Non-asymptotic Analysis of Set Membership Estimation for Linear Systems with Disturbances Bounded by Convex Sets (Supplementary Material)

Haonan Xu and Yingying Li

***Abstract*— This paper revisits the classical set membership estimation (SME) algorithm for linear control systems under bounded, stochastic disturbances and provides non-asymptotic guarantees. Most of the literature analyzing the non-asymptotic behaviors of SME only considers *linear regression* under stochastic and bounded disturbances, which ignores the *correlation* between current and history states introduced by *system dynamics*. Recently, there has been a renewed interest in rigorizing SME's non-asymptotic analysis for control systems by leveraging recent statistical learning techniques. For example, [1] re-establishes SME's convergence for disturbances tightly bounded by a general convex set, while [2] proposes a convergence rate under more restrictive disturbances: the disturbance's support should be an ℓ_∞ ball. This paper generalizes the convergence rates in [2] to allow disturbances bounded by general convex sets as in [1]. In addition, we further relax the assumption from the one in [1] to establish more general convergence and convergence rate guarantees. Our theoretical bounds are validated by several numerical experiments.**

This supplementary material contains the proofs omitted from the main text, which are provided in the Appendix.

I. Introduction

System identification enjoys a long history of research [3], [4]. Recent years have witnessed a revived interest in the non-asymptotic analysis of system identification algorithms for control dynamics through the lens of statistical learning [5], [6], [7], [8]. For example, consider a linear control system

$$x_{t+1} = A^* x_t + B^* u_t + w_t, \tag{1}$$

where (A^*, B^*) are unknown system parameters, $x_t \in \mathbb{R}^{n_x}$ is the state, $u_t \in \mathbb{R}^{n_u}$ is the control input, and $w_t \in \mathbb{R}^{n_x}$ is the process noise/disturbance. Least square estimator (LSE) is a widely adopted point estimator of (A^*, B^*) and its non-asymptotic convergence rates under i.i.d. noises w_t have been extensively studied recently [5], [6], [7], [8].[1] This has further ignited significant research into the non-asymptotic analysis of various LSE-learning-based control algorithms, e.g, regrets [9], sample complexity [10], etc.

Besides LSE, set membership estimation (SME) is another popular estimation algorithm in the control literature [11], [12], [13], [14], [15]. Compared with the point estimator LSE, SME is a set estimator that focuses on bounded noises $w_t \in \mathbb{W}$ and leverages the set $\mathbb{W}$ to characterize the uncertainty set of A^*, B^*. SME is popular among robust adaptive control as it effectively learns and refines uncertainty sets for robust controllers, such as robust adaptive model predictive control (RAMPC) [1], [14], robust adaptive control barrier functions [16], etc.

Compared to LSE, SME has received relatively less attention and analysis from the perspective of statistical learning. Though the convergence (rates) of SME were well-studied for linear regression [17], [18], [12], its convergence analysis for linear control systems has emerged more recently. For example, when measuring set convergence by diameters (a metric particularly useful for non-asymptotic analysis of learning-based controllers that utilize SME), the convergence of SME for general linear dynamics is established in [1] for any convex compact set $\mathbb{W}$.[2] Further, the convergence rate of SME for linear dynamics is recently provided in [2], but only for a special $\mathbb{W}$: ℓ_∞ ball $\mathbb{W} = \{w : \|w\|_\infty \le w_{\max}\}$.

Unlike LSE, the shape of $\mathbb{W}$ plays an important role in SME because SME may not converge unless the bound $\mathbb{W}$ is tight on all directions [19], [1], [2], [12]. Therefore, most convergence results assume such tightness conditions [2], [1], [17]. However, it becomes quite restrictive when [2] also assumes a special shape of $\mathbb{W} = \{w : \|w\|_\infty \le w_{\max}\}$ because even a weighted ℓ_∞ ball do not satisfy both the shape and the tightness assumption in [2]. To see this, consider a true support of the distribution $\mathbb{W}_{\text{true}} = \{w : \max_{1\le i\le n_x} |w_i/i| \le w_{\max}\}$ and its minimum ℓ_∞-ball outer approximation: $\mathbb{W} = \{w : \|w\|_\infty \le n_x w_{\max}\}$. Notice that $\mathbb{W}$ is not tight in the first coordinate because w_t can never visit a neighborhood around $w_t^{[1]} = n_x w_{\max}$. Therefore, the particular shape of $\mathbb{W}$ in [2] significantly limits its applications. For example, consider a power grid with renewable generations as random disturbances on each nodes [20]. Some nodes may have higher renewable penetration, while others may have very little renewable capacity, so the total distribution of the renewable profiles are more likely to be bounded by a weighted ℓ_∞ ball instead of a perfect ℓ_∞ ball, which fails to satisfy the conditions in [2].

Contributions. The major contribution of this paper is providing non-asymptotic convergence rate guarantees for SME under general convex and compact $\mathbb{W}$, bridging the gap between the assumptions used for convergence and those for convergence rates in the existing literature on SME's convergence analysis for linear dynamics. Our theoretical

Haonan Xu and Yingying Li are with the University of Illinois Urbana-Champaign. E-mail: {haonan9,yl101}@illinois.edu.

[1] While LSE's convergence rates for linear regression have long been known, its rates for control systems were developed more recently due to the complications from correlations between current and past states in control dynamics [7], [6].

[2] SME's convergence under other metrics was established in [19], [13].

guarantees also lay foundation for future non-asymptotic analysis of SME-based adaptive controllers, e.g. RAMPC, etc. To establish the non-asymptotic bounds, we first rely on the standard assumption in the convergence literature [1], then propose a relaxed assumption, which not only generalizes the applicability of our bounds but also provides improved convergence rates in certain cases. Finally, we use simulation examples to validate our theoretical guarantees.

Related works. Notice that SME is also widely used for state estimation and system identification in output feedback systems [21], [22], switched systems [23], etc. It is left as our future work to connect and generalize our system estimation results to these cases.

Besides, it is worth mentioning that SME still applies for non-stochastic noises as long as the noise bound $w_t \in \mathbb{W}$ is valid [11], [12], [13]. This attracts applications for SME when the noises do not have nice statistical properties [24], [14]. However, when it comes to non-asymptotic convergence analysis, A lot of SME literature still considers stochastic noises to simplify the analysis and to compare with other stochastic-based system estimation methods [17], [1], [18].

Notations. For two matrices $M_1 \in \mathbb{R}^{k\times t}, M_2 \in \mathbb{R}^{k\times r}$, let (M_1, M_2) denote the concatenated matrix, and the same applies to vector concatenation. Let $||\cdot||_F$ denote the Frobenius norm of a matrix. Let $||\cdot||_p$ denote the ℓ_p vector norm or L_p matrix norm for $1 \le p \le \infty$. Let $A^{\complement}$ denote the complement of event/set A. We use $\tilde{O}(\cdot)$ to hide logarithmic terms. The interior of a set E is denoted by $\mathring{E}$, and the boundary of E is denoted by ∂E. $A \succ B$ means that matrix $A - B$ is positive definite. $\forall\, p \ge 1,\ r > 0, x \in \mathbb{R}^d$, let $\mathbb{B}_p^r(x)$ denote the ℓ_p ball with radius r centered at x including the ℓ_p sphere $\mathbb{S}_p^r(x)$ with radius r centered at x. If $r = 1$ and $x = 0$, we use $\mathbb{B}_p$, $\mathbb{S}_p$ for simplicity.

Mathematical preliminaries. In a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, we say $\{\mathcal{F}_i\}_{i\in\mathbb{N}}$ is a *filtration* if $\forall\, i \in \mathbb{N}$, $\mathcal{F}_i$ is a sub-σ-algebra of $\mathcal{A}$ and $\forall\, i \le j$ one has $\mathcal{F}_i \subseteq \mathcal{F}_j$. We denote the σ-algebra generated by a collection of random variables as $\sigma\{\cdots\}$. A stochastic process $\{X_i\}_{i\in\mathbb{N}}$ is said to be *adapted* to the filtration $\{\mathcal{F}_i\}_{i\in\mathbb{N}}$ if $\forall\, i \in \mathbb{N}$, the random variable X_i is an $\mathcal{F}_i$-measurable function. Let τ be a random variable taking values from $[0, +\infty)$. We say τ is a *stopping time* if $\forall t \ge 0$, we have $\{\tau \le t\} \in \mathcal{F}_t$.

II. Problem Formulation

This paper considers a linear control system:

$$x_{t+1} = A^* x_t + B^* u_t + w_t, \quad t \ge 0, \tag{2}$$

where $x_t \in \mathbb{R}^{n_x}$, $u_t \in \mathbb{R}^{n_u}$, and $w_t \in \mathbb{R}^{n_x}$ respectively denote the state, the control input, and the process noise at time $t \ge 0$. The system parameters A^*, B^* in (2) are unknown and to be estimated. For ease of notation, we denote $\theta^* = (A^*, B^*) \in \mathbb{R}^{n_x\times n_z}$, $z_t = (x_t^\top, u_t^\top)^\top \in \mathbb{R}^{n_z}$, where $n_z = n_x + n_u$. Thus, the system (2) can be written as

$$x_{t+1} = \theta^* z_t + w_t. \tag{3}$$

This paper focuses on a specific estimation algorithm, set membership estimation (SME), to quantify the uncertainty of A^*, B^*. SME is mainly applicable when w_t is bounded and utilizes the bound of w_t to construct the uncertainty sets [11], [12], [13], [14], [15]. This paper studies stage-wise bound $w_t \in \mathbb{W}$ for simplicity, and leave the analysis for more general bounds, e.g. energy constrained bounds across all stages [25], as future work.

We review the SME algorithm in details below. Consider a sequence of single-trajectory data, $\{x_t, u_t, x_{t+1}\}_{t=0}^{T-1}$, generated from (2), where the horizon T can be unknown beforehand. Let Θ_T denote the remaining uncertainty set of θ^* after the T stages of data are revealed. Θ_T generated by SME, also called a membership set, is defined below:

$$\Theta_T = \bigcap_{t=0}^{T-1} \left\{\hat{\theta} \in \mathbb{R}^{n_x\times n_z} \ :\ x_{t+1} - \hat{\theta} z_t \in \mathbb{W}\right\}. \tag{4}$$

Basically, SME tries to rule out any $\hat{\theta}$ that is inconsistent with the linear dynamics with bounded noises $w_t \in \mathbb{W}$. Notice that $\theta^* \in \Theta_T$ as long as $w_t \in \mathbb{W}$ for all $t \le T-1$.

SME is widely used in robust adaptive control to characterize and reduce the uncertainty sets used in robust controller design. Besides, it is also observed from simulations that SME tends to generate smaller uncertainty sets than LSE's confidence regions (see e.g. [2] and Section VI), which explains the wide applications of SME to some extent.

Despite the applications and good performance of SME, its convergence rate analysis for control dynamical systems remain limited. Our goal in this paper is to further study the convergence rate of SME by measuring the size of uncertainty sets with their diameters as defined below.

Definition 1 (Diameter of a set of matrices)**.** *For a set of matrices $\Theta \in \mathbb{R}^{n_x\times n_z}$, we define its diameter to be diam$(\Theta) := \sup_{\theta_1,\theta_2\in\Theta} ||\theta_1 - \theta_2||_F$.*

In particular, when considering linear dynamical systems, though [1] provides the diameter's convergence guarantees of SME for general convex set $\mathbb{W}$, the best convergence rate in the literature only holds when W is a ℓ_∞ ball [2]. In the following, we are going to generalize the convergence rates in [2] to any convex set $\mathbb{W}$ as in [1].

Assumptions throughout this paper. We will introduce two assumptions that are considered throughout this paper. Firstly, we assume the stochastic properties of w_t and the convexity and compactness of set $\mathbb{W}$.

Assumption 1 (Compactly and convexly supported i.i.d. noise)**.** *The additive noise $\{w_t\}_{t\ge0}$ are identically and independently sampled from a compact and convex noise set $\mathbb{W}$ with a non-empty interior (i.e. $\partial\mathbb{W} \ne \emptyset$) such that $\mathbb{E}(w_t) = 0$ and Cov $(w_t) = \Sigma_w \succ \mathbf{0}$.*

Though SME does not need any stochastic properties of w_t to generate valid uncertainty sets that contain θ^*, the i.i.d. assumption is standard in the literature when discussing SME's convergence rates for linear regression [17], [1], [18], as well as LSE's convergence rate analysis [10], [6]. Besides, the convex and compact $\mathbb{W}$ is also commonly assumed in SME's literature [1], [17], [18], [21].

Our next assumption is based on the block-martingale small-ball (BMSB) condition introduced in [6]. BMSB can be viewed as a stochastic version of the persistent excitation (PE) condition that is commonly assumed in the system identification literature [1], [18], [17], [12]. This is because both BMSB and PE require sufficient exploration in all directions to learn accurate system parameters.

Definition 2 (BMSB condition [6])**.** *For a filtration $\{\mathcal{F}_t\}_{t\geq 1}$ and an $\{\mathcal{F}_t\}_{t\geq 1}$-adapted stochastic process $\{Z_t\}_{t\geq 1}$ such that $Z_t \in \mathbb{R}^d$, we say $\{Z_t\}_{t\geq 0}$ satisfies the (k, Γ_{sb}, p)-BMSB condition for a positive integer k, a positive matrix Γ_{sb}, and $p \in [0,1]$ if, for any fixed unit vector $\lambda \in \mathbb{R}^d$, one has $\frac{1}{k}\sum_{i=1}^{k} \mathbb{P}\left(|\lambda^\top Z_{t+i}| \geq \sqrt{\lambda^\top \Gamma_{sb} \lambda} \middle| \mathcal{F}_t\right) \overset{a.s.}{\geq} p$ for all $t \geq 1$.*

Assumption 2 (Bounded z_t & the BMSB condition)**.** *$\exists\, b_z$ such that $\forall\, t \geq 0,\ ||z_t||_2 \overset{a.s.}{\leq} b_z$. For the filtration $\{\mathcal{F}_t\}_{t=0}^{T-1}$, where $\mathcal{F}_t := \sigma\{w_0, \cdots, w_{t-1}, z_0, \cdots, z_t\}$, the adapte process $\{z_t\}_{t\geq 0}$ satisfies the $(1, \sigma^2 \mathbb{I}_{n_z}, p_z)$-BMSB condition.*

It has been shown in [8] that the BMSB condition can be easily achieved by any stabilizing controller adding an i.i.d. exploration noise with a positive definite covariance, i.e. $u_t = \pi(x_t) + \eta_t$, where η_t is i.i.d. The bounded z_t condition can be naturally satisfied if the controller $\pi(\cdot)$ is bounded-input-bounded-output stabilizing since our disturbances are bounded (Sec 4.7 in [26]).

A classical assumption on tight bound $\mathbb{W}$. Here, we review a classical assumption on the tightness of $\mathbb{W}$: pointwise boundary-visiting noises [1]. This assumption is important for the convergence of SME. Later, we will analyze convergence rates based on this assumption in Section III, then discuss how to relax this assumption in Section IV.

Assumption 3 (Pointwise boundary-visiting noise [1])**.** *$\forall\, \epsilon > 0$, $\exists\, q_w(\epsilon) > 0$ such that $\forall\, t \geq 0$, $\forall\, w^0 \in \partial\mathbb{W}$: $\mathbb{P}\left(|w^0 - w_t| < \epsilon\right) \geq q_w(\epsilon)$.*

On "boundary visiting": With the phrase "boundary visiting," we are not requiring the noise to have a non-vanishing probability to reach exactly the boundary of $\mathbb{W}$, but to visit arbitrarily close to the boundary.

Next, we provide three examples on $q_w(\cdot)$.

Example 1 (Weighted ℓ_∞ ball)**.** *We consider that w_t follows a uniform distribution[3] on $\mathbb{W} = \left\{w \in \mathbb{R}^{n_x} : \max_{i\in[n_x]} \left\{\frac{1}{a_i}|w_i|\right\} \leq 1\right\}$ for positive constants $a_1, \cdots, a_{n_x}$. In this case $q_w(\epsilon) = O(\epsilon^{n_x})$.*

Example 2 (Weighted ℓ_1 ball)**.** *We consider w_t uniformly distributed on $\mathbb{W} = \left\{w \in \mathbb{R}^{n_x} : \sum_{i=1}^{n_x} \frac{1}{a_i}|w_i| \leq 1\right\}$ for positive constants $a_1, \cdots, a_{n_x}$. In this case $q_w(\epsilon) = O(\epsilon^{n_x})$.*

Example 3 (ℓ_2 ball)**.** *We consider w_t uniformly distributed on $\mathbb{W} = \mathbb{B}_2^r(0)$ for fixed $r > 0$.In this case $q_w(\epsilon) = O(\epsilon^{n_x})$.*

III. Convergence Rate under Assumption 3

In this section, we propose a non-asymptotic estimation error bound for the SME under the ϵ-ball boundary-visiting noise Assumption 3 and discuss its implications.

Theorem 1. *Under Assumptions 1, 2 and 3: $\forall\, T > m > 0, \delta > 0$, one has*

$$\mathbb{P}\left(\mathit{diam}(\Theta_T) > \delta\right) \leq \underbrace{\frac{T}{m}\tilde{O}(n_z^{5/2}) a_2^{n_z} \exp(-a_3 m)}_{\textit{Term 1}} + \underbrace{\tilde{O}\left((n_x n_z)^{5/2}\right) a_4^{n_x n_z} [1 - q_w(\frac{a_1\delta}{4})]^{\lceil \frac{T}{m} \rceil - 1}}_{\textit{Term 2}} \tag{5}$$

where $a_1 = \frac{1}{4}\sigma_z p_z$, $a_2 = \max\{1, \frac{64 b_z^2}{\sigma^2 p_z^2}\}$, $a_3 = \frac{1}{8}p_z^2$, $a_4 = \max\{1, \frac{4b_z}{a_1}\}$.

A detailed proof of Theorem 1 can be found in Appendix G. It is similar to the proof of Theorem 2 in Section V (Theorem 2 is more general and will be introduced in Section IV).

On Inequality (5): Inequality (5) provides an upper bound on the probability of the "large diameter event" i.e. $\text{diam}(\Theta_T) > \delta$. Therefore, the smaller the upper bound is, the better the SME performs in terms of estimation accuracy. Notice that Inequality (5) involves two terms: $Term\ 1$ bounds the probability that PE does not hold, and $Term\ 2$ bounds the probability that the uncertainty set is still large even though PE holds (refer to Lemmas 1 and 2 in Section V since the proofs of Theorems 1,2 are based on similar ideas).

On the choices of m: Notice that $Term\ 1$ in (5) decays exponentially in m yet increases with T, while $Term\ 2$ decays exponentially in $\frac{T}{m}$. To ensure a small upper bound in (5), one can choose m at a scale of $\log T \leq m \leq T$ (hiding other constant factors for intuitions here). The best bound induced by (5) is essentially the minimum of the upper bound over m.

Though choosing m seems complicated and requires the knowledge of T, it is worth emphasizing that the choice of m only affects our theoretical bound but does not affect the empirical performance of SME. Therefore, a suboptimal choice of m will only increase the gap between our theory and the actual empirical performance, but will not degrade the empirical performance of SME.

On convergence rates. Theorem 1 can be converted to convergence rates of $\text{diam}(\Theta_T)$. Since (5) depends on $q_w(\cdot)$, to provide explicit bounds as illustrating examples, we consider $q_w(\epsilon) = O(\epsilon^p)$ for $p > 0$, which includes many common distributions, e.g. Examples 1-3.

Corollary 1. *If $q_w(\epsilon) = O(\epsilon^p)$ for any $p \neq 0$, given $m \geq O(n_z + \log T - \log\epsilon)$, then with probability no less than $1 - 2\epsilon$, one has*

$$\mathit{diam}(\Theta_T) \leq \tilde{O}\left(\left(\frac{n_x n_z}{T}\right)^{1/p}\right)$$

Convergence rates for Examples 1-3 Recall that we have shown $q_w(\epsilon) = O\left(\epsilon^{n_x}\right)$ in Examples 1-3. By Corollary 1, we have the convergence rate: $\text{diam}(\Theta_T) \leq \tilde{O}\left(\left(\frac{n_x n_z}{T}\right)^{1/n_x}\right)$. Though this provides valid convergence rate bounds for general convex $\mathbb{W}$, these bounds are much worse than LSE's

[3]Though only uniform distribution is considered, other distributions such as truncated Gaussian can also apply.

convergence rate $\frac{1}{\sqrt{T}}$ and do not explain the promising empirical performance of SME (see Section VI and [2]). Therefore, in the next section, we will seek to improve the convergence rate bounds of SME.

IV. Convergence Rate with Relaxed Assumption

This section provides a relaxed version of Assumption 3, which enables tighter convergence rate bounds in some scenarios. In the following, Section IV-A will introduce this relaxed version in Assumption 4. Then Section IV will discuss the corresponding estimation error bound in Theorem 2, followed by discussions on the differences from Theorem 1 as well as examples.

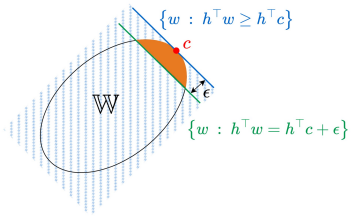
A. A Relaxed Assumption on the Tightness of $\mathbb{W}$

This subsection will introduce a relaxed version of Assumption 3 in Section II. Our assumption relies on supporting half-spaces (SHS) and ϵ-slices induced by SHS for convex sets, so we first review this concept below.

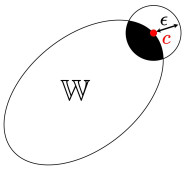
Definition 3 (**Supporting half-spaces & ϵ-slices**). *Consider a convex and compact set $\mathbb{W} \subseteq \mathbb{R}^n$ with a non-empty interior (i.e. $\mathring{\mathbb{W}} \neq \emptyset$). For a boundary point $c \in \partial\mathbb{W}$ and a unit vector h (i.e. $||h||_2 = 1$), we say the half-space*

$$H(c,h) := \left\{x \in \mathbb{R}^n \ : \ h^\top x \geq h^\top c\right\}$$

*is a **supporting half-space (SHS)** of $\mathbb{W}$ at point c with normal vector h, if $\mathbb{W} \subseteq H(c,h)$.*



(a) The ϵ-slice induced by $H(c,h)$.



(b) The ϵ-ball at c.

Fig. 1: (a) and (b) respectively demonstrate the ϵ-slice and the ϵ-neighborhood of a boundary point c. In Figure 1(a), the blue shaded half-space represents the SHS $H(c,h)$, and the orange area is the ϵ-slice induced by $H(c,h)$. In Figure 1(b), the black-filled area is the ϵ-ball in Assumption 3 at c. Observe that with the same ϵ, the ϵ-slice on the left tends to be larger than the ϵ-ball on the right.

*Furthermore, $\forall\, \epsilon > 0$, we define the **ϵ-slice** of $\mathbb{W}$ induced by $H(c,h)$ below:*

$$S^\epsilon_{\mathbb{W}}(c,h) := \left\{x \in \mathbb{R}^n \ : \ h^\top c + \epsilon \geq h^\top x \geq h^\top c\right\} \cap \mathbb{W}.$$

With Definition 3, we can present a relaxed version of Assumption 3 as in Assumption 4 below.

Assumption 4 (ϵ-slice partial-boundary-visiting noise). *There exists a subset of boundary points $\mathcal{C} = \{c_1, \cdots, c_B\} \subseteq \partial\mathbb{W}$[4], and a set of unit vectors $\mathcal{H} = \{h_1, \cdots, h_B\}$ satisfying the following properties.*

(i) $\forall\, i \in \{1, \cdots, B\}$, $H(c_i, h_i)$ is a SHS of $\mathbb{W}$.

(ii) $\bigcap_{i=1}^{B} H(c_i, h_i)$ is a compact set.

(iii) $\forall\, \epsilon > 0$, $\exists\, p_w(\epsilon) > 0$ such that $\forall\, t \geq 0$, $\forall\, i \in \{1, \cdots, B\}$,

$$\mathbb{P}\left(w_t \in S^\epsilon_{\mathbb{W}}(c_i, h_i)\right) \geq p_w(\epsilon) > 0.$$

Assumption 4 is a relaxation of Assumption 3 in two perspectives:

- While Assumption 3 requires a non-vanishing density in the neighborhood of every boundary point, Assumption 4 only requires it on a subset of boundary points.
- The ϵ-ball considered in Assumption 3 is usually a subset of the ϵ-slice considered in Assumption 4 for the same ϵ (see Figure 1 as an example). Therefore, $p_w(\epsilon) \geq q_w(\epsilon)$ in most cases.

To provide more intuitions for Assumption 4, we discuss the three examples in Section II below.

Example 4 (Weighted ℓ_∞ ball). *We consider w_t uniformly sampled from the ℓ_∞ ball in Example 1. With $\mathcal{C}$ including a non-extreme point on each facet of $\mathbb{W}$, and $\mathcal{H}$ including each facet's unit normal vector, the corresponding $p_w(\epsilon)$ is $O(\epsilon)$, which is greater than $q_w(\epsilon)$ in Example 1.*

Example 5 (Weighted ℓ_1 ball). *We consider w_t uniformly sampled from the ℓ_1 ball in Example 2. With $\mathcal{C}$ including a non-extreme point on each facet of $\mathbb{W}$, and $\mathcal{H}$ including each facet's unit normal vector, the corresponding $p_w(\epsilon)$ is $O(\epsilon)$, which is greater than $q_w(\epsilon)$ in Example 2.*

Example 6 (ℓ_2 ball). *We consider w_t uniformly sampled from the ℓ_2 ball in Example 3. $\mathcal{H}$ and $\mathcal{C}$ can be arbitrarily chosen given that $\bigcap_{\mathcal{C},\mathcal{H}} H(c,h)$ is compact, the corresponding $p_w(\epsilon)$ is $O\left(\epsilon^{\frac{n+1}{2}}\right)$, which is of the same order as $q_w(\epsilon)$ in Example 3.*

B. Convergence Rate Analysis under Assumption 4

From now on, we will assume that Assumptions 1, 2, and 4 hold. We provide a new version of the estimation error bound in Theorem 2 based on our relaxed Assumption 4.

Theorem 2. *With Assumptions 1, 2, 4, $\forall\, T > m > 0, \delta > 0$:*

$$\mathbb{P}\left(\mathit{diam}(\Theta_T) > \delta\right) \leq \underbrace{\frac{T}{m}\tilde{O}(n_z^{5/2})a_2^{n_z}\exp(-a_3 m)}_{\textit{Term 1}} + \underbrace{\tilde{O}\left((n_x n_z)^{5/2}\right)a_5^{n_x n_z}[1 - p_w(\frac{a_1\delta\xi}{4})]^{\lceil T/m\rceil - 1}}_{\textit{Term 3}} \tag{6}$$

where $a_5 = \max\{1, \frac{4b_z}{a_1\xi}\}$, and the projection constant $\xi = \min_{||x||_2=1}\max_{h\in\mathcal{H}} h^\top x$.[5]

Differences between Theorems 1 and 2: Both (5) and (6) consist of two terms, where $Term\ 1$ is the same, but the second term is different. $Term\ 3$ differs from $Term\ 2$ in

[4]Though we consider a finite number of boundary points in Assumption 4, the following Theorem 2 remains true when $\mathcal{C}$ is infinite. Meanwhile, Theorem 1 may serve as a special case of when $\mathcal{C}$ is infinite.

[5]In the case where $|\mathcal{H}| = +\infty$, the projection constant is similarly defined by $\xi = \min_{||x||_2=1}\sup_{h\in\mathcal{H}} h^\top x$.

two aspects: (i) $Term\ 3$ depends on $p_w(\cdot)$ while $Term\ 2$ depends on $q_w(\cdot)$, and (ii) there is an additional projection constant ξ in $p_w(\cdot)$ and a_5 of $Term\ 3$. Notice that both p_w and ξ depend on our choices of $\mathcal{C}$ and $\mathcal{H}$, which will be discussed in more details below.

On the projection constant: ξ depends on our choices of $\mathcal{C}$ and $\mathcal{H}$. By definition, we have $\xi := \min_{||x||_2=1} \max_{h\in\mathcal{H}} h^\top x$. It can be shown that $0 < \xi \leq 1$ (See Appendix B). Notice that, by adding more boundary points to $\mathcal{C}$ and more directions to $\mathcal{H}$, we can increase ξ. As an extreme case, if we choose all the points on the boundary and choose all the unit vectors as $\mathcal{H}$, then $\xi = 1$. However, if we choose too many points in $\mathcal{C}$, it might also decrease $p_w(\cdot)$ since it is the uniform lower bound for all points in $\mathcal{C}$. Therefore, there is a tradeoff on the choices of $\mathcal{C}$ and $\mathcal{H}$.

On the choices of $\mathcal{C}$ and $\mathcal{H}$: There is a trade-off on choosing $\mathcal{C}$ and $\mathcal{H}$ to minimize the upper bound in (6). On the one hand, as discussed earlier, since (6) decreases with ξ, it is tempting to add more boundary points to $\mathcal{C}$ to increase ξ to reduce the upper bound in (6). On the other hand, notice that (6) decreases with $p_w(\cdot)$, but having more points in $\mathcal{C}$ might decrease the boundary-visiting probability density $p_w(\cdot)$, which induces a larger upper bound (6). As an example, consider Example 1, if we add a SHS at a degenerate point that is not parallel to any facets of the weighted ℓ_∞ ball, $p_w(\epsilon)$ will decrease to ϵ^{n_x}. Therefore, the best estimation error bound induced by Theorem 2 is by optimizing over the choices of $\mathcal{C}$ and $\mathcal{H}$.

Though it is complicated to optimize the choices of $\mathcal{C}$ and $\mathcal{H}$ in general, it is worth emphasizing that these choices only affect our theoretical bounds and do not affect the empirical performance of SME.

Besides, for some special cases, we have some intuitions on optimizing $\mathcal{C}$ and $\mathcal{H}$. For example, for any polytopic $\mathbb{W}$, we choose an arbitrary non-extreme point as c for every facet of $\mathbb{W}$, then the unit normal vector h at c is unique, and $p_w(\epsilon) = O(\epsilon)$ is the largest possible value.

Convergence rates: Similar to Corollary 1, we discuss the explicit convergence rates of SME induced by Theorem 2 in the following.

Corollary 2. *If $p_w(\epsilon) = O(\epsilon^p)$ for $p > 0$, given $m \geq O(n_z + \log T - \log\epsilon)$, with probability at least $1 - 2\epsilon$, one has*

$$diam(\Theta_T) \leq \tilde{O}\left(\frac{1}{\xi}\left(\frac{n_x n_z}{T}\right)^{\frac{1}{p}}\right)$$

Convergence rates for Examples 4-6: For Examples 4 and 5, we know that $p_w(\epsilon) = O(\epsilon)$ holds for both scenarios. By Corollary 2, we have the convergence rates of both the ℓ_1 and the ℓ_∞ ball supports are $\tilde{O}\left(\frac{n_x n_z}{\xi T}\right)$. Specially, for the ℓ_∞-ball support, the corresponding ξ is $\frac{1}{\sqrt{n_x}}$, which is consistent with the result in [2]. As for Example 6, we have $p_w(\epsilon) = O\left(\epsilon^{\frac{n+1}{2}}\right)$. Hence, for the ℓ_2 ball support, we have a convergence rate of $\tilde{O}\left(\left(\frac{n_x n_z}{T}\right)^{\frac{2}{n+1}}\right)$. The corresponding ξ is 1.

Though only uniform distribution is considered in the above two examples, some other noise types also satisfy the $O(\epsilon)$ boundary assumption (e.g. truncated Gaussian). By Corollary 2, the convergence rates of both the ℓ_∞-ball and the ℓ_1-ball are of $\tilde{O}\left(\frac{n_x n_z}{\xi T}\right)$. In fact, a uniform / truncated Gaussian noise supported on a polytope always results in a $\tilde{O}\left(\frac{n_x n_z}{\xi T}\right)$ convergence rate (simply consider $\mathcal{C}$ to be the set of any non-extreme point on every facet of $\mathbb{W}$). Proofs of Corollary 2 and all the examples can be found in the Appendix (See Appendix A,C,D,E.).

Remark 1. *We first introduce Assumption 3 then Assumption 4 because i) we demonstrate the benefits of Assumption 4 by comparison with the classical one, ii) Assumption 3 and Theorem 1 are simpler, e.g. not involving SHS or ξ.*

V. Proof of Theorem 2

This section presents a proof of Theorem 2. As mentioned earlier, Theorem 1 can be regarded as a special case of Theorem 2 if we allow $B = \infty$. A rigorous proof of Theorem 1 is available in Appendix G. We only focus on the proof of Theorem 2 here in 4 major steps:

(1) We divide the event $\{\text{diam}(\Theta_T) > \delta\}$ by PE and non-PE (see (7)). The probability of non-PE is bounded in [2], so we focus on the event with PE (Lemma 2).
(2) In Section V-B, we divide $\{\text{diam}(\Theta_T) > \delta\}$ with PE (denoted as $\mathcal{E}_1 \cap \mathcal{E}_2$) into sub-events $\{\mathcal{E}_{1,i}\}_{i=1}^{v_\gamma}$ based on a finite discretization of $\mathbb{S}_F$ (Claim 1).
(3) In Section V-C, we further discretize each sub-event $\mathcal{E}_{1,i}$ by time segmentation (see Claim 2), which can be further bounded using Bayes' Rule by utilizing a stopping time $L_{i,k}$ defined in (9) (see Claim 3).
(4) Combining the three Steps above completes the proof.

A. Step 1: Partitioning Based on The Persistent Excitation

Define the estimation error by $\gamma := \hat{\theta} - \theta^*$. $\forall\, t \geq 0$, one has $x_{t+1} - \hat{\theta} z_t = w_t - (\hat{\theta} - \theta^*) z_t$. Define the error set by $\Gamma_T := \{\gamma : \forall 0 \leq t \leq T-1,\ w_t - \gamma z_t \in \mathbb{W}\}$. Notice that Γ_T is attained by translating Θ_T by $-\theta^*$. Hence, $\text{diam}(\Gamma_T) = \text{diam}(\Theta_T)$. In the rest of this proof, we will focus on $\text{diam}(\Gamma_T)$. Next, we define two events as follows:

Definition 4. *Define the event that an large error exists by*

$$\mathcal{E}_1 := \left\{\exists\, \gamma \in \Gamma_T \ s.t.\ ||\gamma||_F \geq \frac{\delta}{2}\right\}$$

For the sake of convenience in the subsequent analysis, let $a_1 = \frac{\sigma_z p_z}{4}$. Define the event of persistence excitation by

$$\mathcal{E}_2 := \left\{\frac{1}{m}\sum_{s=1}^{m} z_{km+s} z_{km+s}^\top \succeq a_1^2 I_{n_z},\ \forall\, 0 \leq k \leq \lceil\frac{T}{m}\rceil - 1\right\}$$

We first show that $\{\text{diam}(\Gamma_T) > \delta\} \subseteq \mathcal{E}_1$. Suppose that $\forall\, \gamma \in \Gamma_T$, $||\gamma||_F < \frac{\delta}{2}$. Then $\forall\, \gamma_1, \gamma_2 \in \Gamma_T$, $||\gamma_1 - \gamma_2||_F \leq ||\gamma_1||_F + ||\gamma_2||_F < \delta$. Taking supremum, we have $\text{diam}(\Gamma_T) = \sup_{\gamma_1,\gamma_2\in\Gamma_T} ||\gamma_1 - \gamma_2||_F \leq \delta$. Thus, $\{\text{diam}(\Gamma_T) > \delta\} \subseteq \mathcal{E}_1$. Hence,

$$\begin{aligned}\mathbb{P}\left(\text{diam}(\Gamma_T) > \delta\right) &\leq \mathbb{P}\left(\mathcal{E}_1\right) = \mathbb{P}\left((\mathcal{E}_1 \cap \mathcal{E}_2) \sqcup (\mathcal{E}_1 \cap \mathcal{E}_2^{\complement})\right) \\ &\leq \mathbb{P}\left(\mathcal{E}_1 \cap \mathcal{E}_2\right) + \mathbb{P}\left(\mathcal{E}_2^{\complement}\right). \end{aligned} \tag{7}$$

The bound on $\mathbb{P}\left(\mathcal{E}_2^{\complement}\right)$ follows directly from [2] as below.

Lemma 1 (Lemma 1 in [2])**.** *Given Assumptions 1, 2, then* $\mathbb{P}\left(\mathcal{E}_2^{\complement}\right) \leq$ *Term 1 in* (6)*.*

Thus, only the following lemma remains to be shown.

Lemma 2. *Given Assumptions 1, 2, and 4, then* $\mathbb{P}\left(\mathcal{E}_1 \cap \mathcal{E}_2\right) \leq$ *Term 3 in* (6)*.*

We prove Lemma 2 by a space-time discretization below.

B. Step 2: Discretization on Space

We take advantage of the theorem from [27] on covering a ball with smaller balls.

Theorem 3 (Ball-covering Theorem in [27])**.** *Consider covering* $\mathbb{B}_F \subseteq \mathbb{R}^{n_x \times n_z}$ *with small balls with radius* $\epsilon_\gamma > 0$*. Denote* v_γ *the minimal number of* ϵ_γ*-balls needed for the covering net* $\mathcal{M}$ *such that* $\forall\, b \in \mathbb{B}_F, \exists\, b_i \in \mathcal{M}$ *with* $||b-b_i||_F \leq 2\epsilon_\gamma$*. Then* $v_\gamma \leq \tilde{O}\left((n_x n_z)^{\frac{5}{2}}\right)\left(\frac{1}{\epsilon_\gamma}\right)^{n_x n_z}$ *if* $\epsilon_\gamma \leq 1$*.*

Denoting $\epsilon_\gamma = \frac{1}{a_5} = \min\{1, \frac{\sigma_z p_z \xi}{16 b_z}\}$, we can find an ϵ_γ-net denoted $\mathcal{M} := \{\gamma_i\}_{i=1}^{v_\gamma}$ where $\forall \bar{\gamma} \in \mathbb{S}_F$, $\exists \gamma_i \in \mathcal{M}$ such that $||\gamma_i - \bar{\gamma}||_F \leq 2\epsilon_\gamma$. We also have:

$$v_\gamma \leq \tilde{O}\left(\left(n_x n_z\right)^{5/2}\right) a_5^{n_x n_z} \tag{8}$$

We define a stopping time

$$L_{i,k} := \min\left\{m+1, \min\left\{l \geq 1 : ||\gamma_i z_{km+l}||_2 \geq a_1\right\}\right\} \tag{9}$$

Define two adapted processes $\{h_{i,t}\}_{t\geq 0}$,$\{c_{i,t}\}_{t\geq 0}$ where

$$h_{i,t} := \arg\max_{h \in \mathcal{H}} h^\top(\gamma_i z_t),$$

and $c_{i,t} \in \mathcal{C}$ is defined to be the boundary point corresponding to $h_{i,t}$. Then we make the following Claim.

Claim 1. $\forall\, i \in \{1, \cdots, v_\gamma\}$*, define*

$$\mathcal{E}_{1,i} := \Bigg\{\exists\, \gamma \in \Gamma_T \text{ such that } \forall\, k \in \left\{0, \cdots, \lceil \frac{T}{m} \rceil - 1\right\}, \\ h_{i,km+L_{i,k}}^\top(\gamma z_{km+L_{i,k}}) \geq \frac{a_1 \delta \xi}{4}\Bigg\}$$

Then $\mathcal{E}_1 \cap \mathcal{E}_2 \subseteq \bigcup_{i=1}^{v_\gamma}\left(\mathcal{E}_{1,i} \cap \mathcal{E}_2\right)$*, and*

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \leq \tilde{O}\left(\left(n_x n_z\right)^{5/2}\right) a_5^{n_x n_z} \max_{i \in \{1, \cdots, v_\gamma\}} \mathbb{P}(\mathcal{E}_{1,i} \cap \mathcal{E}_2).$$

Proof of Claim 1. By $\mathcal{E}_1$, $\exists\, \gamma \in \Gamma_T$ with $||\gamma||_F \geq \frac{\delta}{2}$, define the direction vector by

$$\bar{\gamma} := \frac{\gamma}{||\gamma||_F} \tag{10}$$

Notice that $\forall\, \bar{\gamma} \in \mathbb{S}_F$, $\exists\, \gamma_i \in \mathcal{M}$ such that $||\bar{\gamma} - \gamma_i||_F \leq \frac{2}{a_5}$. By $\mathcal{E}_2$, for any $k \in \left\{0, 1, \cdots, \lceil \frac{T}{m} \rceil - 1\right\}$, we have $\frac{1}{m}\sum_{s=1}^m ||\gamma_i z_{km+s}||_2^2 \geq a_1^2$. By the Pigeonhole Principle, $\forall k \in \left\{0, \cdots, \lceil \frac{T}{m} \rceil - 1\right\}$, $\exists\, L = L(k,i) \in \{1, \cdots, m\}$ such that $||\gamma_i z_{km+l}||_2 \geq a_1$. Hence, $\forall\, i, k$, we have $L_{i,k} = \min L(k,i) \leq m$. Thus, we also have $||\gamma_i z_{km+L_{i,k}}||_2 \geq a_1$. It follows that

$$\begin{aligned}
&h_{i,km+L_{i,k}}^\top(\bar{\gamma} z_{km+L_{i,k}}) \\
=&h_{i,km+L_{i,k}}^\top\left\{\left[\gamma_i - (\gamma_i - \bar{\gamma})\right] z_{km+L_{i,k}}\right\} \\
=&h_{i,km+L_{i,k}}^\top(\gamma_i z_{km+L_{i,k}}) - h_{i,km+L_{i,k}}^\top\left[(\gamma_i - \bar{\gamma}) z_{km+L_{i,k}}\right] \\
\geq&\xi a_1 - ||\gamma_i - \bar{\gamma}||_2 \cdot ||z_{km+L_{i,k}}||_2 \geq \xi a_1 - \frac{2b_z}{a_5} \geq \frac{a_1 \xi}{2}
\end{aligned} \tag{11}$$

By (10) and (11), $\exists\, i \in \{1, \cdots, v_\gamma\}$ such that

$$\forall\, 0 \leq k \leq \lceil \frac{T}{m} \rceil,\ h_{i,km+L_{i,k}}^\top(\gamma z_{km+L_{i,k}}) \geq \frac{a_1 \delta \xi}{4} \tag{12}$$

Consequently, when $\mathcal{E}_1$ holds, there exists $i \in \{1, \cdots, v_\gamma\}$ that $\mathcal{E}_{1,i}$ holds. It follows that $\mathcal{E}_1 \cap \mathcal{E}_2 \subseteq \bigcup_{i=1}^{v_\gamma}\left(\mathcal{E}_{1,i} \cap \mathcal{E}_2\right)$. Consequently, by the sub-additive property of the probability measure, we have $\mathbb{P}\left(\mathcal{E}_1 \cap \mathcal{E}_2\right) \leq \sum_{i=1}^{v_\gamma} \mathbb{P}\left(\mathcal{E}_{1,i} \cap \mathcal{E}_2\right)$. □

With Claim 1 proven, we have discretized the event $\mathcal{E}_1 \cap \mathcal{E}_2$ into sub-events relevant to points on the net $\mathcal{M}$, which leads to a finitely additive upper bound for the probability of $\mathcal{E}_1 \cap \mathcal{E}_2$. We will then focus on bounding the probability of the sub-events from above in the next. Namely, we want to further refine $\mathcal{E}_{1,i} \cap \mathcal{E}_2$ by time segementation.

C. Step 3: Partition on Time

Now we need to further cover the events $\{\mathcal{E}_{1,i} \cap \mathcal{E}_2\}_{i=1}^{v_\gamma}$ to deduce a more explicit bound. Recall Definition 3, we have

$$\mathbb{W} \subseteq \bigcap_{i=1}^{B} H(c_i, h_i). \tag{13}$$

We want to take advantage of Assumption 4 at each stopping time $L_{i,k}$ for each $k \in \left\{0, 1, \cdots, \lceil \frac{T}{m} \rceil - 1\right\}$. The following claim discretizes the event $\mathcal{E}_{1,i} \cap \mathcal{E}_2$ for any $i \in \{1, \cdots, v_\gamma\}$.

Claim 2. $\forall\, 1 \leq i \leq v_\gamma, 0 \leq k \lceil \frac{T}{m} \rceil - 1$*, denote* $G_{i,k} := A_{i,k} \cap \mathcal{E}_2$*, where*

$$A_{i,k} := \left\{h_{i,km+L_{i,k}}^\top\left(w_{km+L_{i,k}} - c_{i,km+L_{i,k}}\right) \geq \frac{a_1 \delta \xi}{4}\right\}.$$

then $\mathcal{E}_{1,i} \cap \mathcal{E}_2 \subseteq \bigcap_{k=0}^{\lceil \frac{T}{m} \rceil - 1} G_{i,k}$*.*

Proof of Claim 2. Recall that by (13) and (4),

$$\forall\, t \geq 0,\ w_t - \gamma z_t = x_{t+1} - \hat{\theta} z_t \in \mathbb{W} \subseteq \bigcap_{i=1}^{B} H(c_i, h_i)$$

Therefore, $\forall\, t \geq 0,\ \forall\, h_j \in \mathcal{H}$, one has $h_j^\top w_t \geq h_j^\top c_j$. For $t = km + L_{i,k}$, $h_j = h_{i,km+L_{i,k}}$, by $\mathcal{E}_{1,i}$, we have, $\forall\, k$:

$$h_{i,km+L_{i,k}}^\top(w_{km+L_{i,k}} - \gamma z_{km+L_{i,k}}) \geq h_{i,km+L_{i,k}}^\top c_{i,km+L_{i,k}}$$

Therefore, when $\mathcal{E}_{1,i}$ holds, we have $h_{i,km+L_{i,k}}^\top(w_{km+L_{i,k}} - c_{i,km+L_{i,k}}) \geq h_{i,km+L_{i,k}}^\top(\gamma z_{km+L_{i,k}}) \geq \frac{a_1 \delta \xi}{4}$ by (12), which is $\bigcap_{k=1}^{\lceil \frac{T}{m} \rceil - 1} A_{i,k}$. □

In the next, we present an upper bound for the relaxed events by the following claim.

Claim 3. *For any* $i \in \{1, \cdots, v_\gamma\}$*:*

$$\mathbb{P}\left(\bigcap_{k=0}^{\lceil \frac{T}{m} \rceil - 1} G_{i,k}\right) \leq \left(1 - p_w\left(\frac{a_1 \delta \xi}{4}\right)\right)^{\lceil \frac{T}{m} \rceil - 1} \tag{14}$$

Proof sketch of Claim 3. Due to page limitation, we present a proof sketch here and defer the complete proof to Appendix F. First, by Bayes' Rule, $\forall\, 1 \le i \le v_\gamma$,

$$\mathbb{P}\left(\bigcap_{k=0}^{\lceil \frac{T}{m} \rceil - 1} G_{i,k}\right) = \mathbb{P}\left(G_{i,0}\right) \prod_{k=1}^{\lceil \frac{T}{m} \rceil - 1} \mathbb{P}\left(G_{i,k} \middle| \bigcap_{\ell=0}^{k-1} G_{i,\ell}\right). \tag{15}$$

Then we bound every term[6] $\mathbb{P}\left(G_{i,k} \middle| \bigcap_{j=0}^{k-1} G_{i,j}\right)$ from above by $1 - p_w\left(\frac{a_1 \delta \xi}{4}\right)$, which finishes the proof. To obtain this, we utilize the stopping time $L_{i,k}$ defined in (9). By Bayes' Rule and the Law of Total Probability, we can show:

$$\mathbb{P}\left(G_{i,k} \middle| \bigcap_{j=0}^{k-1} G_{i,j}\right) \le \sum_{l=1}^{m} \left[\mathbb{P}\left(A_{i,k} \middle| L_{i,k} = l, \bigcap_{j=0}^{k-1} G_{i,j}\right) \times \mathbb{P}\left(L_{i,k} = l \middle| \bigcap_{j=0}^{k-1} G_{i,j}\right)\right] \tag{16}$$

By further utilizing the Bayes' Rule and the Law of Total Probability, it can be shown that

$$\mathbb{P}\left(A_{i,k} \middle| L_{i,k} = l, \bigcap_{j=0}^{k-1} G_{i,j}\right) \le 1 - p_w\left(\frac{a_1 \delta \xi}{4}\right) \tag{17}$$

Claim 3 can be proven by combing (15), (16), and (17). □

D. Step 4: Conclusion & Assembly

By Claim 1 and Claim 2, we conclude that

$$\mathcal{E}_1 \cap \mathcal{E}_2 \subseteq \bigcup_{i=1}^{v_\gamma} \left(\mathcal{E}_{1,i} \cap \mathcal{E}_2\right) \subseteq \bigcup_{i=1}^{v_\gamma} \left[\bigcap_{k=0}^{\lceil \frac{T}{m} \rceil - 1} G_{i,k}\right]$$

Combining (8) and (14), we prove Lemma 2. Ultimately, Theorem 2 is deduced from combining it with Lemma 1.

VI. Numerical Experiments

This section provides numerical examples to test our theoretical bounds and compare SME with LSE's confidence bounds (Theorem 1 [7]) in simulations. We consider a randomly generated example and a linearized Boeing 747 model based on [28].

Experiment settings for LSE: We consider 95% confidence bounds, and use the trace of covariance as a lower bound for the variance proxy L of w_t as shown in [29]. We use $\lambda = 10^{-3}$ as the Ridge regularization parameter.

On the projection constant ξ. Figure 2 considers a randomly generated parameter matrix $\theta^* \in \mathbb{R}^{2\times 4}$, and the noise w_t uniformly distributed on $\mathbb{B}_2$. We consider the SMEs trained using different outer approximations of $\mathbb{B}_2$. Namely, we consider replacing $\mathbb{W}$ in (4) with a regular quadrilateral (4 constraints, yellow curve), a regular octagon (8 constraints, blue curve), and a regular hexadecagon (16 constraints, green curve) circumscribing $\mathbb{B}_2$. In this way, they have same $p_w(\cdot)$ but different ξ so we can visualize the impact of ξ on convergence performance. It can be shown that $0 < \xi_4 < \xi_8 < \xi_{16} < 1$, where ξ_i denotes the ξ for SME with i-constrained $\mathbb{W}$.

[6] $\mathbb{P}\left(G_{i,0}\right)$ can be bounded similarly.

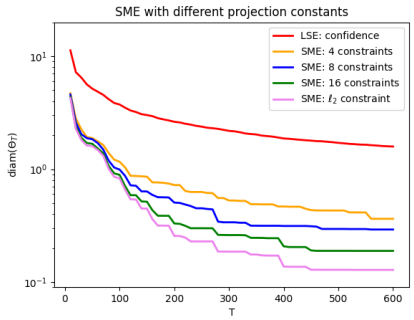


Fig. 2: This plot compares LSE's 95% confidence (red) with SME with different $\mathbb{W}$ that enjoys same $p_w(\cdot)$ but different ξ, i.e. $\xi_4 < \xi_8 < \xi_{16} < 1$. $\mathbb{W}$ for $4/8/16$ constraints respectively correspond to the tightest quadrilateral, octagon, and hexadecagon outer-approximations of $\mathbb{B}_2$ and the pink curve shows SME using exactly $\mathbb{W} = \mathbb{B}_2$.

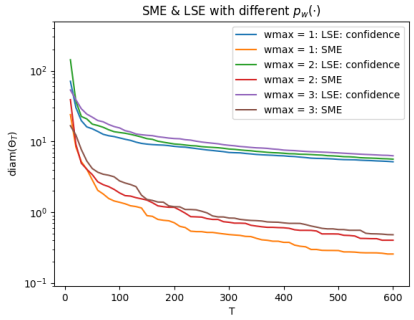
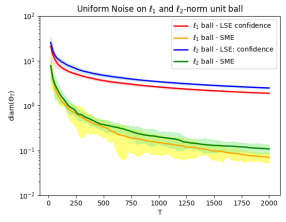


Fig. 3: Comparison of SME and LSE's confidence bounds for truncated Gaussian on ℓ_1 balls with $w_{\max} = 1, 2, 3$.

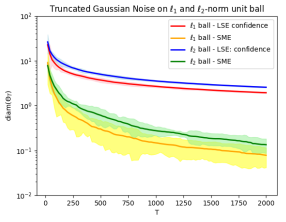
On $p_w(\cdot)$. Figure 3 considers a linearized version of Boeing 747 flight control model [28]. We consider w_t following truncated Gaussian distributions on ℓ_1 balls with the same covariance but different radii $w_{\max} = 1, 2, 3$. We consider that SME knows the exact support of the distributions as $\mathbb{W}$. In this case, different $\mathbb{W}$ has same ξ but different $p_w(\cdot)$. The larger $w_{\max}$ is, the smaller $p_w(\cdot)$ is.

By Theorem 2 and Corollary 2, the estimation error of SME decreases with $p_w(\cdot)$. Notice that this is consistent with Figure 2 because SME with higher $w_{\max}$ provides larger $\text{diam}(\Theta_T)$ due to smaller $p_w(\cdot)$. LSE's confidence bounds follow the same trend because increasing $w_{\max}$ causes larger variances of w_t and thus larger estimation errors in [7].

Discussions on distributions and ℓ_2 balls. Figure 4 also uses the Boeing 747 example. Figure 4's (a) and (b) compare



(a) Uniform distribution.



(b) Truncated Gaussian.

Fig. 4: This plot compares truncated Gaussian and uniform distributions on ℓ_1 and ℓ_2 balls. The plots are based on 5 runs and the shades represent three standard deviations.

SME and LSE with w_t following truncated Gaussian and Uniform distributions. The trends are similar, which is not surprising because these two distributions enjoy the same order of convergence rate according to Corollary 2.

Besides, in Figure 4(a), we compare the differences between ℓ_1 and ℓ_2 balls. The ℓ_2 ball case indeed has a smaller gap between LSE and SME, demonstrating a slightly worse performance, but the performance is still quite similar to the ℓ_1 case. This is quite interesting because our theoretical analysis provides a much worse bound for ℓ_2 case. Therefore, we shows that even though our theoretical bounds successfully predicts the effects of p_w and ξ and explains the good performance of SME on polytopes, our convergence rates for ℓ_2 can still be improved, which is left as our future work.

VII. Conclusions and Future Directions

This paper provides non-asymptotic analysis for set membership estimation for an unknown linear control system with i.i.d. disturbances bounded by general convex sets. We study both the classical assumption in [1] and propose a new relaxed assumption for better bounds.Future directions include: 1) improving the bound for ℓ_2 balls; 2) non-asymptotic analysis of SME with non-stochastic disturbances; 3) proposing and analyzing more computationally efficient SMEs; 4) learning tight bounds $\mathbb{W}$ and its performance bounds; 5) fundamental limits; 6) regrets of robust adaptive controllers using SME, 7) analyzing SME under imperfect state observations, nonlinear systems, etc.

References

[1] X. Lu, M. Cannon, and D. Koksal-Rivet, "Robust adaptive model predictive control: Performance and parameter estimation," *International Journal of Robust and Nonlinear Control*, 2019.

[2] Y. Li, J. Yu, L. Conger, T. Kargin, and A. Wierman, "Learning the uncertainty sets of linear control systems via set membership: A non-asymptotic analysis," *Proceedings of the 41st International Conference on Machine Learning*, vol. 235, pp. 29 234–29 265, 21–27 Jul 2024. [Online]. Available: https://proceedings.mlr.press/v235/li24ci.html

[3] K. J. Åström and P. Eykhoff, "System identification—a survey," *Automatica*, vol. 7, no. 2, pp. 123–162, 1971.

[4] L. Ljung, "Perspectives on system identification," *Annual Reviews in Control*, vol. 34, no. 1, pp. 1–12, 2010.

[5] M. Simchowitz and D. Foster, "Naive exploration is optimal for online lqr," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8937–8948.

[6] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning without mixing: Towards a sharp analysis of linear system identification," in *Conference On Learning Theory*. PMLR, 2018, pp. 439–473.

[7] Y. Abbasi-Yadkori and C. Szepesvári, "Regret bounds for the adaptive control of linear quadratic systems," in *Proceedings of the 24th Annual Conference on Learning Theory*. JMLR Workshop and Conference Proceedings, 2011, pp. 1–26.

[8] Y. Li, T. Zhang, S. Das, J. Shamma, and N. Li, "Non-asymptotic system identification for linear systems with nonlinear policies," *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 1672–1679, 2023.

[9] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "Regret bounds for robust adaptive control of the linear quadratic regulator," *Advances in Neural Information Processing Systems*, vol. 31, pp. 4188–4197, 2018.

[10] S. L. Tu, *Sample complexity bounds for the linear quadratic regulator*. University of California, Berkeley, 2019.

[11] R. L. Kosut, M. K. Lau, and S. P. Boyd, "Set-membership identification of systems with parametric and nonparametric uncertainty," *IEEE Transactions on Automatic Control*, vol. 37, no. 7, pp. 929–941, 1992.

[12] E.-W. Bai, R. Tempo, and H. Cho, "Membership set estimators: size, optimal inputs, complexity and relations with least squares," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 42, no. 5, pp. 266–277, 1995.

[13] M. M. Livstone and M. A. Dahleh, "Asymptotic properties of set membership identification algorithms," *Systems & control letters*, vol. 27, no. 3, pp. 145–155, 1996.

[14] M. Lorenzen, M. Cannon, and F. Allgöwer, "Robust mpc with recursive model update," *Automatica*, vol. 103, pp. 461–471, 2019.

[15] E. Fogel and Y.-F. Huang, "On the value of information in system identification—bounded noise case," *Automatica*, vol. 18, no. 2, pp. 229–238, 1982.

[16] B. T. Lopez, J.-J. E. Slotine, and J. P. How, "Robust adaptive control barrier functions: An adaptive and data-driven approach to safety," *IEEE Control Systems Letters*, vol. 5, no. 3, pp. 1031–1036, 2020.

[17] H. Akçay, "The size of the membership-set in a probabilistic framework," *Automatica*, vol. 40, no. 2, pp. 253–260, 2004.

[18] E.-W. Bai, H. Cho, and R. Tempo, "Convergence properties of the membership set," *Automatica*, vol. 34, no. 10, pp. 1245–1249, 1998.

[19] M. Nayeri, J. Deller Jr, and M. Liut, "Do interpretable optimal bounding ellipsoid algorithms converge? part i-the long-awaited set-convergence proof," *IFAC Proceedings Volumes*, vol. 27, no. 8, pp. 1333–1338, 1994.

[20] O. Smith, O. Cattell, E. Farcot, R. D. O'Dea, and K. I. Hopcraft, "The effect of renewable energy incorporation on power grid stability and resilience," *Science advances*, vol. 8, no. 9, p. eabj6734, 2022.

[21] D. P. Bertsekas, "Control of uncertain systems with a set-membership description of the uncertainty." Ph.D. dissertation, Massachusetts Institute of Technology, 1971.

[22] P. A. Parrilo, M. Sznaier, R. S. Pena, and T. Inanc, "Mixed time/frequency-domain based robust identification," *Automatica*, vol. 34, no. 11, pp. 1375–1389, 1998.

[23] N. Ozay, M. Sznaier, C. M. Lagoa, and O. I. Camps, "A sparsification approach to set membership identification of switched affine systems," *IEEE Transactions on Automatic Control*, vol. 57, no. 3, pp. 634–648, 2011.

[24] X. Lu and M. Cannon, "Robust adaptive model predictive control with persistent excitation conditions," *Automatica*, vol. 152, p. 110959, 2023.

[25] E. Fogel, "System identification via membership set constraints with energy constrained noise," *IEEE Transactions on Automatic Control*, vol. 24, no. 5, pp. 752–758, 1979.

[26] T. Basar, S. Meyn, and W. R. Perkins, "Lecture notes on control system theory and design," *arXiv preprint arXiv:2007.01367*, 2020.

[27] C. Rogers, "Covering a sphere with spheres," *Mathematika*, vol. 10, no. 2, pp. 157–164, 1963.

[28] T. Ishihara, H.-J. Guo, and H. Takeda, "A design of discrete-time integral controllers with computation delays via loop transfer recovery," *Automatica*, vol. 28, no. 3, pp. 599–603, 1992.

[29] J. Arbel, O. Marchal, and H. D. Nguyen, "On strict sub-gaussianity, optimal proxy variance and symmetry for bounded random variables," *ESAIM: Probability and Statistics*, vol. 24, pp. 39–55, 2020.

[30] C. A. Rogers, "Covering a sphere with spheres," *Mathematika*, vol. 10, no. 2, p. 157–164, 1963.

[31] J.-L. Verger-Gaugry, "Covering a ball with smaller equal balls in $\mathbb{R}^n$," *Discrete & Computational Geometry - DCG*, vol. 33, pp. 143–155, 06 2005.

APPENDIX

A. Proof of Corollary 1 and Corollary 2

In this case we have $\tilde{V} = \mathbb{S}_1(0)$, and it follows that $\xi = 1$. We first claim that Term 1 $\leq \epsilon$. To show this, notice that $m \geq O(n_z + \log T - \log \epsilon) = \frac{1}{a_3}\left[O((\log a_2)n_z + \frac{5}{2}\log n_z + O(\log T) - O(\log \epsilon)\right]$. It follows that $\exp(-a_3 m) \leq \frac{a_2^{-n_z} n_z^{-5/2}\epsilon}{T}$. Therefore, Term 1 $\leq \epsilon/m \leq \epsilon$. Next we let Term 2 $= \epsilon$. Then $\epsilon = \tilde{O}\left((n_x n_z)^{5/2}\right)\tilde{a}_4^{n_x n_z}\left[1 - q_w\left(\frac{a_1\delta}{4}\right)\right]^{\lceil \frac{T}{m}\rceil - 1}$. Given $q_w(\epsilon) = O(\epsilon^p)$, we have:

$$\begin{aligned}\delta^p &= O\left(\left(\frac{4}{a_1}\right)^p\right)\left\{1 - \left[\epsilon\tilde{O}\left((n_x n_z)^{5/2}\right)a_4^{n_x n_z}\right]^{\frac{1}{\lceil \frac{T}{m}\rceil - 1}}\right\} \\ &\leq O\left(\left(\frac{4}{a_1}\right)^p\right)\left\{1 - \left[\epsilon\tilde{O}\left((n_x n_z)^{5/2}\right)a_4^{n_x n_z}\right]^{\frac{T}{m}}\right\} \\ &\overset{(b)}{\leq} O\left(\left(\frac{4}{a_1}\right)^p\right)\cdot\frac{m}{T}\cdot\tilde{O}(n_x n_z) = \tilde{O}\left(\left(\frac{4}{a_1}\right)^p\cdot\frac{n_x n_z}{T}\right)\end{aligned}$$

Inequality (b) is deduced from the fact that $\forall\, x > 0,\ x - 1 \geq \log x$. It follows that $\delta \leq \tilde{O}\left(\left(\frac{n_x n_z}{T}\right)^{\frac{1}{p}}\right)$. Then, Corollary 1 is proven. By replacing a_4 with a_5, δ with $\xi\delta$, and $q_w(\epsilon)$ with $q_w(\epsilon)$ in the above proof, we can show Corollary 2.

B. Proof of the strict positivity of ξ

1) $\mathcal{H}$ is finite: Recall that $\xi = \min_{||x||_2=1}\max_{h\in\mathcal{H}} h^\top x$. We show $\forall\, x \in \mathbb{S}_2$, $\max_{h\in\mathcal{H}} h^\top x > 0$ by contradiction. Suppose $\exists x \in \mathbb{S}_2$ such that $\forall\, h \in \mathcal{H}$, $h^\top x \leq 0$. It follows that $\forall w \in \mathbb{W},\ n > 0$, $w - nx \in \mathbb{W}$. This cannot hold since $\mathbb{W}$ is compact.

2) $\mathcal{H}$ is infinite: In this case, recall that $\xi = \min_{||x||_2=1}\sup_{h\in\mathcal{H}} h^\top x$. We prove by contradiction. $\forall\epsilon > 0$, suppose that $\exists\{x_n\}_{n=1}^\infty$ such that $||x||_2 = 1$ and $\forall h \in \mathcal{H}, n \geq 1$, $h^\top x_n \leq \frac{\epsilon}{n}$. Let w be an interior point of $\mathbb{W}$. Then $\forall n \geq 1$, $w + nx_n \in \mathbb{W}$. This cannot hold since $\mathbb{W}$ is assumed to be compact.

C. Proof of Example 1 and Example 4

Recall that the weighted ℓ_∞-norm ball is $\mathbb{W} = \left\{w \in \mathbb{R}^{n_x} : \max_{1\leq i\leq n_x}\left\{\frac{|w_i|}{a_i}\right\} \leq 1\right\}$, and w_t is uniformly distributed on $\mathbb{W}$. The probability density of w_t on $\mathbb{W}$ is $f_w(x) = \frac{1}{2^{n_x}\prod_{i=1}^{n_x} a_i}$.

1) The ϵ-ball: In this case, we consider any vertex w_v of $\mathbb{W}$, the probability that w_t visits $\mathbb{B}_2^\epsilon(w_v) \cap \mathbb{W}$ is $\frac{1}{2^{n_x}}\cdot\frac{\pi^{\frac{n}{2}}\epsilon^{n_x}}{\Gamma\left(\frac{n}{2}+1\right)}\cdot\frac{1}{2^{n_x}\prod_{i=1}^{n_x} a_i} \sim O(\epsilon^{n_x})$.

2) The ϵ-slice: At optimality, we can choose one non-extreme point on each facet of $\mathbb{W}$ as c, and the unit normal vector of this facet at c as h. The ϵ-slice induced by $H(c,h)$ can be then written as $[-a_1, a_1]\times\cdots[a_i - \epsilon, a_i]\times[-a_{i+1}, a_{i+1}]\times[-a_{n_x}, a_{n_x}]$ for some $i \in \{1,\cdots,n_x\}$ (or with $[-a_i, -a_i + \epsilon]$). The probability that w_t visits this ϵ-slice is $\frac{\epsilon\prod_{j\neq i} 2a_j}{2^{n_x}\prod_{j=1}^{n_x} a_j} = \frac{\epsilon}{2a_i} \sim O(\epsilon)$. The according projection constant is $\xi = \frac{1}{n_x}$. Therefore, by Corollary 2, the convergence rate is $\tilde{O}\left(\frac{n_x^{3/2} n_z}{T}\right)$.

D. Proof of Example 2 and Example 5

Recall that the weighted ℓ_1 ball is $\mathbb{W} = \left\{w \in \mathbb{R}^{n_x} : \sum_{i=1}^{n_x}\frac{|w_i|}{a_i} \leq 1\right\}$, and w_t is uniformly distributed on $\mathbb{W}$. The probability density of w_t on $\mathbb{W}$ is $f_w(x) = \frac{n!}{2^n\prod_{i=1}^{n_x} a_i}$.

1) The ϵ-ball: In this case, the least intersection of a boundary ϵ-ball with $\mathbb{W}$ is on one of the vertices of $\mathbb{W}$. Figure 5 demonstrates such an ϵ-ball when $n_x = 2$. Denote

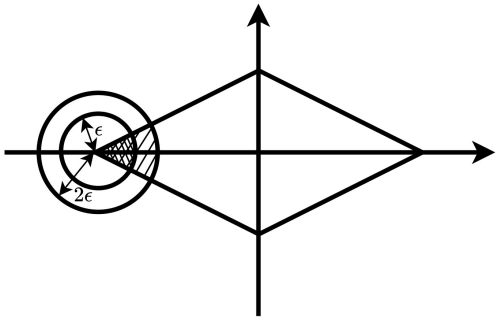


Fig. 5: The shaded area represents the intersection of the 2ϵ-ball with $\mathbb{W}$, while The doubly shaded area represents the intersection of the ϵ-ball with $\mathbb{W}$.

the doubly shaded cone (sector) by S_ϵ. Notice that when we replace ϵ with 2ϵ, the intersection area increases to $2^{n_x}S_\epsilon$. It follows that $q_w(\epsilon) = O\left(\epsilon^{n_x}\right)$.

2) The ϵ-slice: This is similar to Example 4. The best choice of (c,h) is: choose exactly on extreme point of each facet of $\mathbb{W}$ as c, and its according normal vector as h. Any ϵ-slice can be viewed as the intersection of $\mathbb{W}$ and the slab generated by translating a supporting hyperplane by ϵ. Due to the symmetric and parallel property of $\mathbb{W}$, every such ϵ-slice carries a probability measure of $O(\epsilon)$.

E. Proof of Example 3 and Example 6

Recall that the ℓ_2 ball is $\mathbb{W} = \{w \in \mathbb{R}^{n_x} : ||w||_2 \leq 1\}$, and w_t is uniformly distributed on $\mathbb{W}$. The probability density of w_t on $\mathbb{W}$ is $f_w(x) = \frac{\Gamma\left(\frac{n_x}{2}+1\right)}{\pi^{\frac{n_x}{2}}}$.

1) The ϵ-ball: The volume of the intersection of $\mathbb{W}$ and any ϵ-ball is

$$V_{\text{ball}}(\epsilon) = \frac{\pi^{n_x}}{\Gamma\left(\frac{n_x+1}{2}\right)}\left(\int_0^{T_1}\sin^{n_x}\theta d\theta + \epsilon^{n_x}\int_0^{T_2}\sin^{n_x}\theta d\theta\right)$$

where $T_1 = \arccos\left(1 - \frac{\epsilon^2}{2}\right)$, $T_2 = \arccos\left(\frac{\epsilon}{2}\right)$. When ϵ is small, $V_{\text{ball}}(\epsilon) \sim O\left(\epsilon^{n_x}\right)$. It follows that $q_w(\epsilon) = O(\epsilon^{n_x})$.

2) The ϵ-slice: The volume of the ϵ-slice of $\mathbb{W}$ at any boundary point of $\mathbb{W}$ is

$$V_{\text{slice}} = \frac{\pi^{n_x}}{\Gamma\left(\frac{n_x+1}{2}\right)}\int_0^{T_3}\sin^{n_x}\theta d\theta$$

where $T_3 = \arccos(1-\epsilon)$. When ϵ is small enough, we have $V_{\text{slice}}(\epsilon) \sim O\left(\epsilon^{\frac{n_x+1}{2}}\right)$. It follows that $p_w(\epsilon) = O\left(\epsilon^{\frac{n_x+1}{2}}\right)$.

F. Complete Proof of Claim 3

Notice that by the Bayes' Theorem, $\forall\, i$:

$$\mathbb{P}\left(\bigcap_{k=0}^{\lceil\frac{T}{m}\rceil-1} G_{i,k}\right)=\mathbb{P}\left(G_{i,0}\right)\prod_{k=1}^{\lceil\frac{T}{m}\rceil-1}\mathbb{P}\left(G_{i,k}\middle|\bigcap_{\ell=0}^{k-1} G_{i,\ell}\right)$$

We look into an arbitrary term[7] $\mathbb{P}\left(G_{i,k}\middle|\bigcap_{j=0}^{k-1} G_{i,j}\right)$. Recall that when the system is persistently excited (i.e. when $\mathcal{E}_2$ holds), we have: $1\le L_{i,k}\le m$ for all i,k. Hence, $G_{i,k}\subseteq A_{i,k}\cap\{1\le L_{i,k}\le m\}$. It follows that

$$\begin{aligned}
&\mathbb{P}\left(G_{i,k}\middle|\bigcap_{j=0}^{k-1} G_{i,j}\right)\\
\le&\sum_{l=1}^{m}\mathbb{P}\left(A_{i,k};\ L_{i,k}=l\middle|\bigcap_{j=0}^{k-1} G_{i,j}\right)\\
\le&\sum_{l=1}^{m}\left[\mathbb{P}\left(A_{i,k}\middle|L_{i,k}=l,\bigcap_{j=0}^{k-1} G_{i,j}\right)\mathbb{P}\left(L_{i,k}=l\middle|\bigcap_{j=0}^{k-1} G_{i,j}\right)\right]
\end{aligned}\tag{18}$$

Meanwhile, notice that

$$\begin{aligned}
&\mathbb{P}\Bigg(h_{i,km+L_{i,k}}^{\top}\left(w_{km+L_{i,k}}-c_{i,km+L_{i,k}}\right)\ge\frac{a_1\delta\xi}{4}\\
&\quad\Bigg|L_{i,k}=l,\bigcap_{j=0}^{k-1} G_{i,j}\Bigg)\\
\overset{(e)}{=}&\int_{Q_{k,l}}\left[\mathbb{P}\left(h_{i,km+L_{i,k}}^{\top}\left(w_{km+L_{i,k}}-c_{i,km+L_{i,k}}\right)\ge\frac{a_1\delta\xi}{4}\middle|\right.\right.\\
&\quad\left.\left.w_{0:km+l}\right)\cdot\mathbb{P}\left(w_{0:km+l}\middle|L_{i,k}=l,\bigcap_{j=0}^{k-1} G_{i,j}\right)\right]dw_{0:km+l}\\
\le&\left(1-q_w\left(\frac{a_1\delta\xi}{4}\right)\right)\cdot\int_{Q_{k,l}}\mathbb{P}\Bigg(w_{0:km+l}\\
&\quad\Bigg|L_{i,k}=l,\bigcap_{j=0}^{k-1} G_{i,j}\Bigg)dw_{0:km+l}=1-q_w\left(\frac{a_1\delta\xi}{4}\right)
\end{aligned}\tag{19}$$

Here, (e) is deduced by the Law of Total Probability and the Bayes' Rule. Specially, denote a sequence of noise by

$$w_{0:km+l}:=\{w_0,w_1,\cdots,w_{km+l-1}\}$$

and the domain of integration in (e) by

$$Q_{k,l}:=\left\{w_{0:km+l}\text{ such that }L_{i,k}=l\text{ and }\bigcap_{j=1}^{k-1} G_{i,j}\text{ hold}\right\}$$

[7]By a procedure similar to the following reasoning, we can also deduce that $\mathbb{P}\left(G_{i,0}\right)\le 1-q_w\left(\frac{a_1\delta\xi}{4}\right)$.

Combining inequalities (18) and (19), we can induce the following result:

$$\begin{aligned}
&\mathbb{P}\left(G_{i,k}\middle|\bigcap_{j=0}^{k-1} G_{i,j}\right)\\
\le&\left(1-q_w\left(\frac{a_1\delta\xi}{4}\right)\right)\sum_{l=1}^{m}\mathbb{P}\left(L_{i,k}=l\middle|\bigcap_{j=0}^{k-1} G_{i,j}\right)\\
=&1-q_w\left(\frac{a_1\delta\xi}{4}\right)
\end{aligned}$$

It follows that

$$\mathbb{P}\left(\bigcap_{k=0}^{\lceil\frac{T}{m}\rceil-1} G_{i,k}\right)\le\left(1-q_w\left(\frac{a_1\delta\xi}{4}\right)\right)^{\lceil\frac{T}{m}\rceil-1}\tag{20}$$

G. Proof of Theorem 1

We present a proof of Theorem 1. Similar to the proof of Themreom 2, the reasoning can be divided into 4 major steps:

(1) We firstly bound the event $\{\text{diam}(\Theta_T)>\delta\}$ (the "large diameter event") by partitioning it into 2 circumstances: either with or without the persistent excitation (PE) condition. Since the bound for the probability of the absence of the PE condition is identical to its well-elaborated counterpart in [2], it suffices to focus on the scenario where PE is present.

(2) By discretizing the unit sphere $\mathbb{S}_1^F(0)$ in the $n_x\times n_z$-dimensional space equipped with the Frobenius norm, we cover the "large diameter event" by the union of a sequence of events where each of them is related to the realization of a point, at a stopping time during each PE window, in the underlined discretization. The sub-additive property of the probability measure allows us to bound the "large diameter event" but the sum of these discrete events' probabilities.

(3) For each of the points in the discretization, and each window of persistent excitation, we establish a probabilistic relation between the noise realization and a certain boundary point of $\mathbb{W}$, at the corresponding stopping time. Using trivial Bayesian tricks, we can bound each term of the summation in Step (2) from above by a product of conditioning probabilities.

(4) Combining all the 3 Steps above, we can finally assemble the upper bound in Theorem 1.

We propose the detailed proof for each of the 4 steps in the following 4 subsections.

1) Partitioning Based on The Persistent Excitation: Recall that in Section V-A, it is defined that

$$\mathcal{E}_1:=\left\{\exists\,\gamma\in\Gamma_T\ s.t.\ ||\gamma||_F\ge\frac{\delta}{2}\right\}$$

$$\mathcal{E}_2:=\left\{\frac{1}{m}\sum_{s=1}^{m} z_{km+s}z_{km+s}^{\top}\succeq a_1^2 I_{n_z},\,\forall\,0\le k\le\lceil\frac{T}{m}\rceil-1\right\}$$

Also, we have

$$\mathbb{P}\left(\text{diam}(\Theta_T)>\delta\right)\le\mathbb{P}\left(\mathcal{E}_1\cap\mathcal{E}_2\right)+\mathbb{P}\left(\mathcal{E}_2^{\complement}\right)$$

Theorem 1 can then be shown by combing Lemma 1 and Lemma 3. That is, we will concentrate on finding the bound for $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2)$. That is, it remains to be shown that

Lemma 3. *If Assumptions 1, 2, and 3 hold, then*

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \leq \textit{Term 2 in } (5)$$

We will prove Lemma 3 in the next three subsections.

2) Discretization on The Space: Despite the partition above, it remains difficult to bound the "$\exists$" statement in the event $\mathcal{E}_1 \cap \mathcal{E}_2$ due to the continuity of the space $\mathbb{R}^{n_x \times n_z}$. Hence, we want to further refine the event by covering it with a collection of smaller events corresponding to a discrete net on the space. For the $n_x \times n_z$-dimensional unit Frobenius-norm sphere $\mathbb{S}_1^F(0) := \{\bar{\gamma} \in \mathbb{R}^{n_x \times n_z} \ : \ ||\bar{\gamma}||_F = 1\}$, we consider covering it with smaller balls with radius $\tilde{\epsilon}_\gamma = \frac{1}{a_5} = \min\left\{1, \frac{\sigma_z p_z}{16 b_z}\right\}$ where σ_z, p_z, b_z are defined in Assumption 2, and denote the corresponding $\tilde{\epsilon}_\gamma$-net to be $\tilde{\mathcal{M}} := \{\tilde{\gamma}_i\}_{i=1}^{\tilde{v}_\gamma}$. Here, $\tilde{v}_\gamma$ is the number of small $\tilde{\epsilon}_\gamma$-balls required to cover the sphere (i.e. $\forall\, \bar{\gamma} \in \mathbb{S}_1^F(0)$, $\exists \tilde{\gamma}_i \in \tilde{\mathcal{M}}$ such that $||\tilde{\gamma}_i - \bar{\gamma}||_F \leq 2\tilde{\epsilon}_\gamma$). For the theory of covering number, we refer the readers to [30], [31] and Appendix D.1 in [2]. Here we have:

$$\tilde{v}_\gamma \leq \tilde{O}\left((n_x n_z)^{5/2}\right) a_5^{n_x n_z}$$

We define a stopping time

$$\tilde{L}_{i,k} := \min\left\{m+1, \min\left\{l \geq 1 \ : \ ||\tilde{\gamma}_i z_{km+l}||_2 \geq a_1\right\}\right\}$$

$\forall\, t \geq 0$, define the adapted process $\{\tilde{v}_{i,t}\}_{t \geq 0}$

$$\tilde{v}_{i,t} := \arg\max_{v \in \mathbb{S}_1^2(0)} v^\top (\tilde{\gamma}_i z_t)$$

where $\mathbb{S}_1^2(0) := \{w \in \mathbb{R}^{n_x} \ : \ ||w||_2 = 1\}$. Since z_t is $\mathcal{F}_t$-measurable and $\{\tilde{L}_{i,k} = \ell\} \in \mathcal{F}_{km+\ell}$, then $\tilde{v}_{i,t}$ is $\mathcal{F}_t$-measurable, and $\tilde{L}_{i,k}$ a stopping time with respect to the original filtration. Notice that since $\mathbb{S}_1^2(0)$ is compact, the maximizer is well-defined. However, it may not be unique. We can take arbitrary one of such qualified v.

In the next, we want to discretize the event $\mathcal{E}_1 \cap \mathcal{E}_2$ by splitting it into sub-events regarding the $\frac{1}{a_5}$-net of $\mathbb{S}_1^F(0)$, using the sub-additivity property of the probability measure. We have the following claim.

Claim 4. *$\forall\, i \in \{1, \cdots, \tilde{v}_\gamma\}$, define*

$$\tilde{\mathcal{E}}_{1,i} := \{\exists\, \gamma \in \Gamma_T \textit{ such that } \forall\, k \in \{0, \cdots, T/m-1\}, \ \tilde{v}_{i,km+\tilde{L}_{i,k}}^\top (\gamma z_{km+\tilde{L}_{i,k}}) \geq \frac{a_1 \delta}{4}\}$$

Then

$$\mathcal{E}_1 \cap \mathcal{E}_2 \subseteq \bigcup_{i=1}^{\tilde{v}_\gamma} \left(\tilde{\mathcal{E}}_{1,i} \cap \mathcal{E}_2\right)$$

Thus,

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \leq \tilde{O}\left((n_x n_z)^{5/2}\right) a_5^{n_x n_z} \max_{1 \leq i \leq \tilde{v}_\gamma} \mathbb{P}(\tilde{\mathcal{E}}_{1,i} \cap \mathcal{E}_2) \tag{21}$$

Proof of Claim 4. To discretize, $\forall\, \gamma \in \mathbb{R}^{n_x \times n_z} \setminus \{0\}$, denote

$$\bar{\gamma} := \frac{\gamma}{||\gamma||_F}$$

Notice that $\forall\, \bar{\gamma} \in \mathbb{S}_1^F(0)$, $\exists\, \tilde{\gamma}_i \in \tilde{\mathcal{M}}$ such that $||\bar{\gamma} - \tilde{\gamma}_i||_F \leq \frac{2}{a_5}$. By $\mathcal{E}_2$, for any $k \in \{0, 1, \cdots, T/m-1\}$

$$\frac{1}{m}\sum_{s=1}^{m} ||\tilde{\gamma}_i z_{km+s}||_2^2 \geq a_1^2$$

By the Pigeonhole Principle, $\forall k \in \{0, \cdots, T/m-1\}$, $\exists\, l = l(k,i) \in \{1, \cdots, m\}$ such that $||\tilde{\gamma}_i z_{km+l}||_2 \geq a_1$. Hence, we have

$$\forall\, i, k, \ \tilde{L}_{i,k} = \min l(k,i) \leq m$$

Moreover, since $\tilde{L}_{i,k}$ is the minimum of all $l(k,i)$, it also satisfies the following inequality.

$$||\tilde{\gamma}_i z_{km+\tilde{L}_{i,k}}||_2 \geq a_1$$

It follows that

$$\begin{aligned}
&\tilde{v}_{i,km+\tilde{L}_{i,k}}^\top (\bar{\gamma} z_{km+\tilde{L}_{i,k}}) \\
=&\tilde{v}_{i,km+\tilde{L}_{i,k}}^\top \left\{[\tilde{\gamma}_i - (\tilde{\gamma}_i - \bar{\gamma})]\, z_{km+\tilde{L}_{i,k}}\right\} \\
=&\tilde{v}_{i,km+\tilde{L}_{i,k}}^\top (\tilde{\gamma}_i z_{km+\tilde{L}_{i,k}}) - \tilde{v}_{i,km+\tilde{L}_{i,k}}^\top \left[(\tilde{\gamma}_i - \bar{\gamma}) z_{km+\tilde{L}_{i,k}}\right] \\
\geq& a_1 - ||\tilde{\gamma}_i - \bar{\gamma}||_2 \cdot ||z_{km+\tilde{L}_{i,k}}||_2 \\
\geq& a_1 - \frac{2b_z}{a_5} \\
\geq& \frac{a_1}{2}
\end{aligned} \tag{22}$$

Likewise, if $\exists \gamma \in \Gamma_T$ such that $||\gamma||_F \geq \frac{\delta}{2}$, then it can be written as $||\gamma||_F \bar{\gamma}$ for some $\bar{\gamma} \in \mathbb{S}_1^F(0)$. The above reasoning leads to that such a γ also satisfies the following: by (22), there exists $i \in \{1, \cdots, \tilde{v}_\gamma\}$ such that $\forall k \in \{0, \cdots, T/m-1\}$

$$\tilde{v}_{i,km+\tilde{L}_{i,k}}^\top (\gamma z_{km+\tilde{L}_{i,k}}) \geq \frac{a_1 \delta}{4} \tag{23}$$

Therefore, by the sub-additive property of the probability measure,

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \leq \sum_{i=1}^{\tilde{v}_\gamma} \mathbb{P}\left(\tilde{\mathcal{E}}_{1,i} \cap \mathcal{E}_2\right)$$

It follows that $\forall\, i \in \{1, \cdots, \tilde{v}_\gamma\}$,

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \leq \tilde{O}\left((n_x n_z)^{5/2}\right) a_5^{n_x n_z} \max_{1 \leq i \leq \tilde{v}_\gamma} \mathbb{P}(\tilde{\mathcal{E}}_{1,i} \cap \mathcal{E}_2)$$

□

3) Covering by Discretization on The Time:

Claim 5. *$\forall\, i \in \{1, \cdots, \tilde{v}_\gamma\}, k \in \{0, \cdots, \lceil \frac{T}{m} \rceil - 1\}$, we denote*

$$w_{i,km+\tilde{L}_{i,k}}^0 := \arg\min_{w \in \mathbb{W}} \tilde{v}_{i,km+\tilde{L}_{i,k}}^\top w$$

and

$$\tilde{G}_{i,k} := \left\{||w_{km+\tilde{L}_{i,k}} - w_{i,km+\tilde{L}_{i,k}}^0||_2 \geq \frac{a_1 \delta}{4}\right\} \cap \mathcal{E}_2$$

then

$$\tilde{\mathcal{E}}_{1,i} \cap \mathcal{E}_2 \subseteq \bigcap_{k=0}^{\lceil \frac{T}{m} \rceil - 1} \tilde{G}_{i,k}$$

Proof of Claim 5. Recall that by the algorithm 4, $\forall\, t \geq 0$,

$$w_t - \gamma z_t = x_{t+1} - \hat{\theta} z_t \in \mathbb{W}$$

Since $\mathbb{W}$ is convex, compact, and has a non-empty interior, by the Supporting Hyperplane Theorem, $\mathbb{W}$ can be represented with all its supporting hyperplanes. If we denote

$$h(v) := \min_{w \in \mathbb{W}} v^\top w$$

then we can represent $\mathbb{W}$ in the following manner.

$$\mathbb{W} = \{w \in \mathbb{R}^{n_x} \ : \ \forall v \in \mathbb{S}_1^2(0), v^\top w \geq h(v)\}$$

Therefore, $\forall v \in \mathbb{S}_1^2(0)$, $v^\top(w_t - \gamma z_t) \geq h(v)$. Looking at when $t = km + \tilde{L}_{i,k}$, $v = \tilde{v}_{i,km+\tilde{L}_{i,k}}$, we have

$$\tilde{v}_{i,km+\tilde{L}_{i,k}}^\top (w_{km+\tilde{L}_{i,k}} - \gamma z_{km+\tilde{L}_{i,k}}) \geq h(\tilde{v}_{i,km+\tilde{L}_{i,k}})$$

It follows that, when $\tilde{\mathcal{E}}_{1,i}$ holds, by (23), we have

$$\begin{aligned} &\tilde{v}_{i,km+\tilde{L}_{i,k}}^\top w_{km+\tilde{L}_{i,k}} - h(\tilde{v}_{i,km+\tilde{L}_{i,k}}) \\ \geq &\tilde{v}_{i,km+\tilde{L}_{i,k}}^\top (\gamma z_{km+\tilde{L}_{i,k}}) \\ \geq &\frac{a_1 \delta}{4} \end{aligned} \tag{24}$$

Since $\mathbb{W}$ is compact and $v^\top w$ is a linear functional of w with any fixed v, then some of its minimizers for $h(v)$ must be boundary points of $\mathbb{W}$. For all $i \in \{1, \cdots, \tilde{v}_\gamma\}$, $k \in \{0, \cdots, \lceil \frac{T}{m} \rceil - 1\}$, consider

$$\{w_{i,km+\tilde{L}_{i,k}}^0\} \in \arg\min_{w \in \mathbb{W}} \tilde{v}_{i,km+\tilde{L}_{i,k}}^\top w \cap \partial \mathbb{W} \tag{25}$$

Though the minimizer may not be unique, yet we can take any one of them. It follows that

$$h(\tilde{v}_{i,km+\tilde{L}_{i,k}}) = \tilde{v}_{i,km+\tilde{L}_{i,k}}^\top w_{i,km+\tilde{L}_{i,k}}^0$$

By the Cauchy-Schwartz Inequality, we have

$$\begin{aligned} \frac{a_1 \delta}{4} &\overset{(a)}{\leq} |\tilde{v}_{i,km+\tilde{L}_{i,k}}^\top (w_{km+\tilde{L}_{i,k}} - w_{i,km+\tilde{L}_{i,k}}^0)| \\ &\overset{(b)}{\leq} ||\tilde{v}_{i,km+\tilde{L}_{i,k}}||_2 \cdot ||w_{km+\tilde{L}_{i,k}} - w_{i,km+\tilde{L}_{i,k}}^0||_2 \\ &\overset{(c)}{\leq} ||w_{km+\tilde{L}_{i,k}} - w_{i,km+\tilde{L}_{i,k}}^0||_2 \end{aligned}$$

Here, (a) is deduced by (24) and (25); (b) is from the Cauchy-Schwartz Inequality; (c) is implied by the fact that $||\tilde{v}_{i,km+\tilde{L}_{i,k}}||_2 = 1$. Thus, if we define events $\{\tilde{G}_{i,k}\}_{k=0}^{T/m-1}$ in the following manner

$$\tilde{G}_{i,k} := \left\{ ||w_{km+\tilde{L}_{i,k}} - w_{i,km+\tilde{L}_{i,k}}^0||_2 \geq \frac{a_1 \delta}{4} \right\} \cap \mathcal{E}_2$$

Then we can deduce that $\forall\, i \in \{1, \cdots, \tilde{v}_\gamma\}$

$$\tilde{\mathcal{E}}_{1,i} \cap \mathcal{E}_2 \subseteq \bigcap_{k=0}^{\lceil \frac{T}{m} \rceil - 1} \tilde{G}_{i,k}$$

□

Claim 6. *For any $i \in \{1, \cdots, \tilde{v}_\gamma\}$:*

$$\mathbb{P}\left(\bigcap_{k=0}^{\lceil \frac{T}{m} \rceil - 1} \tilde{G}_{i,k} \right) \leq \left(1 - q_w\left(\frac{a_1 \delta}{4}\right)\right)^{\lceil \frac{T}{m} \rceil - 1}$$

Proof of Claim 6. To find an upper bound for the probability of $\bigcap_{k=0}^{\lceil \frac{T}{m} \rceil - 1} \tilde{G}_{i,k}$, notice that by the Bayes' Theorem, $\forall\, i$:

$$\mathbb{P}\left(\bigcap_{k=0}^{\lceil \frac{T}{m} \rceil - 1} \tilde{G}_{i,k} \right) = \mathbb{P}\left(\tilde{G}_{i,0}\right) \prod_{k=1}^{\lceil \frac{T}{m} \rceil - 1} \mathbb{P}\left(\tilde{G}_{i,k} \middle| \bigcap_{\ell=0}^{k-1} \tilde{G}_{i,\ell} \right)$$

We look into an arbitrary factor $\left(\text{i.e. } \mathbb{P}\left(\tilde{G}_{i,k} \middle| \bigcap_{\ell=0}^{k-1} \tilde{G}_{i,\ell} \right)\right)$ of the product on the right hand side. Recall that when the system is persistently excited (i.e. when $\mathcal{E}_2$ holds), we have: $1 \leq \tilde{L}_{i,k} \leq m$ for all i, k. Hence, $\tilde{G}_{i,k} \implies \left\{ ||w_{km+\tilde{L}_{i,k}} - w_{i,km+\tilde{L}_{i,k}}^0||_2 \geq \frac{a_1 \delta}{4};\ 1 \leq \tilde{L}_{i,k} \leq m \right\}$. Thus, for each of these factors:

$$\begin{aligned} &\mathbb{P}\left(\tilde{G}_{i,k} \middle| \bigcap_{j=0}^{k-1} \tilde{G}_{i,j} \right) \\ \leq &\mathbb{P}\left(||w_{km+\tilde{L}_{i,k}} - w_{i,km+\tilde{L}_{i,k}}^0||_2 \geq \frac{a_1 \delta}{4}; \right. \\ &\left. 1 \leq \tilde{L}_{i,k} \leq m \middle| \bigcap_{j=0}^{k-1} \tilde{G}_{i,j} \right) \\ = &\sum_{l=1}^{m} \mathbb{P}\left(||w_{km+l} - w_{i,km+l}^0||_2 \geq \frac{a_1 \delta}{4};\ \tilde{L}_{i,k} = l \middle| \bigcap_{j=0}^{k-1} \tilde{G}_{i,j} \right) \\ \leq &\sum_{l=1}^{m} \left[\mathbb{P}\left(||w_{km+l} - w_{i,km+l}^0||_2 \geq \frac{a_1 \delta}{4} \middle| \tilde{L}_{i,k} = l, \bigcap_{j=0}^{k-1} \tilde{G}_{i,j} \right) \right. \\ &\left. \times \mathbb{P}\left(\tilde{L}_{i,k} = l \middle| \bigcap_{j=0}^{k-1} \tilde{G}_{i,j} \right) \right] \end{aligned}$$

Meanwhile, notice that

$$\begin{aligned} &\mathbb{P}\left(||w_{km+l} - w_{i,km+l}^0||_2 \geq \frac{a_1 \delta}{4} \middle| \tilde{L}_{i,k} = l, \bigcap_{j=0}^{k-1} \tilde{G}_{i,j} \right) \\ \overset{(d)}{=} &\int_{v_{0:km+l}} \mathbb{P}\left(||w_{km+l} - w_{i,km+l}^0||_2 \geq \frac{a_1 \delta}{4}, \right. \\ &\left. w_{0:km+l} = v_{0:km+l} \middle| \tilde{L}_{i,k} = l, \bigcap_{j=0}^{k-1} \tilde{G}_{i,j} \right) dv_{0:km+l} \\ \overset{(e)}{=} &\int_{\tilde{Q}_{k,l}} \left[\mathbb{P}\left(||w_{km+l} - w_{i,km+l}^0||_2 \geq \frac{a_1 \delta}{4} \middle| w_{0:km+l} = v_{0:km+l} \right) \right. \end{aligned}$$

$$\times\mathbb{P}\left(w_{0:km+l}=v_{0:km+l}\middle|\tilde{L}_{i,k}=l,\bigcap_{j=0}^{k-1}\tilde{G}_{i,j}\right)\Bigg]dv_{0:km+l}$$

$$\leq\left(1-q_w\left(\frac{a_1\delta}{4}\right)\right)\times$$

$$\int_{\tilde{Q}_{k,l}}\mathbb{P}\left(w_{0:km+l}=v_{0:km+l}\middle|\tilde{L}_{i,k}=l,\bigcap_{j=0}^{k-1}\tilde{G}_{i,j}\right)dv_{0:km+l}$$

$$=1-q_w\left(\frac{a_1\delta}{4}\right)$$

Here, (d) is deduced by the Law of Total Probability, and (e) comes from the Bayes' Rule. Specially, we denote a sequence of noise (or its realization) by

$$w_{0:km+l}:=\{w_0,w_1,\cdots,w_{km+l-1}\}$$

and the domain of integration by

$$\tilde{Q}_{k,l}:=\{v_{0:km+l}\text{ such that }w_{0:km+l}=v_{0:km+l}\text{ satisfies }\tilde{L}_{i,k}=l\text{ and }\bigcap_{j=1}^{k-1}\tilde{G}_{i,j}\}$$

The above implications induce the following result:

$$\mathbb{P}\left(\tilde{G}_{i,k}\middle|\cap_{j=0}^{k-1}\tilde{G}_{i,j}\right)$$

$$\leq\left(1-q_w\left(\frac{a_1\delta}{4}\right)\right)\sum_{l=1}^{m}\mathbb{P}\left(\tilde{L}_{i,k}=l\middle|\cap_{j=0}^{k-1}\tilde{G}_{i,j}\right)$$

$$=1-q_w\left(\frac{a_1\delta}{4}\right)$$

By a similar procedure, we can also deduce that

$$\mathbb{P}\left(\tilde{G}_{i,0}\right)\leq 1-q_w\left(\frac{a_1\delta}{4}\right)$$

It follows that

$$\mathbb{P}\left(\bigcap_{k=0}^{\lceil\frac{T}{m}\rceil-1}\tilde{G}_{i,k}\right)\leq\left(1-q_w\left(\frac{a_1\delta}{4}\right)\right)^{\lceil\frac{T}{m}\rceil-1}\tag{26}$$

□

4) The Final Assembly: By combining (21) and (26), we proved Lemma 3 which bounds the term $\mathbb{P}\left(\mathcal{E}_1\cap\mathcal{E}_2\right)$ from above, and in Lemma 1 we found an upper bound for the term $\mathbb{P}\left(\mathcal{E}_2^{\complement}\right)$. By combining Lemmas 1 and 3, we finish the proof of Theorem 1.