).$$

Since every term in the sum is nonnegative and $\underline{\mathcal{Q}}_{t,i}$ is obtained by profiling over the row parameter,

$$\mathcal{E}_t \geq \mathcal{Q}_{t,i}\big(\hat{\alpha}_i, \hat{a}_i - a_{i,\star}\big) \geq \underline{\mathcal{Q}}_{t,i}(\hat{\alpha}_i) \geq \frac{t\gamma}{2}. \tag{68}$$

We next obtain the corresponding upper bound from Lemma 8. Set $\tau = \rho = \theta$ and use failure probability $\delta/3$. The Lemma 3 good event retained in its proof also gives

$$X_t X_t^\top \preceq \frac{9\sigma^2 nt}{\delta}\Gamma_t, \qquad \mathrm{tr}(X_t X_t^\top) \leq \frac{9\sigma^2 nt}{\delta}\,\mathrm{tr}(\Gamma_t). \tag{69}$$

Moreover, since both the true and estimated fractional orders lie in $(0,1]$,

$$\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\star\|_\infty \leq 1.$$

Substituting these two bounds into Lemma 8 yields

$$\mathcal{E}_t \leq \mathfrak{B}(\theta, \delta, k).$$

By $\mathfrak{B}(\theta,\delta,k) < \frac{t\gamma}{2}$,

$$\mathcal{E}_t < \frac{t\gamma}{2},$$

contradicting (68). Therefore,

$$|\hat{\alpha}_i - \alpha_{i,\star}| \leq r_{i,\mathrm{sep}}, \qquad i \in [n]. \tag{70}$$

We now verify that this coarse localization places the estimator in the local sets $\mathcal{G}_i$. Recall that

$$\mathfrak{Z}_i = \sum_{j\geq 1}\frac{(1+\log j)^2}{j^{1+\alpha_{i,\star}/2}} \geq \zeta\Big(1 + \frac{\alpha_{i,\star}}{2}\Big) \geq \frac{2}{\alpha_{i,\star}}.$$

Since $G_0 = I_n$, $\widetilde{C}_G \geq 1$. Moreover, $|d_{i,2}| \leq 1/2$ and $|d_{i,3}| \leq 1/3$, so $\kappa_{i,\star} \leq 13/36$. Hence

$$r_{i,\mathrm{sep}} \leq \frac{\alpha_{i,\star}\sqrt{13}}{108e\sqrt{\zeta(2)}} < \frac{\alpha_{i,\star}}{2}.$$

Together with (70), this implies

$$\hat{\alpha}_i > \frac{\alpha_{i,\star}}{2} = \alpha_{i,\mathrm{loc}}.$$

Since $\hat{\alpha}_i \in \mathcal{A}_{\epsilon,i} \subset (0,1]$,

$$\hat{\alpha}_i \in \mathcal{A}_{\epsilon,i} \cap [\alpha_{i,\mathrm{loc}}, 1].$$

Furthermore, the population curvature lower bound gives

$$\mu_{t,i} \geq \mu_{i,\mathrm{lb}}, \qquad r_{i,\mathrm{sep}} = \frac{\sqrt{\mu_{i,\mathrm{lb}}}}{2K_i}.$$

Hence

$$|\hat{\alpha}_i - \alpha_{i,\star}| \le \frac{\sqrt{\mu_{i,\mathrm{lb}}}}{2K_i},$$

and therefore

$$\hat{\alpha}_i \in \mathcal{G}_i, \qquad i \in [n].$$

Step 2: Local lower isometry. Since $\hat{\alpha}_i \in \mathcal{G}_i$ for every row, Lemma 10 gives

$$\begin{aligned}\mathcal{E}_t &= \sum_{i=1}^{n} \mathcal{Q}_{t,i}\big(\hat{\alpha}_i, \hat{a}_i - a_{i,\star}\big) \\ &\ge \frac{t}{8}\sum_{i=1}^{n} \mu_{t,i}|\hat{\alpha}_i - \alpha_{i,\star}|^2 \\ &\ge \frac{t\mu_{\min}}{8}\,\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\star\|_\infty^2 . \end{aligned} \tag{71}$$

Step 3: In-sample upper bound and absorption. We reuse the same in-sample event from Step 1; no additional failure probability is required. Setting $\tau = \rho = \theta$ in Lemma 8 and substituting (69) gives

$$\begin{aligned}\mathcal{E}_t \le{}& \frac{18\theta\sigma^2 n S_1^2 t\,\mathrm{tr}(\Gamma_t)}{\delta}\,\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\star\|_\infty^2 \\ &+ \frac{27\sigma^2 n S_1^2 t\,\mathrm{tr}(\Gamma_t)}{\delta}\epsilon_{\max}^2 \\ &+ \frac{8\sigma^2}{\theta}\sum_{i=1}^{n}\log\frac{9M_i n}{\delta} + 8\sigma^2\log\frac{9}{\delta} \\ &+ \frac{36(1+\theta^{-1})C^2\sigma^2 n\,\mathrm{tr}(\Gamma_t)}{\delta\lambda_{\min}(\Gamma_k)}\Xi_t(\delta,k). \end{aligned} \tag{72}$$

By $0 < \theta \le \min\left\{\frac{1}{2},\ \frac{\delta\mu_{\min}}{576\sigma^2 n S_1^2\,\mathrm{tr}(\Gamma_t)}\right\}$,

$$\frac{18\theta\sigma^2 n S_1^2\,\mathrm{tr}(\Gamma_t)}{\delta} \le \frac{\mu_{\min}}{32}.$$

Combining (71) and (72) yields

$$\begin{aligned}\frac{3t\mu_{\min}}{32}\,\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\star\|_\infty^2 \le{}& \frac{27\sigma^2 n S_1^2 t\,\mathrm{tr}(\Gamma_t)}{\delta}\epsilon_{\max}^2 \\ &+ \frac{8\sigma^2}{\theta}\sum_{i=1}^{n}\log\frac{9M_i n}{\delta} + 8\sigma^2\log\frac{9}{\delta} \\ &+ \frac{36(1+\theta^{-1})C^2\sigma^2 n\,\mathrm{tr}(\Gamma_t)}{\delta\lambda_{\min}(\Gamma_k)}\Xi_t(\delta,k). \end{aligned} \tag{73}$$

Since $\theta \le 1/2$,

$$1 + \theta^{-1} \le \frac{2}{\theta}, \qquad 1 \le \frac{1}{\theta}.$$

Dividing (73) by $3t\mu_{\min}/32$ therefore gives

$$\begin{aligned}\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\star\|_\infty^2 \le{}& \frac{288\sigma^2 n S_1^2\,\mathrm{tr}(\Gamma_t)}{\delta\mu_{\min}}\epsilon_{\max}^2 \\ &+ \frac{256}{3t\mu_{\min}\theta}\Bigg[\sigma^2\left(\sum_{i=1}^{n}\log\frac{9M_i n}{\delta} + \log\frac{9}{\delta}\right) \\ &\qquad + \frac{9C^2\sigma^2 n\,\mathrm{tr}(\Gamma_t)}{\delta\lambda_{\min}(\Gamma_k)}\Xi_t(\delta,k)\Bigg], \end{aligned} \tag{74}$$

which proves (67). □

C Proof of Theorem 2

Proof. Theorem 2 follows immediately by applying Lemmas 3 and 4 to (12). □

C.1 Proof of Lemma 3

Proof.

$$\left\|W_t X_t^\top (X_t X_t^\top)^{-1}\right\| = \left\|\underbrace{W_t X_t^\top (X_t X_t^\top)^{-\frac{1}{2}}}_{:=T_1}\underbrace{(X_t X_t^\top)^{-\frac{1}{2}}}_{:=T_2}\right\|$$

We first prove the process x_t satisfies a block martingale small-ball condition (similar to Proposition 3.1 [34]). Let $\mathcal{F}_s := \sigma(\eta_\tau : \tau \le s-1)$. For any $l \ge 1$,

$$x_{s+l} = \sum_{m=0}^{l-1} G_m \eta_{s+l-1-m} + \sum_{m=l}^{\infty} G_m \eta_{s+l-1-m},$$

so

$$x_{s+l}|\mathcal{F}_s \sim \mathcal{N}(0, \sigma^2 H_l), \qquad H_l = \sum_{m=0}^{l-1} G_m G_m^\top.$$

For any $w \in \mathbb{S}^{n-1}$,

$$\langle w, x_{s+l}\rangle|\mathcal{F}_s \sim \mathcal{N}(0, \sigma^2 w^T H_l w).$$

Since (i) $H_\ell \succeq H_{k'}$ for $\ell \ge k'$, (ii) Paley-Zygmund lower bound, we have

$$\mathbb{P}\Big(|\langle w, x_{s+\ell}\rangle| \ge \sigma\sqrt{w^\top H_{k'} w}\,\Big|\,\mathcal{F}_s\Big) \ge \frac{3}{10}, \qquad \ell \ge k'.$$

Thus,

$$\frac{1}{k}\sum_{l=1}^{k}\mathbb{P}\Big(|\langle w, x_{s+l}\rangle| \ge \sigma\sqrt{w^\top H_{k'} w}\Big) \;\ge\; \frac{1}{k}\sum_{l=k'}^{k}\mathbb{P}\Big(|\langle w, x_{s+l}\rangle| \ge \sigma\sqrt{w^\top H_{k'} w}\Big) \;\ge\; \frac{3}{10}\cdot\frac{k-k'+1}{k}.$$

Pick $k' = \lfloor k/2\rfloor$. Therefore x_t satisfies a $(k, \Gamma_{\rm sb}, p)$-BMSB condition with $\Gamma_{\rm sb} = \sigma^2 H_{\lfloor k/2\rfloor}, p = \frac{3}{20}$. Let $Z_i = \langle w, x_i\rangle$. Then Z_i satisfies the (k, v_w, p), where $v_w = \sqrt{w^\top \Gamma_{sb} w}, p = \frac{3}{20}$. Proposition 2.5 says that If $(Z_i)_{i=1}^T$ satisfies (k, ν, p)-BMSB, then

$$\Pr\left(\sum_{i=1}^{T} Z_i^2 \le \frac{\nu^2 p^2}{8}\, k\lfloor T/k\rfloor\right) \le \exp\big(-\lfloor T/k\rfloor p^2/8\big).$$

Applying this with $T = t$ and $\nu = \nu_w$,

$$\Pr\left(w^\top X_t X_t^\top w \le \frac{p^2}{8}\, k\lfloor t/k\rfloor\, w^\top \Gamma_{\rm sb} w\right) \le \exp\big(-\lfloor t/k\rfloor p^2/8\big).$$

By Lemma 4.1 [34] (in our notation $Q = X_t^\top$) , if $\inf_{w\in\mathcal{T}} w^\top X_t X_t^\top w \ge 1$ and $X_t X_t^\top \lesssim \Gamma_{\max}$, then

$$X_t X_t^\top \succeq \frac{\Gamma_{\min}}{2} = \frac{p^2}{16}\, k\lfloor t/k\rfloor \Gamma_{\rm sb}. \tag{75}$$

Define the following events ($\bar\Gamma = \sum_{m=0}^{t-1} G_m G_m^\top$)

$$\mathcal{E}_1 := \{\|W_t V\|_{\rm op} \ge K\}, \qquad \mathcal{E}_2 := \left\{X_t X_t^\top \succeq \frac{k\lfloor t/k\rfloor p^2 \Gamma_{\rm sb}}{16}\right\}, \qquad \mathcal{E}_3 := \left\{X_t X_t^\top \not\preceq t\bar\Gamma\right\}.$$

$$\mathbb{P}\left[\left\{\left\|W_t X_t^\top (X_t X_t^\top)^{-1}\right\|_{\rm op} \ge \frac{4K}{p\sqrt{k\lfloor t/k\rfloor\,\lambda_{\min}(\Gamma_{\rm sb})}}\right\}\right] \le \mathbb{P}[\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3^c] + \mathbb{P}[\mathcal{E}_2^c\cap\mathcal{E}_3^c] + \mathbb{P}[\mathcal{E}_3].$$

By the same argument of bounding $\mathbb{P}[\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3^c]$ and $\mathbb{P}[\mathcal{E}_2^c\cap\mathcal{E}_3^c]$ and by $\mathbb{P}[\mathcal{E}_3] \le \delta$, we have

$$\mathbb{P}\left[\left\|W_t X_t^\top (X_t X_t^\top)^{-1}\right\|_{\rm op} > \frac{90\sigma}{p}\sqrt{\frac{n + n\log\frac{10}{p} + \log\det\bar\Gamma\Gamma_{\rm sb}^{-1} + \log\big(\frac{1}{\delta}\big)}{t\,\lambda_{\min}(\Gamma_{\rm sb})}}\right] \le 3\delta,$$

if

$$t \geq \frac{10k}{p^2}\left(\log\left(\frac{1}{\delta}\right) + 2d\log(10/p) + \log\det(\Gamma_{\mathrm{sb}}^{-1})\right).$$

Since

$$\mathbb{P}\left[X_t X_t^\top \not\preccurlyeq \frac{\sigma^2 d}{\delta}\, t\Gamma_t\right] \leq \frac{\delta}{d\sigma^2}\,\mathbb{E}\Big[\mathrm{tr}\Big((t\Gamma_t)^{-1/2} X_t X_t^\top (t\Gamma_t)^{-1/2}\Big)\Big] \leq \delta,$$

then there exist universal constants $c, C > 0$ such that

$$\mathbb{P}\left[\left\|W_t X_t^\top (X_t X_t^\top)^{-1}\right\|_{\mathrm{op}} > \frac{C}{\sqrt{T\,\lambda_{\min}(\Gamma_k)}}\sqrt{d\log\frac{d}{\delta} + \log\det(\Gamma_T\Gamma_k^{-1})}\right] \leq \delta,$$

for any k such that $\frac{T}{k} \geq c\big(d\log(d/\delta) + \log\det(\Gamma_T\Gamma_k^{-1})\big)$ holds. □

C.2 Proof of Lemma 4

Proof. For each row $i \in [n]$ and $s = 0, \ldots, t-1$, recall that

$$b_s^{(i)}(\alpha_i) = \big(\Delta^{\alpha_i} - \Delta^{\alpha_{i,\star}}\big)x_{s+1}^{(i)}.$$

Since the $j = 0$ coefficient does not depend on α_i, we have

$$\begin{aligned} b_s^{(i)}(\alpha_i) &= \sum_{j=1}^{s+1}\big(\psi(\alpha_i, j) - \psi(\alpha_{i,\star}, j)\big)x_{s+1-j}^{(i)} \\ &= \int_{\alpha_{i,\star}}^{\alpha_i}\sum_{j=1}^{s+1}\partial_u\psi(u,j)\,x_{s+1-j}^{(i)}\,du. \end{aligned}$$

Let

$$B_t^{(i)}(\alpha_i) := \big(b_0^{(i)}(\alpha_i), \ldots, b_{t-1}^{(i)}(\alpha_i)\big) \in \mathbb{R}^{1\times t},$$

and

$$B_t(\boldsymbol{\alpha}) := \begin{bmatrix} B_t^{(1)}(\alpha_1) \\ \vdots \\ B_t^{(n)}(\alpha_n) \end{bmatrix} \in \mathbb{R}^{n\times t}.$$

For each row i, define

$$S_{1,i} := \sup_{u\in\mathcal{A}_i}\sum_{j\geq 1}|\partial_u\psi(u,j)|, \qquad S_1 := \max_{i\in[n]} S_{1,i}.$$

By Minkowski's integral inequality and Young's convolution inequality,

$$\begin{aligned} \|B_t^{(i)}(\alpha_i)\|_2 &\leq |\alpha_i - \alpha_{i,\star}|\sup_{u\in\mathcal{A}_i}\left(\sum_{s=0}^{t-1}\left|\sum_{j=1}^{s+1}\partial_u\psi(u,j)x_{s+1-j}^{(i)}\right|^2\right)^{1/2} \\ &\leq |\alpha_i - \alpha_{i,\star}|\sup_{u\in\mathcal{A}_i}\left(\sum_{j=1}^{t}|\partial_u\psi(u,j)|\right)\left(\sum_{s=0}^{t-1}|x_s^{(i)}|^2\right)^{1/2} \\ &\leq S_{1,i}|\alpha_i - \alpha_{i,\star}|\left(\sum_{s=0}^{t-1}|x_s^{(i)}|^2\right)^{1/2}. \end{aligned}$$

Consequently,

$$\|B_t(\boldsymbol{\alpha})\|_F^2 = \sum_{i=1}^{n}\|B_t^{(i)}(\alpha_i)\|_2^2 \leq S_1^2\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_\star\|_\infty^2\sum_{i=1}^{n}\sum_{s=0}^{t-1}|x_s^{(i)}|^2 = S_1^2\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_\star\|_\infty^2\|X_t\|_F^2.$$

On the event $X_t X_t^\top \succ 0$,

$$\begin{aligned}\left\|B_t(\boldsymbol{\alpha}) X_t^\top (X_t X_t^\top)^{-1}\right\|_{\mathrm{op}} &\le \|B_t(\boldsymbol{\alpha})\|_F \left\|X_t^\top (X_t X_t^\top)^{-1}\right\|_{\mathrm{op}} \\ &= \frac{\|B_t(\boldsymbol{\alpha})\|_F}{\sigma_{\min}(X_t)} \\ &\le S_1 \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_\star\|_\infty \frac{\|X_t\|_F}{\sigma_{\min}(X_t)}. \end{aligned} \tag{76}$$

It remains to control $\|X_t\|_F / \sigma_{\min}(X_t)$. Since $x_0 = 0$, the finite-past representation gives, for $s \ge 1$,

$$x_s = \sum_{m=0}^{s-1} G_m \eta_{s-1-m}.$$

Therefore,

$$\mathbb{E}\|X_t\|_F^2 = \sum_{s=0}^{t-1} \mathbb{E}\|x_s\|_2^2 = \sigma^2 \sum_{s=1}^{t-1} \sum_{m=0}^{s-1} \|G_m\|_F^2 \le t\sigma^2 \sum_{m=0}^{\infty} \|G_m\|_F^2.$$

Let

$$\mathcal{G}_2 := \sum_{m=0}^{\infty} \|G_m\|_F^2.$$

By Markov's inequality, with probability at least $1 - \delta$,

$$\|X_t\|_F^2 \le \frac{t\sigma^2 \mathcal{G}_2}{\delta}. \tag{77}$$

Moreover, by the block small-ball lower bound, with probability at least $1 - \delta$,

$$X_t X_t^\top \succeq \frac{p^2}{16} k \lfloor t/k \rfloor \Gamma_{\mathrm{sb}}. \tag{78}$$

Hence, on the intersection of (77) and (78),

$$\frac{\|X_t\|_F^2}{\sigma_{\min}^2(X_t)} = \frac{\|X_t\|_F^2}{\lambda_{\min}(X_t X_t^\top)} \le \frac{16 t\sigma^2 \mathcal{G}_2}{\delta p^2 k \lfloor t/k \rfloor \lambda_{\min}(\Gamma_{\mathrm{sb}})} \le \frac{16 t\sigma^2 \mathcal{G}_2}{\delta p^2 (t-k) \lambda_{\min}(\Gamma_{\mathrm{sb}})}.$$

If $t \ge 10k$, then $t/(t-k) \le 10/9$, and thus

$$\frac{\|X_t\|_F}{\sigma_{\min}(X_t)} \le \sqrt{\frac{160\sigma^2 \mathcal{G}_2}{9\delta p^2 \lambda_{\min}(\Gamma_{\mathrm{sb}})}}. \tag{79}$$

Combining (76) and (79), we obtain, with probability at least $1 - 2\delta$,

$$\left\|B_t(\boldsymbol{\alpha}) X_t^\top (X_t X_t^\top)^{-1}\right\|_{\mathrm{op}} \le S_1 \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_\star\|_\infty \sqrt{\frac{160\sigma^2 \mathcal{G}_2}{9\delta p^2 \lambda_{\min}(\Gamma_{\mathrm{sb}})}}.$$

Finally, Lemma 14 gives

$$\mathcal{G}_2 \le n\widetilde{C}_G^2 \sum_{m=0}^{\infty} (m+1)^{-2(1+\alpha_{\min})} = n\tilde{C}_G^2 \zeta(2 + 2\alpha_{\min}) \le 2n\tilde{C}_G^2,$$

and therefore

$$\left\|B_t(\boldsymbol{\alpha}) X_t^\top (X_t X_t^\top)^{-1}\right\|_{\mathrm{op}} \le S_1 \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_\star\|_\infty \sqrt{\frac{320 n\sigma^2 \tilde{C}_G^2}{9\delta p^2 \lambda_{\min}(\Gamma_{\mathrm{sb}})}}. \tag{80}$$

This proves the claim. □

D Auxiliary Lemmas

Lemma 14. *Suppose Assumption 1 holds and let $\alpha_{\min} := \min_{i\in[n]} \alpha_{i,\star}$. Then there exists a constant $\widetilde{C}_G < \infty$, depending only on $(A_\star, \alpha_\star)$, such that*

$$\|G_m\|_{\mathrm{op}} \le \widetilde{C}_G (m+1)^{-(1+\alpha_{\min})}, \qquad m \ge 0. \tag{81}$$

Consequently, for every $p \ge 1$ and $k \ge 0$,

$$\|\operatorname{Cov}(x_p, x_{p+k})\|_{\mathrm{op}} \le \widetilde{C}_x (k+1)^{-(1+\alpha_{\min})}, \tag{82}$$

where

$$\widetilde{C}_x := \sigma^2 \widetilde{C}_G^2 \left(1 + \frac{1}{\alpha_{\min}} + \frac{1}{2\alpha_{\min}+1}\right).$$

Proof. we first bound G_m. Let $\Psi_j := \operatorname{diag}\big(\psi(\alpha_{1,\star}, j), \dots, \psi(\alpha_{n,\star}, j)\big)$. We have

$$\mathcal{A}(z) = A_\star + \operatorname{diag}(\boldsymbol{\alpha}_\star) - \sum_{j=1}^{\infty} \Psi_{j+1} z^j.$$

For each coordinate i, $\sum_{k=0}^{\infty} \psi(\alpha_{i,\star}, k)\, z^k = (1-z)^{\alpha_{i,\star}}$. Therefore

$$\sum_{j=1}^{\infty} \psi(\alpha_{i,\star}, j+1)\, z^j = \frac{(1-z)^{\alpha_{i,\star}} - 1 + \alpha_{i,\star} z}{z}.$$

Define $D_{\boldsymbol{\alpha}_\star}(z) := \operatorname{diag}\big((1-z)^{\alpha_{1,\star}}, \dots, (1-z)^{\alpha_{n,\star}}\big)$. Then

$$\mathcal{G}(z) = \big(D_{\boldsymbol{\alpha}_\star}(z) - z A_\star\big)^{-1}.$$

Let $w = 1 - z$ and $D(w) := \operatorname{diag}\big(w^{\alpha_{1,\star}}, \dots, w^{\alpha_{n,\star}}\big)$. Then

$$\mathcal{G}(z) = -\big(I - B(w)\big)^{-1} A_\star^{-1},$$

where $B(w) := wI + A_\star^{-1} D(w)$. Define

$$R(z) := B(w)^2 (I - B(w))^{-1} A_\star^{-1}.$$

Then, we have

$$G_m = -A_\star^{-1} \operatorname{diag}\big(\psi(\alpha_{1,\star}, m), \dots, \psi(\alpha_{n,\star}, m)\big) A_\star^{-1} - [z^m] R(z), \quad m \ge 2.$$

For $|w| \le 1$,

$$\|D(w)\|_{\mathrm{op}} = \max_i |w|^{\alpha_{i,\star}} \le |w|^{\alpha_{\min}},$$

and therefore

$$\|B(w)\|_{\mathrm{op}} \le \big(1 + \|A_\star^{-1}\|_{\mathrm{op}}\big)|w|^{\alpha_{\min}}.$$

Choose $\rho > 0$ sufficiently small so that $\|B(w)\|_{\mathrm{op}} \le 1/2$ whenever $|w| \le \rho$. Then

$$\|(I - B(w))^{-1}\|_{\mathrm{op}} \le 2,$$

and thus

$$\|R(z)\|_{\mathrm{op}} \le 2\|A_\star^{-1}\|_{\mathrm{op}} \|B(w)\|_{\mathrm{op}}^2 \le 2\|A_\star^{-1}\|_{\mathrm{op}} \big(1 + \|A_\star^{-1}\|_{\mathrm{op}}\big)^2 |w|^{2\alpha_{\min}}.$$

Using the singularity transfer argument in Lemma 15, we have

$$\|[z^m] R(z)\|_{\mathrm{op}} \le C_I \|A_\star^{-1}\|_{\mathrm{op}} \big(1 + \|A_\star^{-1}\|_{\mathrm{op}}\big)^2 m^{-(1+2\alpha_{\min})},$$

where C_I depends only on the interval for $\alpha_{i,\star}$.

We next bound the fractional coefficient $\psi(\alpha, j)$. For $0 < \alpha < 1$,

$$\psi(\alpha, j) = \frac{\Gamma(j-\alpha)}{\Gamma(-\alpha)\Gamma(j+1)} = (-1)^j \binom{\alpha}{j}.$$

For $j \geq 2$,

$$|\psi(\alpha, j)| = \frac{\alpha}{j} \prod_{k=1}^{j-1} \left(1 - \frac{\alpha}{k}\right).$$

Using $\log(1-u) \leq -u$ for $0 < u < 1$ and $\sum_{k=1}^{j-1} k^{-1} \geq \log j$, we obtain

$$|\psi(\alpha, j)| \leq \alpha j^{-(1+\alpha)}, \qquad j \geq 1.$$

For $\alpha = 1$, the same bound follows directly from $\psi(1,1) = -1$ and $\psi(1,j) = 0$ for $j \geq 2$.

Hence

$$\max_{i \in [n]} |\psi(\alpha_{i,\star}, m)| \leq m^{-(1+\alpha_{\min})}, \qquad m \geq 1.$$

Then, we have

$$\|G_m\|_{\mathrm{op}} \leq \|A_\star^{-2}\|_{\mathrm{op}}\, m^{-(1+\alpha_{\min})} + C_I \|A_\star^{-1}\|_{\mathrm{op}} (1 + \|A_\star^{-1}\|_{\mathrm{op}})^2 m^{-(1+2\alpha_{\min})}.$$

Thus

$$\|G_m\|_{\mathrm{op}} \leq \tilde{C}_G (m+1)^{-(1+\alpha_{\min})},\ m \geq 0, \tag{83}$$

where $\tilde{C}_G = C_I' \Big(\|A_\star^{-2}\|_{\mathrm{op}} + \|A_\star^{-1}\|_{\mathrm{op}} (1 + \|A_\star^{-1}\|_{\mathrm{op}})^2 \Big)$.

Therefore,

$$\begin{aligned} \|\operatorname{Cov}(x_p, x_{p+k})\|_{\mathrm{op}} &\leq \sigma^2 \sum_{m=0}^{p-1} \|G_m\|_{\mathrm{op}} \|G_{m+k}\|_{\mathrm{op}} \\ &\leq \sigma^2 \widetilde{C}_G^2 \sum_{m=0}^{\infty} (m+1)^{-(1+\alpha_{\min})} (m+k+1)^{-(1+\alpha_{\min})}. \end{aligned}$$

Splitting the last sum into $0 \leq m \leq k$ and $m > k$ gives

$$\sum_{m=0}^{\infty} (m+1)^{-(1+\alpha_{\min})} (m+k+1)^{-(1+\alpha_{\min})} \leq \left(1 + \frac{1}{\alpha_{\min}} + \frac{1}{2\alpha_{\min}+1}\right) (k+1)^{-(1+\alpha_{\min})}.$$

Thus

$$\|\operatorname{Cov}(x_p, x_{p+k})\|_{\mathrm{op}} \leq \widetilde{C}_x (k+1)^{-(1+\alpha_{\min})}.$$

The case $q < p$ follows by covariance symmetry, which proves the claim for all $p, q \geq 0$. □

Lemma 15. *Fix $\phi \in (0, \pi/2)$ and $\beta_0 > 0$. Then there exists a constant $K_{\phi,\beta_0} < \infty$ such that the following holds.*

Let $\rho \in (0, 1]$, and let

$$\Delta(\phi, \rho) := \{z : |z| < 1 + \rho,\ z \neq 1,\ |\arg(z-1)| > \phi\}.$$

If f is analytic in $\Delta(\phi, \rho)$ and satisfies

$$|f(z)| \leq M |1 - z|^{\beta}, \qquad z \in \Delta(\phi, \rho),$$

for some $\beta \in [0, \beta_0]$, then for every $m \geq 1$,

$$|[z^m] f(z)| \leq K_{\phi,\beta_0}\, M\, \rho^{-(\beta+1)}\, m^{-(\beta+1)}.$$

Proof. We follow the standard truncated Hankel-contour proof of the Flajolet–Odlyzko transfer theorem [6], keeping the ρ-dependence explicit.

For $m < 2\rho^{-1}$, the right-hand side is $\geq 2^{-(\beta+1)} \geq 2^{-(\beta_0+1)}$. Since f is analytic on a fixed contour inside $\Delta(\phi, \rho)$ enclosing the origin, Cauchy's formula gives $|[z^m] f(z)| \leq C_{\phi,\beta_0} M$, so the claim follows after enlarging K_{ϕ,β_0}. One choice is $C_{\phi,\beta_0} = 2^{\beta_0}$ and $K_{\phi,\beta_0} \geq 2^{2\beta_0+1}$.

Hence, it suffices to consider $m \geq 2\rho^{-1}$. Fix $\vartheta = \frac{\phi + \pi/2}{2} \in (\phi, \pi/2)$ and let $r_{\rho,\vartheta} > 0$ be the unique solution to

$$|1 + r e^{i\vartheta}| = 1 + \frac{\rho}{2},$$

namely, $r_{\rho,\vartheta} = -\cos\vartheta + \sqrt{\cos^2\vartheta + \rho + \frac{\rho^2}{4}}$. Consider the contour $H_{m,\rho,\vartheta} \in \Delta(\phi,\rho)$ defined by

$$\Gamma_\pm := \{1 + re^{\pm i\vartheta} : m^{-1} \le r \le r_{\rho,\vartheta}\},$$

$$\Gamma_0 := \{1 + m^{-1}e^{i\theta} : \vartheta \le \theta \le 2\pi - \vartheta\},$$

and the outer arc $\Gamma_{\mathrm{out}} \subset \{|z| = 1 + \rho/2\} \cap \Delta(\phi,\rho)$ joining the endpoints of Γ_+ and Γ_-. Therefore, $[z^m]f(z) = \frac{1}{2\pi i}\int_{H_{m,\rho,\vartheta}} f(z)z^{-m-1}dz$. On $\Gamma_\pm$, $|1-z| = r$, $|dz| = dr$, and

$$|z| = |1 + re^{\pm i\vartheta}| \ge 1 + r\cos\vartheta.$$

Therefore, on $\Gamma_\pm$, $|z|^{-m} \le (1 + r\cos\vartheta)^{-m} = e^{-m\log(1+r\cos\vartheta)} \le e^{-\frac{mr\cos\vartheta}{1+r_{1,\vartheta}\cos\vartheta}} = e^{-c_\phi mr}$, where $c_\phi := \frac{\cos\vartheta}{1+r_{1,\vartheta}\cos\vartheta}$. Using $|f(z)| \le Mr^\beta$ there, we obtain

$$\int_{\Gamma_\pm} |f(z)z^{-m-1}\,dz| \le M\int_{m^{-1}}^{r_{\rho,\vartheta}} r^\beta e^{-c_\phi mr}\,dr \le M(c_\phi m)^{-(\beta+1)}\Gamma(\beta+1) \le C_{\phi,\beta_0}Mm^{-(\beta+1)}.$$

On Γ_0, we have $|1-z| = m^{-1}$ and $|f(z)| \le Mm^{-\beta}$. The arc length satisfies $(2\pi - 2\vartheta)m^{-1} \le C_\phi m^{-1}$. Since the smallest possible value of $|z|$ on the arc occurs at $\theta = \pi$, we have $|z| \ge 1 - \frac{1}{m}$. Since $m \ge 2\rho^{-1} \ge 2$, we can uniformly upper bound $|z|^{-m-1}$ by $|z|^{-m-1} \le (1-\frac{1}{m})^{-(m+1)} \le 8$. Hence

$$\int_{\Gamma_0} |f(z)z^{-m-1}\,dz| \le 8Mm^{-\beta}C_\phi m^{-1} = C'_\phi Mm^{-(\beta+1)}.$$

On the outer arc Γ_{out}, $|z| = 1 + \rho/2$, so

$$|z|^{-m} = (1+\rho/2)^{-m} \le e^{-c\rho m}$$

for a universal $c > 0$. Also, since $\beta \le \beta_0$ and $|1-z| \le 3$ on $|z| = 1+\rho/2$ with $\rho \le 1$, we have $|f(z)| \le 3^{\beta_0}M$. Hence

$$\int_{\Gamma_{\mathrm{out}}} |f(z)z^{-m-1}\,dz| \le C_{\phi,\beta_0}Me^{-c\rho m}.$$

Finally, because

$$\sup_{x>0} x^{\beta_0+1}e^{-cx} < \infty$$

and $\rho \le 1$, we have

$$e^{-c\rho m} \le C_{\beta_0}\rho^{-(\beta+1)}m^{-(\beta+1)}.$$

Combining the three contour bounds with Cauchy's coefficient formula proves the claim.

□

Lemma 16. *Suppose Assumption 1 holds and consider the stationary two-sided solution of the system. Define*

$$s_G := \inf_{|z|=1} \sigma_{\min}(G(z)), \qquad \alpha_{\max} := \max_{i\in[n]} \alpha_{i,\star}.$$

Then, for every $i \in [n]$,

$$\mu_{t,i} \ge \sigma^2 s_G^2 \left[\left(\alpha_{i,\star} - \frac{1}{2}\right)^2 + \left(\frac{-3\alpha_{i,\star}^2 + 6\alpha_{i,\star} - 2}{6}\right)^2\right]$$

$$\ge \frac{\sigma^2}{\left(2^{\alpha_{\max}} + \|A_\star\|_{\mathrm{op}}\right)^2}\left[\left(\alpha_{i,\star} - \frac{1}{2}\right)^2 + \left(\frac{-3\alpha_{i,\star}^2 + 6\alpha_{i,\star} - 2}{6}\right)^2\right].$$

Proof. Let

$$d_{i,j} := \partial_\alpha\psi(\alpha_{i,\star}, j), \qquad j \ge 1,$$

and define

$$g_s^{(i)} := \sum_{j\ge1} d_{i,j}\, x_{s+1-j}^{(i)}.$$

By the definition of the population profiled derivative curvature,

$$\mu_{t,i} = \inf_{a_i \in \mathbb{R}^{1\times n}} \frac{1}{t} \sum_{s=0}^{t-1} \mathbb{E} \left| g_s^{(i)} - a_i x_s \right|^2 .$$

Under stationarity, every term in the preceding average has the same distribution. Hence

$$\mu_{t,i} = \inf_{a_i \in \mathbb{R}^{1\times n}} \mathbb{E} \left| g_0^{(i)} - a_i x_0 \right|^2 . \tag{84}$$

Using the stationary moving-average representation

$$x_s = \sum_{m=0}^{\infty} G_m \eta_{s-1-m},$$

we have

$$g_0^{(i)} = \sum_{j=1}^{\infty} d_{i,j} x_{1-j}^{(i)} = \sum_{\ell \geq 1} h_{i,\ell}^{\top} \eta_{-\ell},$$

where

$$h_{i,\ell}^{\top} := \sum_{j=1}^{\ell} d_{i,j}\, e_i^{\top} G_{\ell-j}.$$

Similarly,

$$a_i x_0 = \sum_{\ell=1}^{\infty} a_i G_{\ell-1} \eta_{-\ell}.$$

Therefore,

$$g_0^{(i)} - a_i x_0 = \sum_{\ell=1}^{\infty} \big(h_{i,\ell}^{\top} - a_i G_{\ell-1}\big) \eta_{-\ell}.$$

Since the innovations are independent with covariance $\sigma^2 I_n$,

$$\mathbb{E} \left| g_0^{(i)} - a_i x_0 \right|^2 = \sigma^2 \sum_{\ell=1}^{\infty} \left\| h_{i,\ell}^{\top} - a_i G_{\ell-1} \right\|_2^2 . \tag{85}$$

Next define

$$D_i(z) := \sum_{j=1}^{\infty} d_{i,j} z^j.$$

Since

$$\sum_{j=0}^{\infty} \psi(\alpha, j) z^j = (1-z)^{\alpha},$$

we obtain

$$D_i(z) = \partial_{\alpha} (1-z)^{\alpha} \big|_{\alpha=\alpha_{i,\star}} = (1-z)^{\alpha_{i,\star}} \log(1-z).$$

Moreover,

$$\sum_{\ell \geq 1} h_{i,\ell}^{\top} z^{\ell} = D_i(z) e_i^{\top} G(z),$$

whereas

$$\sum_{\ell \geq 1} a_i G_{\ell-1} z^{\ell} = a_i z G(z).$$

Thus,

$$\sum_{\ell \geq 1} \big(h_{i,\ell}^{\top} - a_i G_{\ell-1}\big) z^{\ell} = \big(D_i(z) e_i^{\top} - a_i z\big) G(z).$$

By Parseval's identity,

$$\sum_{\ell\geq 1}\left\|h_{i,\ell}^{\top}-a_iG_{\ell-1}\right\|_2^2=\frac{1}{2\pi}\int_0^{2\pi}\left\|\left(D_i(e^{\mathrm{i}\theta})e_i^{\top}-a_ie^{\mathrm{i}\theta}\right)G(e^{\mathrm{i}\theta})\right\|_2^2\,d\theta. \tag{86}$$

By Assumption 1, $G(z)$ is invertible for every $|z|=1$. Since G is continuous on the unit circle,

$$s_G=\inf_{|z|=1}\sigma_{\min}(G(z))>0.$$

Hence, using

$$\|MN\|_2\geq\sigma_{\min}(N)\|M\|_2,$$

(86) gives

$$\sum_{\ell\geq 1}\left\|h_{i,\ell}^{\top}-a_iG_{\ell-1}\right\|_2^2\geq s_G^2\frac{1}{2\pi}\int_0^{2\pi}\left\|D_i(e^{\mathrm{i}\theta})e_i^{\top}-a_ie^{\mathrm{i}\theta}\right\|_2^2\,d\theta.$$

Applying Parseval's identity once more,

$$\sum_{\ell\geq 1}\left\|h_{i,\ell}^{\top}-a_iG_{\ell-1}\right\|_2^2\geq s_G^2\left[\left\|d_{i,1}e_i^{\top}-a_i\right\|_2^2+\sum_{j=2}^{\infty}d_{i,j}^2\right]. \tag{87}$$

Combining (84), (85), and (87), and minimizing over $a_i\in\mathbb{R}^{1\times n}$, yields

$$\begin{aligned}\mu_{t,i}&\geq\sigma^2s_G^2\inf_{a_i\in\mathbb{R}^{1\times n}}\left[\left\|d_{i,1}e_i^{\top}-a_i\right\|_2^2+\sum_{j=2}^{\infty}d_{i,j}^2\right]\\&=\sigma^2s_G^2\sum_{j=2}^{\infty}d_{i,j}^2.\end{aligned} \tag{88}$$

For the first two terms in this sum,

$$\psi(\alpha,2)=\frac{\alpha(\alpha-1)}{2},\qquad\psi(\alpha,3)=-\frac{\alpha(\alpha-1)(\alpha-2)}{6}.$$

Therefore,

$$d_{i,2}=\alpha_{i,\star}-\frac{1}{2},$$

and

$$d_{i,3}=\frac{-3\alpha_{i,\star}^2+6\alpha_{i,\star}-2}{6}.$$

It follows from (88) that

$$\mu_{t,i}\geq\sigma^2s_G^2\left[\left(\alpha_{i,\star}-\frac{1}{2}\right)^2+\left(\frac{-3\alpha_{i,\star}^2+6\alpha_{i,\star}-2}{6}\right)^2\right]. \tag{89}$$

It remains to lower bound s_G. For $|z|=1$,

$$G(z)^{-1}=\operatorname{diag}\left((1-z)^{\alpha_{1,\star}},\ldots,(1-z)^{\alpha_{n,\star}}\right)-zA_\star.$$

Hence

$$\begin{aligned}\|G(z)^{-1}\|_{\mathrm{op}}&\leq\max_{i\in[n]}|1-z|^{\alpha_{i,\star}}+\|A_\star\|_{\mathrm{op}}\\&\leq 2^{\alpha_{\max}}+\|A_\star\|_{\mathrm{op}},\end{aligned}$$

because $|1-z|\leq 2$ on the unit circle. Therefore,

$$s_G=\inf_{|z|=1}\frac{1}{\|G(z)^{-1}\|_{\mathrm{op}}}\geq\frac{1}{2^{\alpha_{\max}}+\|A_\star\|_{\mathrm{op}}}.$$

Substituting this bound into (89) completes the proof. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflects the main contributions, which are later more formally stated in main theorems (Section 4) and justified by proofs (Section 5) and experiments (Section 6).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses both the theoretical and experimental limitations of this work (Section 7).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper offers an explicit statement of the assumption (Section 4) as well as a full proof (Appendices B and C).

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The code is included in a github repository, which exactly reproduces the desired result. The paper also extensively discusses how to reproduce the result (Appendix A).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code is available at: `https://github.com/Eric0312/FO-GS`.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides all experimental details (Appendix A), along with the full code needed to reproduce the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Figure 2 shows shaded regions representing 95% confidence intervals. Tables 1 and 2 also report 95% confidence intervals.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments were run on an AMD 9800X3D CPU and an RTX 5080 GPU, but the experiments are not device-dependent.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We confirm that our work conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Our work falls into the category of foundational research; we do not believe it poses any specific and/or unique potential harms.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper contains no risky data.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: This paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New assets introduced in the paper, such as code, are well documented. The documentation is provided alongside the assets in the supplementary material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs do not constitute an important, original, or non-standard component of our work.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.