Nearly Instance-Optimal Parameter Recovery from Many Trajectories via Hellinger Localization

Eliot Shekhtman[*1], Yichen Zhou[*2], Ingvar Ziemann[1], Nikolai Matni[1], and Stephen Tu[2]

[1]Department of Electrical and Systems Engineering, University of Pennsylvania
[2]Department of Electrical and Computer Engineering, University of Southern California

October 7, 2025

Abstract

Learning from sequential, temporally-correlated data is a core facet of modern machine learning and statistical modeling. Yet our fundamental understanding of sequential learning remains incomplete, particularly in the multi-trajectory setting where data consists of many independent realizations of a time-indexed stochastic process. This important regime both reflects modern training pipelines such as for large language and vision-language models, and offers the potential for learning without the typical mixing assumptions (e.g., geometric/polynomial ergodicity) made in the classical single-trajectory case. However, sharp instance-optimal bounds are known only for least-squares regression problems with dependent covariates [1, 2]; for more general models or loss functions, the only broadly applicable guarantees result from a simple reduction to either (i) i.i.d. learning, with effective sample size scaling only in the number of trajectories, or (ii) an existing single-trajectory result when each individual trajectory mixes, with effective sample size scaling as the full data budget deflated by a factor of the mixing-time.

In this work, we significantly broaden the scope of instance-optimal rates for parameter recovery in multi-trajectory settings via the *Hellinger localization framework*, a general approach for maximum likelihood estimation. Our method proceeds by first controlling the squared Hellinger distance at the *path-measure* level via a reduction to i.i.d. learning, followed by localization as a quadratic form in parameter space weighted by the trajectory-level Fisher information matrix. This yields instance-optimal parameter recovery bounds that scale with the full data budget, i.e., number of trajectories times trajectory length, under a broad set of conditions. We instantiate our framework across diverse case studies, including a simple mixture of Markov chains example, dependent linear regression under non-Gaussian noise (i.e., non-square losses), generalized linear models with non-monotonic activations, and linear-attention sequence models. In all cases, our parameter recovery bounds nearly match the instance-optimal rates implied by asymptotic normality, substantially improving over bounds from standard reductions.

---

[*]Equal contribution, order chosen randomly.

Contents

1 Introduction

Learning from sequential data is central to modern machine learning (ML) and statistical modeling, underpinning applications such as language modeling [3], speech recognition [4], time-series forecasting [5], generalist robotics [6], neurological sequence analysis [7], and many other examples. Yet, despite its importance and prevalence, our fundamental understanding of when learning from sequential streams is possible—and what the sharp, problem-specific sample complexities are—remains far less developed compared with the classical i.i.d. setting.

There are two predominant approaches to analyzing non-i.i.d. sequential learning setups. The first approach is to consider consuming data from a single stochastic process indexed by time t, and study what happens to the estimator as time progresses forward. In this paper, we will refer to a realization from a time-index stochastic process as a trajectory, and hence we denote this approach as the *single-trajectory* setting. This setting is challenging, as simple examples illustrate the necessity of imposing non-trivial assumptions about the long-running behavior of the underlying process. The strongest results hold under assumptions about the rate of convergence (typically quantified via its mixing-time [8]) of the stochastic process to its time-marginal distributions, and provide bounds on e.g., the excess risk of the empirical risk minimizer (ERM) as time t moves forward (see e.g., [9–12]). However, these type of results suffer from some key drawbacks: (i) Many processes—e.g., human dialogue, periodic locomotion gaits, wearable sensor data—are interesting precisely because they do *not* mix. Even for processes that do mix, analytical bounds on the mixing-time can be quite conservative in high-dimension [13], and challenging to estimate numerically without assuming extra structure [14]. (ii) Typical bounds suffer from *sample deflation*, where the non-i.i.d. sequential rate is a factor of the mixing-time larger than its corresponding i.i.d. rate (i.e., its effective sample size is deflated by the mixing-time); furthermore, the non-i.i.d. rates are only valid after time t exceeds some factor of the mixing-time, typically referred to as a burn-in time.

The second approach to studying sequential learning sits in-between the classic i.i.d. setting, where every data point is independent, and the single-trajectory setting, where every data point is correlated. Instead, one assumes that many independent realizations of a stochastic process are observed, a setup we will refer to as the *multi-trajectory* setting (see e.g., [1, 15–17]). In the multi-trajectory setting, data within a trajectory is temporally correlated as usual, but importantly, data across different trajectories is independent. The latter fact is crucial, as it enables reductions to i.i.d. learning only using minimal assumptions about the underlying process. A further benefit of the multi-trajectory data model is that it is a much more accurate description of the underlying training data for modern ML models which ingest sequential data—such as vision-language models (VLMs) [18], large language models (LLMs) [3], and generalist behavior policies for robotics [6]—compared with the single-trajectory model.

Nevertheless, despite these advantages, many open questions still remain regarding the multi-trajectory model. A naïve reduction to i.i.d. learning, where each trajectory is treated as a single data point, only yields rates where the effective sample size is the m, the number of trajectories. Here, the length T of each trajectory[1] is absent from this sample size, which is in general not the correct scaling. On the other hand, embedding a multi-trajectory process into a single trajectory and appealing to a single-trajectory result yields either (i) similar bounds as the i.i.d. reduction if no assumption on the mixing-time of the individual trajectories is made, or (ii) bounds where the effective sample size scales as the mT/κ—i.e., *total* number of data points available to the

---

[1]We assume for ease of exposition that each trajectory has the same length.

learner divided by the mixing-time κ of the individual trajectories. While (ii) improves upon the i.i.d. reduction, the deflation factor is in general not optimal for multi-trajectory settings. Indeed, a recent line of work [1, 2, 19] shows that for square-loss regression from dependent covariates, one can obtain—under a set of conditions which do not generally require e.g., bounded mixing-times—finite-sample rates where the effective sample size not only improves to mT, but also the bounds nearly match those prescribed by asymptotic normality of maximum likelihood estimation (MLE). However, as their proof techniques are tailored specifically for square-loss, it is unclear how to generalize these results more broadly to e.g., MLE settings.

In this work, we significantly broaden our understanding of learning in the multi-trajectory setting by providing a general framework—which we call the *Hellinger localization framework*—for deriving sharp instance-optimal parameter recovery rates for general maximum-likelihood estimation. At a high-level, our framework proceeds in two main phases: In the first phase, we utilize a reduction to i.i.d. learning which controls the squared Hellinger distance between the *path measures* of the MLE estimate and the underlying distribution, at a rate where the sample size is m, the number of trajectories. In the second phase, we utilize the fact that squared Hellinger distance is locally quadratic in the parameter space, weighted by the Fisher information matrix of the underlying path measure. This has two key consequences: First, it allows us to extract out an additional scaling factor of T, the length of each trajectory, whenever the process contains sufficient excitation. Second, the Fisher information matrix allows us to derive instance-specific rates that match, up to logarithmic factors, those prescribed by asymptotic normality of MLE. Our framework thus yields instance-optimal bounds where the effective sample size contains all the observed data (i.e., scales as mT), and applies broadly to maximum likelihood estimation problems. Furthermore, as our framework relies only on bounded growth conditions of both the score function and the observed information matrix for localization, it applies beyond the usual mixing processes and stable dynamics typically assumed in sequential learning.

To demonstrate the generality of our approach, we instantiate the Hellinger localization framework in four case studies: (i) a simple mixture of Markov chains example, (ii) a dependent linear regression setting under general (i.e., non-Gaussian) product-noise distributions, (iii) a non-monotonic generalized linear model (GLM) example, and (iv) a linear-attention [20] sequence modeling problem. For each of these case studies, our framework obtains near-optimal parameter recovery error rates, yielding significant improvements over the rates obtained by standard reductions.

Paper organization. This manuscript is organized as follows. In Section 2, we review related work. Section 3 describes the multi-trajectory MLE problem setup, reviews the standard i.i.d. and single-trajectory reductions in more detail, and presents the Hellinger localization framework, including a step-by-step guide describing how to instantiate the framework for a general problem. Section 4 contains the results of our four specific case studies; for each case study, we also conduct a more thorough literature review on the specific problem beyond what is described in Section 2. Section 5 concludes the paper, with the appendices containing the deferred proofs.

2 Related Work

We first review the relevant literature for learning from non-i.i.d. sequential data in the single-trajectory (single realization of a time-indexed stochastic process) setting; as already discussed briefly and will be reviewed in more detail in Section 3.1, single-trajectory results can generally be

used to analyze multi-trajectory settings. The most common approach to analyzing single-trajectory learning is to impose mixing-time assumptions on the underlying process (see e.g., [9–12, 21–23]); as a result, excess-risk or parameter recovery rates are typically a factor of mixing-time worse than their corresponding i.i.d. rates, as the standard blocking technique [9] can only utilize one data point in every mixing-time size chunk. Recently, there have been a few improvements for various problem settings. A line of work studying parameter recovery in linear dynamical systems allows for the transition matrix A to be marginally stable (i.e., $\rho \leqslant 1$) [24] or even unstable (i.e., $\rho > 1$) [25, 26]; both situations correspond to unbounded mixing-times, with the former marginally stable setting also extended for a class of GLMs [27]. For realizable non-linear regression problems with square loss, [19] shows that assuming a certain trajectory-level hyper-contractivity condition holds, the sample size deflation in the excess risk bound can actually be removed, leaving the mixing-time dependence to only the burn-in time. The work [2] further improves upon this result by obtaining variance-optimal rates under a (weakly) sub-Gaussian class (cf. [28]) assumption; as this result is important context for our work, we review it in detail in Section 3.1. Despite improvements, these works either (a) only apply to a limited class of models, or (b) require the underlying process to mix, to satisfy additional technical assumptions (e.g., trajectory hyper-contractivity or sub-Gaussian class) that can be difficulty to verify, and only apply for the square-loss.

Next, we address literature directly studying the multi-trajectory setting (see e.g., [1, 15–17, 29–33]). Most relevant to our work is [1], which studies parameter recovery for linear least-squares regression from dependent covariates, and derives instance-optimal rates that scale with the full dataset size mT while requiring no stability/mixing assumptions; we review these results in more detail in Section 3.1. While this work also provides important context and motivation for our study, their proof techniques—self-normalized martingales [34] and extensions of small-ball inequalities [35]—are tailored for the specific closed-form structure of the linear least-squares regression solution, and do not admit obvious extensions to more general setups. On the other hand, our approach is based more on information-theoretic concepts, building on a combination of techniques for analyzing density estimation [36–38], in conjunction with locally quadratic expansions of f-divergences [39] (specifically, the squared Hellinger distance in our setting).

We next briefly remark on other recent progress in learning from non-i.i.d. data sources. One line of work studies learning either regression functions or filters from the perspective of online learning and regret minimization [40–43], constructing a online predictor of future observations that is competitive over a family of fixed predictors given perfect hindsight knowledge. We view our results as complementary to this line of work—an interesting question for future research is to study these regret minimization formulations in settings where multi-trajectory data is revealed online to the player. Another line of work considers non-temporal data correlations, specifically learning Ising models from either a single sample (e.g., [44, 45]), or multiple independent samples (e.g., [46, 47]). While these problem setups are not directly comparable, we believe it is interesting future work to study whether our techniques can also be applied in such a setting.

Finally, the case studies we consider in Section 4 are special cases and/or natural extensions of problem setups previously considered in the literature; we provide detailed overview of problem-specific related work for each case study in its corresponding sub-section.

3   Problem Setup and General Framework

In this section, we review background, outline our general problem formulation and describe our new Hellinger localization framework. We first describe the notation used in our work.

Notation.   For a vector $x \in \mathbb{R}^d$, we let $\|x\|_p$ denote its ℓ_p norm; for $p = 2$, we drop the subscript, i.e., $\|x\| = \|x\|_2$. The notation $x^{\otimes 2} = xx^\mathsf{T}$ is shorthand for the outer product matrix. Also, the notation $\mathrm{diag}(x) \in \mathbb{R}^{d \times d}$ is the diagonal matrix satisfying $\mathrm{diag}(x)_{ii} = x_i$ for $i \in [d]$. Given a positive definite matrix $\Sigma \in \mathbb{R}^{d \times d}$, we let $\|x\|_\Sigma = \sqrt{x^\mathsf{T} \Sigma x}$ denote its weighed ℓ_2 norm. For a matrix $M \in \mathbb{R}^{d \times k}$, we let $\|M\|_{\mathrm{op}}$, $\|M\|_F$ denote its operator (maximum singular value) and Frobenius norm, respectively. If $d = k$ and M is positive semi-definite, we let $M^{1/2}$ denote its PSD square root. If $M = M^\mathsf{T}$ is symmetric, we let the eigenvalues of M be denoted as $\lambda_i(M)$, $i \in [d]$, listed in non-increasing order. The notation $\mathrm{vec}(M) \in \mathbb{R}^{dk}$ denotes vectorization of M; we follow the convention that vectorization is done in column-order, so that for size conforming matrices A, M, and B, the identity $\mathrm{vec}(AMB) = (B^\mathsf{T} \otimes A)\mathrm{vec}(M)$ holds, where $\otimes$ denotes the Kronecker product. We use $\mathrm{mat}(\cdot)$ to denote the inverse of $\mathrm{vec}(\cdot)$, i.e., $\mathrm{mat}(\mathrm{vec}(M)) = M$; the output dimension $d \times k$ of $\mathrm{mat}(\cdot)$ will be implicit from context. Given a real-valued random variable X, we let $\|X\|_{\mathcal{L}^p(\rho)} = (\mathbb{E}_\rho[|X|^p])^{1/p}$ denote the $\mathcal{L}^p(\rho)$ norm. For a measure μ, we let $\mu^{\otimes k}$ to denote its k-fold product measure. Finally, the unit sphere in $\mathbb{R}^d$ is denoted $\mathbb{S}^{d-1} \coloneqq \{x \in \mathbb{R}^d \mid \|x\| = 1\}$.

3.1   Maximum Likelihood Estimation in Multi-Trajectory Settings

We fix an index $T \in \mathbb{N}_+$ and consider a stochastic process $z_{1:T} \coloneqq (z_t)_{t=1}^T$ taking values in Z. Let $p_\star(z_{1:T})$ denote the joint distribution over $z_{1:T}$. We emphasize that the process $z_{1:T}$ is not necessarily stationary nor ergodic, nor does it necessarily have bounded mixing-times. Our learner observes $m \in \mathbb{N}_+$ independent trajectories $\mathcal{D}_{m,T} \coloneqq ((z_t^{(i)})_{t=1}^T)_{i=1}^m$ with each $z_{1:T}^{(i)} \sim p_\star(\cdot)$. Fix a parametric class $\mathcal{P} \coloneqq \{p_\theta(z_{1:T}) \mid \theta \in \Theta\}$ of distributions and consider the maximum-likelihood (MLE) estimator $\hat{p}_{m,T} \in \mathcal{P}$ given the dataset $\mathcal{D}_{m,T}$ as:

$$\hat{p}_{m,T} \in \arg\max_{p_\theta \in \mathcal{P}} \sum_{i=1}^m \log p_\theta(z_{1:T}^{(i)}). \tag{3.1}$$

In this work, we are interested in the finite-sample behavior of the MLE estimator $\hat{p}_{m,T}$ in the *realizable* setting, i.e., where $p_\star \in \mathcal{P}$. We impose some regularity conditions to make the analysis well-posed. First, we endow Z with a base measure μ (e.g., counting measure for discrete Z or Lebesgue measure when Z is a subset of Euclidean space), and we overload p_θ to also denote the Radon-Nikodym density w.r.t. the corresponding base measure μ on Z. We also assume that (a) for $\mu^{\otimes T}$-a.e. $z_{1:T} \in \mathsf{Z}^T$, the map $\theta \mapsto p_\theta(z_{1:T})$ is $C^2(\Theta_0)$ where $\Theta_0 \supseteq \Theta$ is an open set, (b) that there exists a unique $\theta_\star \in \Theta$ such that $p_{\theta_\star}(\cdot) = p_\star(\cdot)$ (almost surely), and that (c) the parameter set Θ is star-convex around $\theta_\star$.[2] To set the stage for our results, let us first review what is known about the MLE estimator $\hat{p}_{m,T}$.

Asymptotic normality.   First, we can use the lens of asymptotic normality to understand limiting behavior as $m \to \infty$ (but T is fixed). To do this, we recall that the Fisher information (FI) matrix

---

[2]That is for any $\theta \in \Theta$, $s\theta_\star + (1-s)\theta \in \Theta$ for all $s \in [0,1]$. As our analysis relies on local quadratic expansions, this technical assumption ensures that such expansions are indeed valid.

for the *trajectory* $z_{1:T}$ is defined as:

$$\mathcal{I}(\theta) := -\mathbb{E}_{z_{1:T} \sim p_\theta} \left[\nabla_\theta^2 \log p_\theta(z_{1:T}) \right] = \mathbb{E}_{z_{1:T} \sim p_\theta} \left[\nabla_\theta \log p_\theta(z_{1:T})^{\otimes 2} \right]. \tag{3.2}$$

Under the assumption that $\theta_\star \in \mathrm{int}(\Theta)$ and that the estimator $\hat{\theta}_{m,T}$ is consistent (i.e., $\hat{\theta}_{m,T} \to \theta_\star$ a.s. as $m \to \infty$), standard asymptotic normality [see e.g., 48] for M-estimators yields:

$$\sqrt{m} \cdot \mathcal{I}(\theta_\star)^{1/2} (\hat{\theta}_{m,T} - \theta_\star) \overset{\mathrm{d}}{\rightsquigarrow} \mathsf{N}(0, I_p), \tag{3.3}$$

where $\overset{\mathrm{d}}{\rightsquigarrow}$ denotes convergence in distribution. The condition (3.3) depends implicitly on the trajectory length T through the FI matrix $\mathcal{I}(\theta_\star)$. However, as the FI matrix $\mathcal{I}(\theta)$ factorizes nicely across time:

$$\mathcal{I}(\theta) = -\mathbb{E}_{p_\theta} \left[\nabla_\theta^2 \log p_\theta(z_{1:T}) \right] = -\sum_{t=1}^{T} \mathbb{E}_{p_\theta} \left[\nabla_\theta^2 \log p_\theta(z_t \mid z_{1:t-1}) \right],$$

we generically expect that $\mathcal{I}(\theta)$ grows at least linearly with T, i.e., $\lambda_{\min}(\mathcal{I}(\theta)) \geqslant \Omega(T)$. Hence, if we define the *normalized* Fisher information matrix $\bar{\mathcal{I}}(\theta) := T^{-1} \cdot \mathcal{I}(\theta)$, the result (3.3) implies that the limiting behavior as $m \to \infty$ scales with high probability as:

$$\|\hat{\theta}_{m,T} - \theta_\star\|_{\bar{\mathcal{I}}(\theta_\star)}^2 \lesssim \frac{p}{mT} \quad \text{and} \quad \|\hat{\theta}_{m,T} - \theta_\star\|^2 \lesssim \frac{p}{mT \cdot \lambda_{\min}(\bar{\mathcal{I}}(\theta_\star))}. \tag{3.4}$$

Therefore, as long as the normalized FI matrix provides sufficient excitation so that $\lambda_{\min}(\bar{\mathcal{I}}(\theta_\star))$ does *not* vanish to zero as the trajectory length T increases, then (3.4) implies that the squared parameter error decreases at a $1/(mT)$ rate, a rate which not involves *all* the data points available in the training set, but as importantly is also *instance-optimal*, containing instance-specific scaling through the FI matrix $\bar{\mathcal{I}}(\theta_\star)$. Hence, showing that (3.4) holds in a non-asymptotic, finite number of trajectories regime under general conditions serves as one of the main goals of this work.

As we will discuss in detail in the remainder of this sub-section, finite-sample rates of the form (3.4) are known to hold for the setting of least-squares regression over dependent covariates under various assumptions [1, 2]; unfortunately, these analysis techniques heavily utilize the structure of the square-loss, and do not readily extend to more general losses such as the log-loss for MLE. Beyond the square-loss, the most general rates comes from reductions to either (i) standard i.i.d. learning results or (ii) existing single-trajectory results; the former yields rates exhibiting sub-optimal $1/m$ scaling, whereas the latter inherits the single-trajectory stability assumptions that are often unnecessary in the multi-trajectory case, and suffers from sample-deflation issues in the rates. This motivates the need for developing a new approach for establishing finite-sample instance-optimal rates for the multi-trajectory setting, which we turn to in Section 3.2.

Linear least-squares regression and linear system identification. One case where a non-asymptotic rate of the form (3.4) is shown to hold in the literature is in the setting of linear least-squares regression over dependent covariates [1]. Specifically, consider the following linear dynamical system (LDS) parameterized by $A \in \mathbb{R}^{d \times d}$:

$$z_{t+1} = A z_t + w_t, \quad w_t \sim \mathsf{N}(0, \sigma^2 I_d), \quad z_0 = 0. \tag{3.5}$$

For any $\theta = \mathrm{vec}(A) \in \mathbb{R}^{d^2}$, the FI matrix $\mathcal{I}(\theta)$ takes on the form:

$$\mathcal{I}(\theta) = \frac{1}{\sigma^2} \left(\sum_{t=1}^{T-1} \Sigma_t(\mathrm{mat}(\theta)) \right) \otimes I_d, \quad \Sigma_t(A) = \mathbb{E}_{z_t \sim p_A} [z_t z_t^\mathsf{T}] = \sigma^2 \sum_{s=0}^{t-1} A^s (A^s)^\mathsf{T}.$$

Hence letting $\hat{A}_{m,T} \in \arg\min_{A \in \mathbb{R}^{d \times d}} \sum_{i=1}^{m} \sum_{t=1}^{T-1} \|z_{t+1}^{(i)} - Az_t^{(i)}\|^2$ denote the MLE (3.1) and $A_\star$ denote the true dynamics matrix generating the data via (3.5), plugging the FI matrix into (3.4) yields

$$\|\hat{A}_{m,T} - A_\star\|_{\Gamma_T(A_\star)}^2 \lesssim \frac{\sigma^2 d^2}{mT} \quad \text{and} \quad \|\hat{A}_{m,T} - A_\star\|_F^2 \lesssim \frac{\sigma^2 d^2}{mT \cdot \lambda_{\min}(\Gamma_T(A_\star))}, \tag{3.6}$$

where $\Gamma_t(A) = \frac{1}{t-1} \sum_{s=1}^{t-1} \Sigma_s(A)$. In [1], it is shown that this rate (3.6) holds in expectation whenever $m \gtrsim d$, and that this cut-off is sharp. Similar results hold for the more general linear regression from LDS covariates. We emphasize here that the rate from (3.6) is truly a multi-trajectory phenomenon, and is *not* possible for arbitrary $A_\star$ from a single trajectory; as shown in [25], the MLE is not generally consistent in the single-trajectory setting when $A_\star$ is unstable.

Reductions to existing i.i.d./single-trajectory results. Beyond the linear least-squares regression setting, a simple generic approach for deriving rates in the multi-trajectory setting is to invoke an existing i.i.d. and/or single-trajectory result. As an example of an i.i.d. reduction, using the standard Rademacher complexity machinery for deriving risk bounds in independent settings [49],[3] we have that the excess risk (in KL-divergence)

$$\mathsf{ER}(\theta) \coloneqq \frac{1}{T}\mathrm{KL}(p_\star(z_{1:T}) \parallel p_\theta(z_{1:T})) - \frac{1}{T} \inf_{\theta' \in \Theta} \mathrm{KL}(p_\star(z_{1:T}) \parallel p_{\theta'}(z_{1:T})) \tag{3.7}$$

of the MLE $\hat{\theta}_{m,T}$ satisfies with probability at least $1 - \delta$,

$$\mathsf{ER}(\hat{\theta}_{m,T}) \leqslant \frac{2}{T} \sum_{t=1}^{T} \mathcal{R}_m(\mathcal{G}_t) + c_0 B \sqrt{\frac{\log(2/\delta)}{m}}, \tag{3.8}$$

where B bounds the log-likelihood $T^{-1} \cdot \log p_\theta(z_{1:T})$ a.s., $\mathcal{R}_m(\mathcal{G}_t)$ is the Rademacher complexity of the function class $\mathcal{G}_t \coloneqq \{z_{1:T} \mapsto \log p_\theta(z_t \mid z_{1:t-1}) \mid \theta \in \Theta\}$, and c_0 is a universal constant. Assuming that each conditional $|\log p_\theta(z_t \mid z_{1:t-1})| \leqslant O(1)$, then we have $B \leqslant O(1)$ as well. We also generically expect that $\mathcal{R}_m(\mathcal{G}_t) \leqslant O(\sqrt{p/m})$, which is the usual rate for parametric function classes. Furthermore, $\frac{1}{T}\mathrm{KL}(p_\star(z_{1:T}) \parallel p_{\hat{\theta}_{m,T}}(z_{1:T})) \approx \frac{1}{2}\|\theta_\star - \hat{\theta}_{m,T}\|_{\mathcal{I}(\theta_\star)}^2$ asymptotically as $\hat{\theta}_{m,T} \to \theta_\star$, and hence the scaling w.r.t. T in the excess risk (3.7) is the correct one for comparison to (3.4). Therefore, the general scaling for the RHS of (3.8) is of order $\sqrt{p/m}$, i.e., the effective sample size is the number of trajectories m. Note that in the realizable setting when $\inf_{\theta' \in \Theta} \mathrm{KL}(p_\star(z_{1:T}) \parallel p_{\theta'}(z_{1:T})) = 0$, the bound for (3.8) can be improved to a fast-rate p/m scaling with local Rademacher complexities [51].

Single-trajectory results can also be used for reductions, by embedding the trajectories $\{z_{1:T}^{(i)}\}$ into one single trajectory $\bar{z}_{1:mT} \coloneqq (z_{1:T}^{(1)}, \ldots, z_{1:T}^{(m)})$. For this discussion, we focus on results relying on β-mixing[4] for concreteness, noting that our discussion also applies to results that rely on other definitions of mixing (e.g., ϕ-mixing) in the literature. We also assume the process $\{z_t\}$ is Markovian, as this makes the reduction simpler to state. Because $z_{1:T}^{(i)} \perp z_{1:T}^{(j)}$ whenever $i \neq j$, then we have that the β-mixing coefficients $\bar{\beta}(k)$ of $\{\bar{z}_t\}$ satisfy $\bar{\beta}(k) = \beta(k) \cdot \mathbb{1}\{k < T\}$, where $\beta(k)$ are the β-mixing

---

[3]For analyzing MLE, there are much sharper non-asymptotic analysis in the i.i.d. setting (e.g., [50]), which do not require almost sure bounds on log-likelihoods, contain the correct variance-optimal scaling, and also capture fast-rates in realizable settings. We present the simplest result here to make our point clear.

[4]The β-mixing coefficients (cf. [8, 9]) for $\{z_t\}_{t=1}^{\infty}$ are defined as $\beta(k) \coloneqq \sup_{j \in \mathbb{N}_+} \mathbb{E}_{z_{1:j}}[\|\mathbb{P}_{z_{j+k:\infty}}(\cdot \mid z_{1:j}) - \mathbb{P}_{z_{j+k:\infty}}\|_{\mathrm{TV}}]$. The process is called β-mixing if $\beta(k) \to 0$ as $k \to \infty$.

coefficients of $\{z_t\}$. Hence, the embedded trajectory $\{\bar{z}_t\}$ is trivially β-mixing with mixing-time equal to T without any assumption on $\beta(k)$. Using this mixing-time and invoking a single-trajectory β-mixing result, such as from [12], without any further assumption on $\beta(k)$ yields a similar result to (3.8). However, if we further assume that $\beta(k) \leqslant C \exp(-\rho k)$ for some $\rho > 0$, and that $T/(2\kappa) \in \mathbb{N}_+$ for $\kappa := \lceil \rho^{-1} \log(CmT/\delta) \rceil$, then we have the improved result: with probability at least $1 - \delta$,

$$
\mathsf{ER}(\hat{\theta}_{m,T}) \leqslant \frac{c_0'}{2\kappa} \sum_{j=1}^{2\kappa} \bar{\mathcal{R}}^j_{mT/\kappa} + c_1' B \sqrt{\frac{\kappa \log(c_2'/\delta)}{mT}}, \tag{3.9}
$$

where $\bar{\mathcal{R}}^j_{mT/\kappa}$ denotes the *de-coupled* Rademacher complexity:

$$
\bar{\mathcal{R}}^j_{mT/\kappa} := \mathbb{E} \sup_{\theta \in \Theta} \frac{\kappa}{mT} \sum_{i=1}^m \sum_{\ell=1}^\kappa \varepsilon_{i,\ell} \log p_\theta(\tilde{z}^{(i)}_{(\ell-1)2\kappa+j} \mid \tilde{z}^{(i)}_{(\ell-1)2\kappa+j-1}), \quad j \in [2\kappa],
$$

with the pair $(\tilde{z}^{(i)}_{(\ell-1)2\kappa+j-1}, \tilde{z}^{(i)}_{(\ell-1)2\kappa+j})$ drawn from the same distribution as $(z_{(\ell-1)2\kappa+j-1}, z_{(\ell-1)2\kappa+j})$, but *independently* across $i \in [m]$ and $\ell \in [\kappa]$. Similar to before, we generically expect that $\bar{\mathcal{R}}^j_{mT/\kappa}$ scales as order $\sqrt{\kappa p/(mT)}$. Hence, the general scaling of the RHS of (3.9) is of order $\sqrt{\kappa p/(mT)}$. Furthermore, as with the i.i.d. reduction, in the realizable setting local Rademacher arguments can also be used to improve the scaling of (3.9) to the fast-rate $\kappa p/(mT)$. This is an improvement over (3.8), as the effective sample size increases from m to mT/κ; however, this sample size still remains deflated by κ, as a consequence of the standard blocking technique used for de-coupling.

Regression with square-loss. A recent line of work [2, 19] has shown that the sample size deflation described previously can be removed in the special case of non-linear regression with the square-loss, in both parametric and non-parametric regimes. To make their results concrete, we consider the following parametric family of distribution $\mathcal{P}$ over trajectories in $\mathbb{R}^d$:

$$
z_{t+1} = f_\theta(z_t) + w_t, \quad w_t \sim \mathsf{N}(0, \sigma^2 I_d), \tag{3.10}
$$

coupled with the non-linear least-squares estimator $\hat{\theta}_{m,T} \in \arg\min_{\theta \in \Theta} \sum_{i=1}^m \sum_{t=1}^{T-1} \|z^{(i)}_{t+1} - f_\theta(z^{(i)}_t)\|^2$ with $\Theta \subseteq \mathbb{R}^p$, which is precisely the MLE estimator for (3.10). We now specialize the main result of [2, Theorem 3.1] to the problem (3.10). Suppose that the following assumptions hold:[5]

(a) *(Realizable):* The process $\{z_t\}$ is generated by (3.10) for some $\theta_\star \in \Theta$.

(b) *(Stationary process):* The process $\{z_t\}$ has a stationary measure ν, and $z_1 \sim \nu$.

(c) *(Weakly sub-Gaussian):* The function class $\mathcal{F}' := \{f_{\theta_1} - f_{\theta_2} \mid \theta_1, \theta_2 \in \Theta\}$ satisfies a *weak-sub-Gaussian* condition [cf. 2, Def. 2.1]: there exists $\eta \in (0, 1]$ and $L \geqslant 1$ such that $\|f\|_{\Psi_p} \leqslant L \|f\|^\eta_{\mathcal{L}^2(\nu)}$ for all $f \in \mathcal{F}'$, where $\|f\|_{\Psi_p} := \sup_{k \in \mathbb{N}_+} k^{-1/p} \|f\|_{\mathcal{L}^p(\nu)}$.

(d) *(Function class regularity):* For every $x \in \mathbb{R}^d$, the map $\theta \mapsto f_\theta(x)$ is $L(x)$-Lipschitz, with $\|L(x)\|_{\mathcal{L}^2(\nu)} < \infty$. Furthermore, the set $\Theta \subseteq \mathbb{R}^p$ is a bounded set.

---

[5]We state a clear set of assumptions, but not the most minimal, as [2, Theorem 3.1] is stated in broad generality.

(e) *(Burn-in):* Either (i) $mT \gtrsim \mathrm{poly}_\eta(T,p)$ or (ii) $\{z_t\}$ is β-mixing with $\beta(k) \leqslant C\exp(-\rho k)$ and $T \gtrsim \kappa := \lceil \rho^{-1}\log(CmT/\delta)\rceil$, $mT \gtrsim \mathrm{poly}_\eta(\kappa,p)$, where $\mathrm{poly}_\eta(\cdot)$ denotes that the polynomial dependence is a function of η.

Then, the MLE estimator $\hat{\theta}_{m,T}$ satisfies with probability at least $1 - \delta$,

$$\sigma^2 \cdot \mathsf{ER}(\hat{\theta}_{m,T}) \asymp \|f_{\hat{\theta}_{m,T}} - f_{\theta_\star}\|_{\mathcal{L}^2(\nu)}^2 \lesssim \frac{\sigma_{\mathrm{prox}}^2(p + \log(1/\delta))}{mT}, \tag{3.11}$$

where $\sigma_{\mathrm{prox}}^2 \geqslant \sigma^2$ is a variance proxy which is determined from the specific choice of (p,η) in Assumption (c). In the case where $p = \infty$, we have $\sigma_{\mathrm{prox}}^2 = \sigma^2$. Furthermore, when $p < \infty$ and $\eta = 1$, we have that $\sigma_{\mathrm{prox}}^2 = C_p\sigma^2 + o_{mT}(1)$ by the martingale Rosenthal inequality (cf. Theorem A.6), where C_p is a constant only depending on p. On the other hand, when $p < \infty$ and $\eta < 1$, the precise relationship between σ_{prox}^2 and σ^2 is more complex. We observe that the rate (3.11) is order-wise optimal from asymptotic normality (up to the variance proxy σ_{prox}^2); importantly, the rate (3.11) has the correct dependence on the entire dataset size mT, compared with the deflated rate mT/κ from the previous single-trajectory reduction.

However, Assumptions (a)-(e) can be restrictive and/or challenging to verify. We first note that the stationary process Assumption (b) can be removed by using [19, Theorem 4.1] instead of [2, Theorem 3.1], although the downside of this is that the weakly sub-Gaussian Assumption (c) is then replaced with a trajectory-level hyper-contractivity condition (cf. [19, Def. 4.1]) which is more challenging to verify. On the other hand, while the weakly sub-Gaussian Assumption (c) holds broadly if f_θ is bounded and smooth in its input (cf. [2, Prop. 4.1]), the constants (L,η) provided depend poorly on the process dimension d, which yields burn-in times for m, T that can depend exponentially in d; sharp control on the (L,η) constants is only currently available for simple function classes, e.g., linear function classes.

Summary. The finite-sample behavior of the MLE in multi-trajectory settings is currently most broadly available through reduction to either an existing i.i.d. or single-trajectory result. In either case, there is a gap between the resulting bound (cf. (3.8) for the i.i.d. reduction and (3.9) for the single-trajectory reduction) compared with the optimal bound (3.4) in terms of effective sample sizes. In the case of least-squares regression (both for linear and more general parametric models), however, the finite-sample rate (3.6) for linear regression and (3.11) for more general parametric regression matches the CLT-optimal bound up to constant factors in the former, and up to a variance-proxy factor in the latter. This naturally raises the question whether optimal finite-sample rates can be derived beyond the square-loss setting. The proof techniques used for analyzing the square-loss (e.g., self-normalized martingales, small-ball inequalities) take advantage of either the closed-form nature of the linear regression solution, or specific properties of the square-loss such as the offset basic inequality [see e.g., 52], and hence do not readily generalize. This motivates the need for a different approach for establishing error bounds of the form (3.4) in more general settings.

3.2 Analyzing MLE via Localization in Hellinger Distance

We next develop a set of tools and a general five-step framework for analyzing the MLE over a diverse set of problems. The roadmap for the remainder of this section is as follows. We first develop tools in Section 3.2.1 to control the Hellinger distance of the MLE solution to the true

solution in terms of their length-T trajectory (path) measures. Next, we study in Section 3.2.2 how we can localize the Hellinger distance so that it approximately behaves like a weighted Euclidean norm over the parameters, where the weight is determined by the Fisher information matrix at optimality. Importantly, given sufficient trajectory-level excitation, the FI matrix scales with the trajectory length T, providing the correct scaling with length of each trajectory. Building on these mathematical tools, in Section 3.3 we work through a simple illustrative example combining these tools to derive a sharp rate for parameter recovery in a two-state Markov chain. Finally, we present our general Hellinger localization framework in Section 3.4.

Divergence measures. For two measures p, q over the same probability space, we define the Total-Variation (TV) distance, Hellinger distance, and Kullback-Leibler (KL) divergence as:

$$\|p - q\|_{\mathrm{TV}} := \frac{1}{2} \int |p - q| \, \mathrm{d}\mu, \quad \mathrm{d}_H(p,q) := \sqrt{\int \left(\sqrt{p} - \sqrt{q}\right)^2 \mathrm{d}\mu}, \quad \mathrm{KL}(p \parallel q) := \mathbb{E}_p\left[\log \frac{p}{q}\right].$$

Note that the last definition of KL divergence require the absolute continuity condition $p \ll q$. For what follows, we will often overload notation and write e.g., $\mathrm{d}_H(\theta_1, \theta_2) = \mathrm{d}_H(p_{\theta_1}, p_{\theta_2})$ for $\theta_1, \theta_2 \in \Theta$ (and similarly for TV distance and KL divergences).

3.2.1 Control of Trajectory Measures in Hellinger Distance

Our main approach is based on techniques used for studying density estimation with maximum-likelihood. To set the stage for what follows, we first state a prototypical non-asymptotic result from the study of maximum-likelihood estimators. The following instantiation is from [53] and applied directly to our problem setting (3.1), although it traces its roots back to the work of [36, 37]. Similar instantiations of the following result can also be found in more recent works [54–56].

Theorem 3.1 (cf. [53, Proposition B.1]). *We have with probability at least $1 - \delta$,*

$$\mathrm{d}_H^2(\hat{p}_{m,T}, p_\star) \leqslant \inf_{\varepsilon > 0} \left\{ \frac{6 \log(2\mathcal{N}_\infty(\mathcal{P}, \varepsilon)/\delta)}{m} + 4\varepsilon \right\},$$

where $\mathcal{N}_\infty(\mathcal{P}, \varepsilon)$ is the ε-covering number of $\mathcal{P}$ in the max divergence.[6]

Theorem 3.1 is a powerful result in that it controls the Hellinger divergence of the *trajectory (path)* distributions between the MLE estimator $\hat{p}_{m,T}(z_{1:T})$ and the true data-generating trajectory distribution $p_\star(z_{1:T})$ at a $1/m$ rate, under fairly minimal assumptions on $\mathcal{P}$. In fact, the only assumption made on $\mathcal{P}$ is that its max divergence covering number is bounded. However, powerful as this result may be, extracting trajectory information out of the Hellinger divergence in order to obtain $1/(mT)$ rates is non-trivial, as the Hellinger distance does *not* in general tensorize nicely over non-product measures, unlike the KL-divergence.

Fortunately, some form of tensorization is indeed possible when $\hat{p}_{m,T}$ is close enough to $p_\star$. In particular, by an asymptotic argument [see e.g. 39, Theorem 7.23] for $\theta_0, \theta_1 \in \Theta$, the following local expansion holds:

$$\mathrm{d}_H^2(\theta_0, \theta_1) = \frac{1}{4}\|\theta_0 - \theta_1\|_{\mathcal{I}(\theta_0)}^2 + o(\|\theta_0 - \theta_1\|^2). \tag{3.12}$$

---

[6]Specifically, a set $\mathcal{P}' \subseteq \mathcal{P}$ is an ε-covering in max divergence if for every $p \in \mathcal{P}$ there exists a $p' \in \mathcal{P}'$ such that for a.e. $z_{1:T} \in \mathsf{Z}^T$, $\log(p(z_{1:T})/p'(z_{1:T})) \leqslant \varepsilon$. The quantity $\mathcal{N}_\infty(\mathcal{P}, \varepsilon)$ denotes the cardinality of the smallest such ε-covering.

Combining (3.12) with Theorem 3.1 yields a bound which resembles that of the CLT (3.4). Our key result is to quantify the region for which such a local expansion (3.12) holds, using a second-order Taylor expansion argument. The argument proceeds in two steps. Due to the nature of Taylor's theorem, we first need to uniformly control the performance of parameters within $\mathrm{conv}\{\hat{\theta}_{m,T}, \theta_\star\}$,[7] which we do via a star-shaped variation of Theorem 3.1. We then Taylor expand the squared Hellinger distance and characterize the necessary radius conditions for (3.12) to hold.

To proceed, we first require the definition of an ε-cover in Hellinger distance.

Definition 3.2 (Hellinger cover). *A set $\mathcal{P}' \subseteq \mathcal{P}$ is an ε-covering of $\mathcal{P}$ in Hellinger distance if for every $p \in \mathcal{P}$, there exists a $p' \in \mathcal{P}'$ such that $\mathrm{d}_H(p, p') \leqslant \varepsilon$. The ε-covering number of $\mathcal{P}$ in Hellinger distance, denoted $\mathcal{N}_H(\mathcal{P}, \varepsilon)$, is defined as the cardinality of the smallest such ε-covering.*

One issue which arises with either Hellinger or squared Hellinger distance is that it is not convex in its parameterization, i.e., in general we have neither $\theta \mapsto \mathrm{d}_H(\theta, \theta_1)$ nor $\theta \mapsto \mathrm{d}_H^2(\theta, \theta_1)$ is convex for a fixed θ_1; f-divergences are jointly convex in the space of *probability measures*, but not necessarily the specific parameterization. Thus, it will be necessary to consider another type of divergence.

For what follows, given $\theta_0, \theta_1 \in \Theta$, we define $\mathcal{I}(\theta_0, \theta_1)$ as the matrix:

$$\mathcal{I}(\theta_0, \theta_1) := \int_0^1 \mathcal{I}(\theta(s))\mathrm{d}s, \quad \theta(s) := (1-s)\theta_0 + s\theta_1.$$

Note that $\mathcal{I}(\theta_0, \theta_1)$ is symmetric, i.e., $\mathcal{I}(\theta_0, \theta_1) = \mathcal{I}(\theta_1, \theta_0)$. We also assume there exists a positive definite matrix $\mathcal{I}_{\max}$ such that $\mathcal{I}(\theta) \preccurlyeq \mathcal{I}_{\max}$ for all $\theta \in \Theta$. We use this to define both a symmetric averaged Fisher Information, and a max Fisher Information divergence measure:

$$\mathrm{d}_{\mathrm{FI}}(p_{\theta_0}, p_{\theta_1}) := \|\theta_0 - \theta_1\|_{\mathcal{I}(\theta_0, \theta_1)}, \quad \mathrm{d}_{\mathcal{I}_{\max}}(p_{\theta_0}, p_{\theta_1}) := \|\theta_0 - \theta_1\|_{\mathcal{I}_{\max}}. \tag{3.13}$$

The relationship $\mathrm{d}_{\mathrm{FI}}(p_{\theta_0}, p_{\theta_1}) \leqslant \mathrm{d}_{\mathcal{I}_{\max}}(p_{\theta_0}, p_{\theta_1})$ is clear by definition. Furthermore, the max FI measure exhibits the necessary convexity of $\theta \mapsto \mathrm{d}_{\mathcal{I}_{\max}}(\theta, \theta_1)$ for all fixed θ_1, via the convexity of the weighted ℓ_2 norm. We now show the following connection that these two FI distances dominate the Hellinger distance, with proof deferred to Appendix A.

Proposition 3.3. *For any $\theta_0, \theta_1 \in \Theta$ such that $\mathrm{conv}(\theta_0, \theta_1) \subseteq \Theta$, we have:*

$$\mathrm{d}_H(p_{\theta_0}, p_{\theta_1}) \leqslant \frac{1}{2}\mathrm{d}_{\mathrm{FI}}(p_{\theta_0}, p_{\theta_1}) \leqslant \frac{1}{2}\mathrm{d}_{\mathcal{I}_{\max}}(p_{\theta_0}, p_{\theta_1}).$$

Parallel to Definition 3.2, we also define a covering in terms of the max FI divergence as follows.

Definition 3.4 (Max FI cover). *A set $\mathcal{P}' \subseteq \mathcal{P}$ is an ε-covering of $\mathcal{P}$ in the max Fisher Information divergence if for every $p_\theta \in \mathcal{P}$, there exists a $p_{\theta'} \in \mathcal{P}'$ such that $\|\theta - \theta'\|_{\mathcal{I}_{\max}} \leqslant \varepsilon$. The ε-covering number of $\mathcal{P}$ in the max Fisher Information divergence, denoted $\mathcal{N}_{\mathcal{I}_{\max}}(\mathcal{P}, \varepsilon)$, is defined as the cardinality of the smallest such ε-covering.*

Using this definition of ε-covering, we next introduce a *discretized* version of the MLE estimator (3.1). For $\varepsilon \geqslant 0$, we let $\mathcal{P}_\varepsilon \subseteq \mathcal{P}$ denote a minimal ε-covering of $\mathcal{P}$ in either the Hellinger divergence

---

[7]We define $\mathrm{conv}\{\theta_0, \theta_1\} := \{(1-s)\theta_0 + s\theta_1 \mid s \in [0,1]\}$.

(cf. Definition 3.2) or max FI divergence (cf. Definition 3.4);[8] the specific divergence will be clear from context. We then define the MLE over this set as:

$$\hat{p}_{m,T}^{\varepsilon} \in \arg\max_{p \in \mathcal{P}_{\varepsilon}} \sum_{i=1}^{m} \log p(z_{1:T}^{(i)}). \tag{3.14}$$

We also denote the parameters $\hat{\theta}_{m,T}^{\varepsilon} \in \Theta$ so that $\hat{p}_{m,T}^{\varepsilon} = p_{\hat{\theta}_{m,T}^{\varepsilon}}$. We further introduce the definition of the log-concavity of a parameterization of a density class.

Definition 3.5. *We say that $\mathcal{P}$ is* log-concave *if Θ is convex, and furthermore for every $\theta_0, \theta_1 \in \Theta$, $s \in [0,1]$, and $\mu^{\otimes T}$-a.e. $z \in \mathsf{Z}^T$,*

$$\log p_{s\theta_0 + (1-s)\theta_1}(z) \geqslant s \log p_{\theta_0}(z) + (1-s) \log p_{\theta_1}(z).$$

That is, for $\mu^{\otimes T}$-a.e. $z \in \mathsf{Z}^T$, the function $\theta \mapsto \log p_\theta(z)$ is concave over Θ.

Finally, we define the max FI-diameter of Θ as:

$$\mathrm{diam}(\Theta) := \sup_{\theta_0, \theta_1 \in \Theta} \|\theta_0 - \theta_1\|_{\mathcal{I}_{\max}}.$$

We are now in a position to state the main result of the section.

Theorem 3.6. *Fix $\delta \in (0,1)$ and resolution $\varepsilon \in [0, \delta/(2\sqrt{2m})]$. We have the following:*

(a). *With probability at least $1 - \delta$ over the data $\mathcal{D}_{m,T}$, the Hellinger divergence discretized MLE estimator $\hat{\theta}_{m,T}^{\varepsilon}$ satisfies:*

$$\mathrm{d}_H^2(\hat{\theta}_{m,T}^{\varepsilon}, \theta_\star) \leqslant \frac{4 \log(2\mathcal{N}_H(\mathcal{P}, \varepsilon)/\delta)}{m} + 2\varepsilon^2. \tag{3.15}$$

Furthermore, the same bound (3.15) holds for the max FI divergence discretized MLE estimator with $\mathcal{N}_H(\mathcal{P}, \varepsilon)$ replaced with $\mathcal{N}_{\mathcal{I}_{\max}}(\mathcal{P}, \varepsilon)$.

(b). *If we further assume that $\mathcal{P}$ is log-concave (cf. Definition 3.5), then with probability at least $1 - \delta$ over the data $\mathcal{D}_{m,T}$, the max FI divergence discretized MLE estimator $\hat{\theta}_{m,T}^{\varepsilon}$ satisfies:*

$$\sup_{s \in [0,1]} \mathrm{d}_H^2((1-s)\theta_\star + s\hat{\theta}_{m,T}^{\varepsilon}, \theta_\star) \leqslant \inf_{\eta > 0} \left\{ \frac{6}{m} \log\left(\frac{2\mathcal{N}_{\mathcal{I}_{\max}}(\mathcal{P}, \varepsilon)}{\delta} \left\lceil \frac{1}{2\eta} \right\rceil \right) + \frac{3\eta^2}{4} \mathrm{diam}^2(\Theta) + 3\varepsilon^2 \right\}. \tag{3.16}$$

Before we turn to the proof of Theorem 3.6, several remarks are in order.

Remark 3.7. One key difference between Theorem 3.1 and Theorem 3.6 is that the former applies directly to the MLE estimator $\hat{\theta}_{m,T}$ (3.1) over $\mathcal{P}$, whereas the latter applies to the discretized MLE estimator $\hat{\theta}_{m,T}^{\varepsilon}$ (3.14) over $\mathcal{P}_{\varepsilon}$. In practice there is no difference between these two estimators at a sufficiently small ε below floating point resolution. However, from a theoretical perspective, the discrete estimator seems to exhibit more favorable properties than the exact MLE estimator. One

---

[8]If there is not a unique minimal ε-covering, then we break ties in an arbitrary way so that $\mathcal{P}_{\varepsilon}$ is not ambiguous.

of these properties is allowing one to relax the covering requirement on $\mathcal{P}$ to either Hellinger (3.15) or max FI-divergence (3.16), both which are less stringent than the max divergence covering in Theorem 3.1, which requires an almost sure bound on the log-density ratio. This is an important relaxation, as it allows us to handle trajectory distributions where the paths $z_{1:T}$ are not bounded almost surely; in such situations the Hellinger/max-FI divergences can still be finite as we will see in the sequel. We leave open the question of whether rates of the form (3.15) and (3.16) are possible for $\hat{p}_{m,T}$ without relying on max divergence coverings, noting that some extra tail conditions on $\mathcal{P}$ would be needed to control the behavior of the empirical log likelihood $\frac{1}{m}\sum_{i=1}^{m}\log p(z_{1:T}^{(i)})$.

Remark 3.8. The key difference between (3.15) and (3.16) is that the former only controls $d_H(\hat{\theta}_{m,T}^{\varepsilon}, \theta_{\star})$, whereas the latter controls $d_H(\theta, \theta_{\star})$ along the entire ray $\theta \in \operatorname{conv}\{\hat{\theta}_{m,T}^{\varepsilon}, \theta_{\star}\}$. Note that since in general neither Hellinger nor squared Hellinger distance is convex in the *parameter space*, the former in general does *not* imply the latter. Thus, (3.16) is a strictly stronger conclusion than (3.15), and therefore requires a stronger set of assumptions (e.g., log-concavity of $\mathcal{P}$). As we will see in the sequel, the conclusion (3.16) will play an important role in allowing $m \gtrsim \operatorname{polylog}(T)$ instead of $m \gtrsim T \cdot \operatorname{polylog}(T)$ minimum number of trajectories for our CLT rates to hold.

Proof of Theorem 3.6. The proof follows the general structure of [53, Proposition B.1], but includes a few crucial modifications. Before we begin, we state the following upper bound on squared Hellinger distance which holds generally for two distributions p, q, which follows from the inequality $\log(1+x) \leqslant x$ for $x > -1$:

$$
\begin{aligned}
\frac{1}{2}d_H^2(p,q) &\leqslant -\log\left(1 - \frac{1}{2}d_H^2(p,q)\right) \\
&= -\log\left(\mathbb{E}_p\left[\exp\left(-\frac{1}{2}\log\frac{p}{q}\right)\right]\right) = -\log\left(\mathbb{E}_q\left[\exp\left(-\frac{1}{2}\log\frac{q}{p}\right)\right]\right). \quad (3.17)
\end{aligned}
$$

(a). Let $\mathcal{P}_{\varepsilon} \subseteq \mathcal{P}$ denote a minimal ε-covering of $\mathcal{P}$ in the Hellinger distance (cf. Definition 3.2). Let us abbreviate $\hat{p}^{\varepsilon} := \hat{p}_{m,T}^{\varepsilon}$, and for $p \in \mathcal{P}$ let $\varphi_{\varepsilon}[p] \in \mathcal{P}_{\varepsilon}$ denote the closest element in the Hellinger cover, i.e., $d_H(\varphi_{\varepsilon}[p], p) \leqslant \varepsilon$. We first consider a hypothetical scenario where each $z^{(i)} := z_{1:T}^{(i)}$ in $\mathcal{D}_{m,T}$ is drawn i.i.d. from $p_{\star}^{\varepsilon} := \varphi_{\varepsilon}[p_{\star}]$ instead of $p_{\star}$. By combining (3.17) and Proposition A.5 with a union bound over $\mathcal{P}_{\varepsilon}$, we have with probability at least $1 - \delta/2$ over $(p_{\star}^{\varepsilon})^{\otimes m}$,

$$
\frac{1}{2}d_H^2(\hat{p}^{\varepsilon}, p_{\star}^{\varepsilon}) \leqslant -\log\left(\mathbb{E}_{p_{\star}^{\varepsilon}}\left[\exp\left(-\frac{1}{2}\log\frac{p_{\star}^{\varepsilon}}{\hat{p}^{\varepsilon}}\right)\right]\right) \leqslant \frac{1}{2m}\sum_{i=1}^{m}\log\frac{p_{\star}^{\varepsilon}(z^{(i)})}{\hat{p}^{\varepsilon}(z^{(i)})} + \frac{1}{m}\log\left(\frac{2|\mathcal{P}_{\varepsilon}|}{\delta}\right) \leqslant \frac{1}{m}\log\left(\frac{2|\mathcal{P}_{\varepsilon}|}{\delta}\right),
$$

where the last inequality is since $\hat{p}^{\varepsilon}$ is the MLE over $\mathcal{P}_{\varepsilon}$ and $p_{\star}^{\varepsilon} \in \mathcal{P}_{\varepsilon}$. On the other hand, by triangle inequality for Hellinger distance followed by the inequality $(a+b)^2 \leqslant 2(a^2+b^2)$ for $a, b \in \mathbb{R}$,

$$
d_H^2(\hat{p}^{\varepsilon}, p_{\star}) \leqslant 2d_H^2(\hat{p}^{\varepsilon}, p_{\star}^{\varepsilon}) + 2d_H^2(p_{\star}^{\varepsilon}, p_{\star}) \leqslant \frac{4}{m}\log\left(\frac{2|\mathcal{P}_{\varepsilon}|}{\delta}\right) + 2\varepsilon^2.
$$

Hence, we have shown that:

$$
\mathbb{P}_{\mathcal{D}_{m,T} \sim (p_{\star}^{\varepsilon})^{\otimes m}}\left\{d_H^2(\hat{p}^{\varepsilon}[\mathcal{D}_{m,T}], p_{\star}) > \frac{4}{m}\log\left(\frac{2|\mathcal{P}_{\varepsilon}|}{\delta}\right) + 2\varepsilon^2\right\} \leqslant \delta/2,
$$

14

where $\hat{p}^\varepsilon[\mathcal{D}_{m,T}]$ is notation to emphasize that $\hat{p}^\varepsilon$ is a function of the data $\mathcal{D}_{m,T}$. Recall that $\|p - q\|_{\mathrm{TV}} \leqslant \mathrm{d}_H(p,q)$ for two measures p, q [cf. 39, Section 7.3]. Hence we can change measure between $\mathcal{D}_{m,T} \sim (p_\star^\varepsilon)^{\otimes m}$ and $\mathcal{D}_{m,T} \sim p_\star^{\otimes m}$ as follows:

$$\mathbb{P}_{\mathcal{D}_{m,T} \sim p_\star^{\otimes m}} \left\{ \mathrm{d}_H^2(\hat{p}^\varepsilon[\mathcal{D}_{m,T}], p_\star) > \frac{4}{m} \log\left(\frac{2|\mathcal{P}_\varepsilon|}{\delta}\right) + 2\varepsilon^2 \right\} \leqslant \|p_\star^{\otimes m} - (p_\star^\varepsilon)^{\otimes m}\|_{\mathrm{TV}} + \delta/2$$
$$\leqslant \mathrm{d}_H(p_\star^{\otimes m}, (p_\star^\varepsilon)^{\otimes m}) + \delta/2$$
$$\leqslant \delta,$$

where the last inequality follows from Proposition A.2 since we have $\mathrm{d}_H(p_\star, p_\star^\varepsilon) \leqslant \varepsilon \leqslant \delta/(2\sqrt{2m})$. This establishes (3.15) for the Hellinger divergence discretized $\hat{\theta}_{m,T}^\varepsilon$. The proof for the max FI divergence discretized estimator is nearly identical, and hence omitted.

(b). For $p \in \mathcal{P}$, let $\varphi_\varepsilon[p] \in \mathcal{P}_\varepsilon$ denote the closest element in the max FI-divergence ε-covering $\mathcal{P}_\varepsilon \subseteq \mathcal{P}$, so that $\mathrm{d}_{\mathcal{I}_{\max}}(\varphi_\varepsilon[p], p) \leqslant \varepsilon$ for all $p \in \mathcal{P}$. We define the following parameterization of $\mathrm{conv}\{\theta, \hat{\theta}_{m,T}^\varepsilon\}$:

$$\hat{\theta}^\varepsilon(s; \theta) := (1 - s)\theta + s\hat{\theta}_{m,T}^\varepsilon.$$

We also abbreviate $\hat{\theta}^\varepsilon = \hat{\theta}_{m,T}^\varepsilon$ and $\theta_\star^\varepsilon = \varphi_\varepsilon[\theta_\star]$. Next, we let $s_1, \ldots, s_N \in [0,1]$ be a minimal η-covering of $[0,1]$ in absolute value. For any $s \in [0,1]$, letting s_η denote its nearest element in the cover, by the triangle inequality for Hellinger distance:

$$\mathrm{d}_H(\hat{\theta}^\varepsilon(s; \theta_\star), \theta_\star) \leqslant \mathrm{d}_H(\hat{\theta}^\varepsilon(s; \theta_\star), \theta_\star^\varepsilon) + \mathrm{d}_H(\theta_\star^\varepsilon, \theta_\star)$$
$$\leqslant \mathrm{d}_H(\hat{\theta}^\varepsilon(s; \theta_\star^\varepsilon), \theta_\star^\varepsilon) + \mathrm{d}_H(\hat{\theta}^\varepsilon(s; \theta_\star^\varepsilon), \hat{\theta}^\varepsilon(s; \theta_\star)) + \mathrm{d}_H(\theta_\star^\varepsilon, \theta_\star)$$
$$\leqslant \mathrm{d}_H(\hat{\theta}^\varepsilon(s_\eta; \theta_\star^\varepsilon), \theta_\star^\varepsilon) + \mathrm{d}_H(\hat{\theta}^\varepsilon(s; \theta_\star^\varepsilon), \hat{\theta}^\varepsilon(s_\eta; \theta_\star^\varepsilon))$$
$$\qquad + \mathrm{d}_H(\hat{\theta}^\varepsilon(s; \theta_\star^\varepsilon), \hat{\theta}^\varepsilon(s; \theta_\star)) + \mathrm{d}_H(\theta_\star^\varepsilon, \theta_\star)$$
$$\leqslant \mathrm{d}_H(\hat{\theta}^\varepsilon(s_\eta; \theta_\star^\varepsilon), \theta_\star^\varepsilon) + \frac{1}{2}\|\hat{\theta}^\varepsilon(s; \theta_\star^\varepsilon) - \hat{\theta}^\varepsilon(s_\eta; \theta_\star^\varepsilon)\|_{\mathcal{I}_{\max}} \qquad \text{[using Proposition 3.3]}$$
$$\qquad + \frac{1}{2}\|\hat{\theta}^\varepsilon(s; \theta_\star^\varepsilon) - \hat{\theta}^\varepsilon(s; \theta_\star)\|_{\mathcal{I}_{\max}} + \frac{1}{2}\|\theta_\star^\varepsilon - \theta_\star\|_{\mathcal{I}_{\max}}$$
$$\leqslant \mathrm{d}_H(\hat{\theta}^\varepsilon(s_\eta; \theta_\star^\varepsilon), \theta_\star^\varepsilon) + \frac{|s - s_\eta|}{2}\|\hat{\theta}^\varepsilon - \theta_\star^\varepsilon\|_{\mathcal{I}_{\max}} + \varepsilon$$
$$\leqslant \mathrm{d}_H(\hat{\theta}^\varepsilon(s_\eta; \theta_\star^\varepsilon), \theta_\star^\varepsilon) + \frac{\eta}{2}\mathrm{diam}(\Theta) + \varepsilon. \tag{3.18}$$

As in the proof of (a), we first consider a hypothetical scenario where each $z^{(i)} := z_{1:T}^{(i)}$ in $\mathcal{D}_{m,T}$ is drawn i.i.d. from $p_\star^\varepsilon := p_{\theta_\star^\varepsilon}$. By combining (3.17) and Proposition A.5 with a union bound over both $\mathcal{P}_\varepsilon$ and $\{s_k\}_{k=1}^N$, we have with probability at least $1 - \delta/2$ over $(p_\star^\varepsilon)^{\otimes m}$, abbreviating $\hat{\theta}^\varepsilon(s_\eta) := \hat{\theta}^\varepsilon(s_\eta; \theta_\star^\varepsilon)$,

$$\frac{1}{2}\mathrm{d}_H^2(\hat{\theta}^\varepsilon(s_\eta), \theta_\star^\varepsilon) \leqslant -\log\left(\mathbb{E}_{p_\star^\varepsilon}\left[\exp\left(-\frac{1}{2}\log\frac{p_\star^\varepsilon}{p_{\hat{\theta}^\varepsilon(s_\eta)}}\right)\right]\right)$$
$$\leqslant \frac{1}{2m}\sum_{i=1}^m \log\frac{p_\star^\varepsilon(z^{(i)})}{p_{\hat{\theta}^\varepsilon(s_\eta)}(z^{(i)})} + \frac{1}{m}\log\left(\frac{2|\mathcal{P}_\varepsilon|}{\delta}\left\lceil\frac{1}{2\eta}\right\rceil\right) \leqslant \frac{1}{m}\log\left(\frac{2|\mathcal{P}_\varepsilon|}{\delta}\left\lceil\frac{1}{2\eta}\right\rceil\right), \tag{3.19}$$

15

where the last inequality holds from the following arguments. First note that log-concavity of $\mathcal{P}$ means that for $\mu^{\otimes T}$ a.e. $z \in \mathsf{Z}^T$, $-\log p_{\hat{\theta}^\varepsilon(s_\eta)}(z) \leqslant -(1-s_\eta)\log p_{\theta_\star^\varepsilon}(z) - s_\eta \log p_{\hat{\theta}^\varepsilon}(z)$. Hence,

$$\log \frac{p_\star^\varepsilon(z)}{p_{\hat{\theta}^\varepsilon(s_\eta)}(z)} = \log p_\star^\varepsilon(z) - \log p_{\hat{\theta}^\varepsilon(s_\eta)}(z)$$

$$\leqslant \log p_\star^\varepsilon(z) - (1-s_\eta)\log p_{\theta_\star^\varepsilon}(z) - s_\eta \log p_{\hat{\theta}^\varepsilon}(z) = s_\eta \log \frac{p_{\theta_\star^\varepsilon}(z)}{p_{\hat{\theta}^\varepsilon}(z)},$$

and therefore we have that the empirical log-likelihood ratio satisfies:

$$\sum_{i=1}^m \log \frac{p_\star^\varepsilon(z^{(i)})}{p_{\hat{\theta}^\varepsilon(s_\eta)}(z^{(i)})} \leqslant s_\eta \sum_{i=1}^m \log \frac{p_\star^\varepsilon(z^{(i)})}{p_{\hat{\theta}^\varepsilon}(z^{(i)})} \leqslant 0,$$

where the last inequality holds since $\hat{\theta}^\varepsilon$ is a MLE over $\mathcal{P}_\varepsilon$ and $p_\star^\varepsilon \in \mathcal{P}_\varepsilon$. Let us denote the event that (3.19) holds as $\mathcal{E}_1$. On this event, we have by (3.18) and $(a+b+c)^2 \leqslant 3(a^2+b^2+c^2)$ for $a, b, c \in \mathbb{R}$, on $\mathcal{E}_1$, for every $s \in [0,1]$,

$$\mathrm{d}_H^2(\hat{\theta}^\varepsilon(s;\theta_\star),\theta_\star) \leqslant 3\mathrm{d}_H^2(\hat{\theta}^\varepsilon(s_\eta;\theta_\star^\varepsilon),\theta_\star) + \frac{3\eta^2}{4}\mathrm{diam}^2(\Theta) + 3\varepsilon^2$$

$$\leqslant \frac{6}{m}\log\left(\frac{2|\mathcal{P}_\varepsilon|}{\delta}\left\lceil\frac{1}{2\eta}\right\rceil\right) + \frac{3\eta^2}{4}\mathrm{diam}^2(\Theta) + 3\varepsilon^2.$$

Hence, we have shown that:

$$\mathbb{P}_{\mathcal{D}_{m,T}\sim(p_\star^\varepsilon)^{\otimes m}}\left\{\sup_{s\in[0,1]}\mathrm{d}_H^2(\hat{\theta}^\varepsilon[\mathcal{D}_{m,T}](s;\theta_\star),\theta_\star) > \frac{6}{m}\log\left(\frac{2|\mathcal{P}_\varepsilon|}{\delta}\left\lceil\frac{1}{2\eta}\right\rceil\right) + \frac{3\eta^2}{4}\mathrm{diam}^2(\Theta) + 3\varepsilon^2\right\} \leqslant \delta/2.$$

Above, as in part (a), we use the notation $\hat{\theta}^\varepsilon[\mathcal{D}_{m,T}]$ to emphasize the dependence of the estimator on the data $\mathcal{D}_{m,T}$. We now take similar steps as in part (a) to change measure between $\mathcal{D}_{m,T} \sim (p_\star^\varepsilon)^{\otimes m}$ and $\mathcal{D}_{m,T} \sim p_\star^{\otimes m}$:

$$\mathbb{P}_{\mathcal{D}_{m,T}\sim(p_\star)^{\otimes m}}\left\{\sup_{s\in[0,1]}\mathrm{d}_H^2(\hat{\theta}^\varepsilon[\mathcal{D}_{m,T}](s;\theta_\star),\theta_\star) > \frac{6}{m}\log\left(\frac{2|\mathcal{P}_\varepsilon|}{\delta}\left\lceil\frac{1}{2\eta}\right\rceil\right) + \frac{3\eta^2}{4}\mathrm{diam}^2(\Theta) + 3\varepsilon^2\right\}$$

$$\leqslant \|p_\star^{\otimes m} - (p_\star^\varepsilon)^{\otimes m}\|_{\mathrm{TV}} + \delta/2 \leqslant \mathrm{d}_H(p_\star^{\otimes m},(p_\star^\varepsilon)^{\otimes m}) + \delta/2 \leqslant \delta,$$

where the last inequality follows from Proposition A.2 since we have $\mathrm{d}_H(p_\star,p_\star^\varepsilon) \leqslant \varepsilon \leqslant \delta/(2\sqrt{2m})$. This establishes the result. $\qquad\square$

3.2.2 Equivalence of Hellinger Distance and Fisher-weighted Metric

Theorem 3.6 is, at its core, a result about i.i.d. learning. It however contains a rich amount of information about trajectories *within* the divergence term $\mathrm{d}_H^2(\theta_{m,T}^\varepsilon,\theta_\star)$. In this section, we study how to extract this information out of the Hellinger divergence. As previously discussed, this is challenging as neither the Hellinger nor squared Hellinger distance tensorizes across $z_{1:T}$ in non-i.i.d. settings. However, when $\hat{\theta}_{m,T}^\varepsilon$ is close to $\theta_\star$, such a tensorization is indeed possible, as observed via the asymptotic expansion (3.12). Our next result quantifies the radius of validity for this expansion, through a second-order Taylor expansion analysis.

Proposition 3.9. *Fix any $\theta_0, \theta_1 \in \Theta$ where for all $\theta \in \text{conv}\{\theta_0, \theta_1\}$, we have (a) $\theta \in \Theta$ and (b) $\mathcal{I}(\theta) > 0$. Define the quantities:*

$$B_1(\theta_0, \theta_1) := \sup_{\theta \in \text{conv}\{\theta_0, \theta_1\}} \sup_{v \in \mathbb{S}^{p-1}} \|\langle v, \mathcal{I}(\theta)^{-1/2} \nabla_\theta \log p_\theta(z_{1:T})\rangle\|_{\mathcal{L}^4(p_\theta)}, \tag{3.20}$$

$$B_2(\theta_0, \theta_1) := \sup_{\theta \in \text{conv}\{\theta_0, \theta_1\}} \sup_{v \in \mathbb{S}^{p-1}} \|\langle v, \mathcal{I}(\theta)^{-1/2} \nabla_\theta^2 \log p_\theta(z_{1:T}) \mathcal{I}(\theta)^{-1/2} v\rangle\|_{\mathcal{L}^2(p_\theta)}. \tag{3.21}$$

(a). *Suppose that the following condition holds:*

$$\sup_{\theta \in \text{conv}\{\theta_0, \theta_1\}} \mathrm{d}_H(\theta_0, \theta) \leqslant \frac{1}{16\sqrt{2}} \min\left\{\frac{1}{B_1^2(\theta_0, \theta_1)}, \frac{1}{B_2^2(\theta_0, \theta_1)}\right\}. \tag{3.22}$$

Then the following inequalities hold:

$$\frac{3}{16}\|\theta_0 - \theta_1\|_{\mathcal{I}_2(\theta_0, \theta_1)}^2 \leqslant \mathrm{d}_H^2(\theta_0, \theta_1) \leqslant \frac{5}{16}\|\theta_0 - \theta_1\|_{\mathcal{I}_2(\theta_0, \theta_1)}^2, \quad \mathcal{I}_2(\theta_0, \theta_1) := 2\int_0^1 (1-s)\mathcal{I}(\theta(s))\mathrm{d}s, \tag{3.23}$$

where $\theta(s) := (1-s)\theta_0 + s\theta_1$.

(b). *If in addition to (3.22) holding, we furthermore have that:*

$$\sup_{\theta \in \text{conv}\{\theta_0, \theta_1\}} \|\mathcal{I}(\theta_0)^{-1/2}\mathcal{I}(\theta)\mathcal{I}(\theta_0)^{-1/2} - I_p\|_{\text{op}} \leqslant \frac{1}{2}, \tag{3.24}$$

then we also have the following inequalities:

$$\frac{3}{32}\|\theta_0 - \theta_1\|_{\mathcal{I}(\theta_0)}^2 \leqslant \mathrm{d}_H^2(\theta_0, \theta_1) \leqslant \frac{15}{32}\|\theta_0 - \theta_1\|_{\mathcal{I}(\theta_0)}^2. \tag{3.25}$$

Remark 3.10. The constants in (3.23) and (3.25) can be made arbitrarily close to $1/4$ (cf. (3.12)) at the expense of decreasing the constants in the local radius conditions (3.22) and (3.24).

Proof of Proposition 3.9. For $\theta \in \Theta$ and $z \in \mathsf{Z}^T$, define the function $h(\theta; z) := \sqrt{p_\theta(z)}$. Abbreviating $\mu = \mu^{\otimes T}$, we take the first and second derivatives of both $\theta \mapsto h(\theta; z)$ and $\theta \mapsto \mathrm{d}_H^2(\theta_0, \theta)$:

$$\nabla_\theta h(\theta; z) = \frac{1}{2}\sqrt{p_\theta(z)}\nabla_\theta \log p_\theta(z),$$

$$\nabla_\theta^2 h(\theta; z) = \sqrt{p_\theta(z)}\left[\frac{1}{2}\nabla_\theta^2 \log p_\theta(z) + \frac{1}{4}\nabla_\theta \log p_\theta(z)\nabla_\theta \log p_\theta(z)^\mathsf{T}\right],$$

$$\nabla_\theta \mathrm{d}_H^2(\theta_0, \theta) = 2\int (h(\theta; z) - h(\theta_0; z))\nabla_\theta h(\theta; z)\mathrm{d}\mu,$$

$$\nabla_\theta^2 \mathrm{d}_H^2(\theta_0, \theta) = 2\int (h(\theta; z) - h(\theta_0; z))\nabla_\theta^2 h(\theta; z)\mathrm{d}\mu + 2\int \nabla_\theta h(\theta; z)\nabla_\theta h(\theta; z)^\mathsf{T}\mathrm{d}\mu =: \mathcal{H}(\theta; \theta_0).$$

We therefore have the identity for the Hessian $\mathcal{H}(\theta; \theta_0)$:

$$\mathcal{H}(\theta; \theta_0) = 2\int (h(\theta; z) - h(\theta_0; z))\nabla_\theta^2 h(\theta; z)\mathrm{d}\mu + \frac{1}{2}\int \nabla_\theta \log p_\theta(z)\nabla_\theta \log p_\theta(z)^\mathsf{T} p_\theta(z)\mathrm{d}\mu$$

$$= 2\int (h(\theta; z) - h(\theta_0; z))\nabla_\theta^2 h(\theta; z)\mathrm{d}\mu + \frac{1}{2}\mathcal{I}(\theta).$$

17

Using the second-order integral version of Taylor's theorem and expanding $\theta \mapsto d_H^2(\theta_0, \theta)$ around $\theta = \theta_0$, with shorthand notation $\theta_s := \theta(s)$ and $\mathcal{I}_s := \mathcal{I}(\theta(s))$ for $s \in [0,1]$, we have:

$$
\begin{aligned}
d_H^2(\theta_0, \theta_1) &= \int_0^1 (1-s) \Delta^\mathsf{T} \mathcal{H}(\theta_s; \theta_0) \Delta \, ds \\
&= \int_0^1 (1-s) \Delta^\mathsf{T} \left[2 \int (h(\theta_s; z) - h(\theta_0; z)) \nabla_\theta^2 h(\theta_s; z) d\mu + \frac{1}{2} \mathcal{I}_s \right] \Delta \, ds \\
&= \frac{1}{2} \int_0^1 (1-s) \Delta^\mathsf{T} \mathcal{I}_s \Delta \, ds + 2 \int_0^1 (1-s) \int (h(\theta_s; z) - h(\theta_0; z)) \Delta^\mathsf{T} \nabla_\theta^2 h(\theta_s; z) \Delta \, d\mu \, ds.
\end{aligned}
$$

$$(3.26)$$

(a). Fix a vector $q \in \mathbb{R}^p$ and $s \in [0,1]$. We first bound:

$$
\left| \int (h(\theta_s; z) - h(\theta_0; z)) q^\mathsf{T} \mathcal{I}_s^{-1/2} \nabla_\theta^2 h(\theta_s; z) \mathcal{I}_s^{-1/2} q \, d\mu \right|
$$

$$
\overset{(a)}{\leqslant} \sqrt{\int (h(\theta_s; z) - h(\theta_0; z))^2 d\mu} \sqrt{\int (q^\mathsf{T} \mathcal{I}_s^{-1/2} \nabla_\theta^2 h(\theta_s; z) \mathcal{I}_s^{-1/2} q)^2 d\mu}
$$

$$
= d_H(\theta_0, \theta_s) \sqrt{\int (q^\mathsf{T} \mathcal{I}_s^{-1/2} \nabla_\theta^2 h(\theta_s; z) \mathcal{I}_s^{-1/2} q)^2 d\mu}
$$

$$
\overset{(b)}{\leqslant} \|q\|^2 d_H(\theta_0, \theta_s) \sqrt{\frac{1}{2} B_2^2 + \frac{1}{8} B_1^4} \overset{(c)}{\leqslant} \sqrt{2} \|q\|^2 d_H(\theta_0, \theta_s) \max\{B_1^2, B_2\} \overset{(d)}{\leqslant} \frac{1}{16} \|q\|^2,
$$

where (a) is Cauchy-Schwarz, (b) follows by the following bounds with $q_s := \mathcal{I}_s^{-1/2} q$ and using the inequality $(a+b)^2 \leqslant 2(a^2 + b^2)$ for $a, b \in \mathbb{R}$:

$$
\begin{aligned}
\int (q_s^\mathsf{T} \nabla_\theta^2 h(\theta_s; z) q_s)^2 d\mu &= \int p_{\theta_s}(z) \left[\frac{1}{2} q_s^\mathsf{T} \nabla_\theta^2 \log p_{\theta_s}(z) q_s + \frac{1}{4} \langle q_s, \nabla_\theta \log p_{\theta_s}(z) \rangle^2 \right]^2 d\mu \\
&\leqslant \int \left[\frac{1}{2} (q_s^\mathsf{T} \nabla_\theta^2 \log p_{\theta_s}(z) q_s)^2 + \frac{1}{8} \langle q_s, \nabla_\theta \log p_{\theta_s}(z) \rangle^4 \right] p_{\theta_s}(z) d\mu \\
&= \frac{1}{2} \| \langle q, \mathcal{I}_s^{-1/2} \nabla_\theta^2 \log p_{\theta_s}(z) \mathcal{I}_s^{-1/2} q \rangle \|_{\mathcal{L}^2(p_{\theta_s})}^2 + \frac{1}{8} \| \langle q, \mathcal{I}_s^{-1/2} \nabla_\theta \log p_{\theta_s}(z) \rangle \|_{\mathcal{L}^4(p_{\theta_s})}^4 \\
&\leqslant \|q\|^4 \left[\frac{1}{2} B_2^2 + \frac{1}{8} B_1^4 \right],
\end{aligned}
$$

(c) follows from the basic inequalities:

$$
\sqrt{\frac{1}{2} B_2^2 + \frac{1}{8} B_1^4} \leqslant B_1^2 / \sqrt{8} + B_2 / \sqrt{2} \leqslant (B_1^2 + B_2) / \sqrt{2} \leqslant \sqrt{2} \max\{B_1^2, B_2\},
$$

and (d) follows by our stated assumption (3.26). Hence, setting $q = \mathcal{I}_s^{1/2} \Delta$, we have:

$$
\left| \int (h(\theta_s; z) - h(\theta_0; z)) \Delta^\mathsf{T} \nabla_\theta^2 h(\theta_s; z) \Delta \, d\mu \right|
$$

$$
= \left| \int (h(\theta_s; z) - h(\theta_0; z)) (\mathcal{I}_s^{1/2} \Delta)^\mathsf{T} \mathcal{I}_s^{-1/2} \nabla_\theta^2 h(\theta_s; z) \mathcal{I}_s^{-1/2} (\mathcal{I}_s^{1/2} \Delta) \, d\mu \right| \leqslant \frac{1}{16} \|\Delta\|_{\mathcal{I}_s}^2.
$$

Now, utilizing the second order expansion from (3.26),

$$
d_H^2(\theta_0, \theta_1) \geqslant \frac{1}{2} \int_0^1 (1-s) \Delta^\mathsf{T} \mathcal{I}_s \Delta \, ds - 2 \int_0^1 (1-s) \left| \int (h(\theta_s; z) - h(\theta_0; z)) \Delta^\mathsf{T} \nabla_\theta^2 h(\theta_s; z) \Delta \, d\mu \right| ds
$$

18

$$\geqslant \frac{1}{2}\int_0^1 (1-s)\|\Delta\|_{\mathcal{I}_s}^2 \,\mathrm{d}s - \frac{1}{8}\int_0^1 (1-s)\|\Delta\|_{\mathcal{I}_s}^2 \,\mathrm{d}s = \frac{3}{8}\int_0^1 (1-s)\|\Delta\|_{\mathcal{I}_s}^2 \,\mathrm{d}s = \frac{3}{16}\|\Delta\|_{\mathcal{I}_2(\theta_0,\theta_1)}^2.$$

The upper bound is established in a nearly identical way, which yields (3.23).

(b). Using (3.24), we conclude that for all $s \in [0,1]$,

$$\frac{1}{2}I_p \preccurlyeq \mathcal{I}_0^{-1/2}\mathcal{I}_s\mathcal{I}_0^{-1/2} \preccurlyeq \frac{3}{2}I_p.$$

From this, we conclude that:

$$\frac{1}{2}\mathcal{I}(\theta_0) \preccurlyeq \mathcal{I}_2(\theta_0,\theta_1) \preccurlyeq \frac{3}{2}\mathcal{I}(\theta_0),$$

from which (3.25) follows from plugging the above semidefinite inequalities into (3.23). $\qquad\square$

Proposition 3.9 shows that the region for which the asymptotic expansion (3.12) holds is governed by two key conditions: (i) the value of $\sup_{\theta\in\mathrm{conv}\{\theta_0,\theta_1\}} \mathrm{d}_H(\theta_0,\theta)$ being small enough relative to the inverse of the moment bounds $B_1^2(\theta_0,\theta_1)$ and $B_2(\theta_0,\theta_1)$ (cf. (3.22)), and (ii) the parameters θ_0,θ_1 being close enough as measured through the corresponding FI matrices (cf. (3.24)). In Section 3.4, we describe the Hellinger localization framework, which gives a general recipe for verifying these conditions so that Theorem 3.6 can be used in conjunction with Proposition 3.9 to establish non-asymptotic rates for the MLE which exhibit the CLT scaling (3.4). Before describing our general framework, we first work through a specific example next in Section 3.3, which will set the stage for the general recipe.

3.3 A Two-State Markov Chain Example

We now demonstrate how the combination of Theorem 3.6 and Proposition 3.9 gives us a nearly optimal parameter recovery bound via a simple example. We consider a two-state discrete-time Markov chain, where $\mathsf{Z} = \{0,1\}$, $\Theta = [\mu, 1-\mu]$ for some $\mu \in (0,1/2)$, and $\mathcal{P}$ is the set of all two-state Markov chains with $z_1 \sim \rho_1$ (independent of θ) and one-step transition probability:

$$p_\theta(z_{t+1} \mid z_t) = \theta\mathbb{1}\{z_{t+1} = z_t\} + (1-\theta)\mathbb{1}\{z_{t+1} \neq z_t\}, \quad \theta \in \Theta.$$

For what follows, we will assume that $T \geqslant 2$ (otherwise no information about the Markov chain transition probabilities is revealed). We assume $p_\star \in \mathcal{P}$ with parameter $\theta_\star \in \Theta$. While this specific problem is simple enough that its MLE estimator can be studied via only elementary concentration inequalities (which we discuss at the end), we utilize our framework to analyze this problem in order to illustrate both the mechanics and relative sharpness of our arguments.

Roadmap. We first compute the FI matrix $\mathcal{I}(\theta)$ in addition to its uniform bound $\mathcal{I}_{\max}$. From these quantities, we bound the covering number $\mathcal{N}_{\mathcal{I}_{\max}}(\mathcal{P},\varepsilon)$ and invoke Theorem 3.6 (b) for log-concave $\mathcal{P}$; this yields control of the Hellinger distance between $\theta_\star$ and every element in $\mathrm{conv}\{\theta_\star, \hat\theta_{m,T}^\varepsilon\}$, where $\hat\theta_{m,T}^\varepsilon$ denotes the discretized MLE estimator.[9] With this bound in hand, we then estimate the quantities B_1, B_2 (cf. (3.20), (3.21)); a key intermediate step is to show that control of the

---

[9]Specially, $\hat\theta_{m,T}^\varepsilon$ denotes the max FI divergence discretized MLE estimator at resolution $\varepsilon = 1/(2\sqrt{2m})$.

Hellinger distance from Theorem 3.6 implies a direct $O(1/m)$ bound on the squared parameter error, which we then use to localize the B_1, B_2 computation in a small neighborhood around $\theta_\star$. At this point, we are now able to invoke Proposition 3.9 (a) to boost our bound on the squared parameter error to $O(1/(mT))$; the only missing piece is that this bound is not variance-optimal. However, by using this bound to establish that the condition (3.24) holds, we finally conclude by invoking Proposition 3.9 (b), which yields the instance-optimal rate.

Towards carrying out this plan, we first introduce some notation. Let us denote $\sigma_\star^2 := \theta_\star(1 - \theta_\star)$, which is the variance of a $\mathrm{Bern}(\theta_\star)$ distribution, and governs the curvature of the FI matrix $\mathcal{I}(\theta_\star)$. We also define the set $\Theta_{\sigma_\star} := \{\theta \in \Theta \mid |\theta - \theta_\star| \leqslant \sigma_\star^2/2\}$ for localization purposes: observe that for $\theta \in \Theta_{\sigma_\star}$, we have $\frac{1}{\theta(1-\theta)} \leqslant \frac{2}{\sigma_\star^2}$, a key inequality we will use in our computations.

FI matrix, covering number, and Hellinger bound. We first gather the results of some straightforward computations in Appendix B:

$$\mathcal{I}(\theta) = \frac{T-1}{\theta(1-\theta)}, \quad \sup_{\theta \in \Theta} \mathcal{I}(\theta) \leqslant \mathcal{I}_{\max} := \frac{T-1}{\mu(1-\mu)}, \quad \mathrm{diam}(\Theta) \leqslant \frac{T-1}{\mu(1-\mu)}. \tag{3.27}$$

From this, we bound covering number of $\mathcal{P}$ in the max FI-divergence (cf. Definition 3.4) as:

$$\mathcal{N}_{\mathcal{I}_{\max}}(\mathcal{P}, \varepsilon) \leqslant \mathcal{N}_{|\cdot|}\left([\mu, 1 - \mu], \frac{\mu(1-\mu)\varepsilon}{T-1}\right) \leqslant \left\lceil \frac{T-1}{2\mu(1-\mu)\varepsilon} \right\rceil.$$

Now we are in a position to apply Theorem 3.6. Setting $\varepsilon = \delta/(2\sqrt{2m})$ and $\eta = 1/\mathrm{diam}(\Theta)$, we obtain with probability at least $1 - \delta$, the max FI divergence MLE estimator satisfies:

$$\sup_{s \in [0,1]} \mathrm{d}_H^2((1-s)\theta_\star + s\hat{\theta}_{m,T}^\varepsilon, \theta_\star) \lesssim \frac{\log(mT/(\mu\delta))}{m}. \tag{3.28}$$

Let us denote the event in (3.28) by $\mathcal{E}_1$.

Estimate B_1 and B_2. We will estimate $B_1(\theta_0, \theta_1)$ and $B_2(\theta_0, \theta_1)$ over $\theta_0, \theta_1 \in \Theta_{\sigma_\star}$. Before we proceed, we first utilize (3.28) to construct a condition on m such that $\hat{\theta}_{m,T}^\varepsilon \in \Theta_{\sigma_\star}$ on $\mathcal{E}_1$. Specifically:

$$\mathrm{d}_H(\hat{p}_{m,T}^\varepsilon, p_\star) \overset{(a)}{\geqslant} \mathrm{d}_H(\hat{p}_{m,T}^\varepsilon(z_1, z_2), p_\star(z_1, z_2)) \overset{(b)}{\geqslant} \|\hat{p}_{m,T}^\varepsilon(z_1, z_2) - p_\star(z_1, z_2)\|_{\mathrm{TV}}$$

$$\overset{(c)}{=} \mathbb{E}_{z_1 \sim \rho_1}[\|\mathrm{Bern}(\hat{\theta}_{m,T}^\varepsilon) - \mathrm{Bern}(\theta_\star)\|_{\mathrm{TV}}] = |\hat{\theta}_{m,T}^\varepsilon - \theta_\star|,$$

where (a) uses the data processing inequality for f-divergences, (b) uses the inequality $\|p - q\|_{\mathrm{TV}} \leqslant \mathrm{d}_H(p, q)$ for two measures p, q, and (c) uses the fact that $z_1 \sim \rho_1$ irregardless of θ. Hence, if $m \gtrsim \sigma_\star^{-4} \log(mT/(\mu\delta))$, then we have that $\hat{\theta}_{m,T}^\varepsilon \in \Theta_{\sigma_\star}$ on $\mathcal{E}_1$. By convexity of $\Theta_{\sigma_\star}$, this implies that $\mathrm{conv}\{\hat{\theta}_{m,T}^\varepsilon, \theta_\star\} \subset \Theta_{\sigma_\star}$. By Proposition A.1, it suffices to take $m \gtrsim \sigma_\star^{-4} \log(T/(\mu\delta))$. In summary:

$$m \gtrsim \sigma_\star^{-4} \log(T/(\mu\delta)) \implies \mathrm{conv}\{\hat{\theta}_{m,T}^\varepsilon, \theta_\star\} \subset \Theta_{\sigma_\star} \text{ on } \mathcal{E}_1. \tag{3.29}$$

Next, we show in Appendix B that:

$$\mathbb{E}_{p_\theta}[(\partial_\theta \log p_\theta(z_{1:T}))^4] = (T-1)\left(\frac{1}{\theta^3} + \frac{1}{(1-\theta)^3}\right) + 3(T-1)(T-2)\frac{1}{\theta^2(1-\theta)^2},$$

20

$$\mathbb{E}_{p_\theta}[(\partial_\theta^2 \log p_\theta(z_{1:T}))^2] = (T-1)\left(\frac{1}{\theta^3} + \frac{1}{(1-\theta)^3}\right) + (T-1)(T-2)\frac{1}{\theta^2(1-\theta)^2}.$$

Hence, for any $\theta \in \Theta_{\sigma_\star}$, we have:

$$\mathcal{I}(\theta)^{-2}\max\{\mathbb{E}_{p_\theta}[(\partial_\theta \log p_\theta(z_{1:T}))^4], \mathbb{E}_{p_\theta}[(\partial_\theta^2 \log p_\theta(z_{1:T}))^2]\} \lesssim \max\left\{\frac{1}{T\sigma_\star^2}, 1\right\}. \tag{3.30}$$

Therefore, we have established

$$\sup_{\theta_1,\theta_2 \in \Theta_{\sigma_\star}} \max\{B_1^2(\theta_1,\theta_2), B_2(\theta_1,\theta_2)\} \lesssim \max\left\{\frac{1}{\sigma_\star\sqrt{T}}, 1\right\}. \tag{3.31}$$

Parameter error bound. We first verify the condition in (3.22) for $\theta_0 = \theta_\star$ and $\theta_1 = \hat{\theta}_{m,T}^\varepsilon$. By combining (3.28), (3.31), and Proposition A.1, it suffices to choose an m satisfying:

$$mT \gtrsim \frac{1}{\sigma_\star^2}\log(1/(\mu\delta)), \quad m \gtrsim \log(T/(\mu\delta)). \tag{3.32}$$

Thus combining all requirements on m, T from (3.28), (3.29), and (3.32):

$$m \gtrsim \sigma_\star^{-4}\log(T/(\mu\delta)) \implies (3.22) \text{ holds on } \mathcal{E}_1 \text{ for } \theta_0 = \theta_\star \text{ and } \theta_1 = \hat{\theta}_{m,T}^\varepsilon.$$

Therefore by (3.23) from Proposition 3.9, we have on $\mathcal{E}_1$:

$$\mathcal{I}(\theta_\star, \hat{\theta}_{m,T}^\varepsilon)|\hat{\theta}_{m,T} - \theta_\star|^2 \lesssim d_H^2(\hat{\theta}_{m,T}^\varepsilon, \theta_\star) \lesssim \frac{\log(mT/(\mu\delta))}{m}.$$

To lower bound the LHS, observe that for any $\theta \in \Theta$, $\mathcal{I}(\theta) \geq 4(T-1)$. Hence for any $\theta_0, \theta_1 \in \Theta$, $\mathcal{I}(\theta_0, \theta_1) \gtrsim T$, which implies that on $\mathcal{E}_1$,

$$|\hat{\theta}_{m,T}^\varepsilon - \theta_\star|^2 \lesssim \frac{\log(mT/(\mu\delta))}{mT}. \tag{3.33}$$

Verify FI radius. We first observe for any $\theta_0, \theta_1 \in \Theta_{\sigma_\star}$,

$$|\mathcal{I}(\theta_0)^{-1}\mathcal{I}(\theta_1) - 1| = \left|\frac{\theta_0(1-\theta_0)}{\theta_1(1-\theta_1)} - 1\right| \leq \frac{2|\theta_0 - \theta_1|}{\sigma_\star^2}. \tag{3.34}$$

From (3.33) and (3.34), we have on $\mathcal{E}_1$:

$$\sup_{\theta \in \text{conv}\{\hat{\theta}_{m,T}^\varepsilon, \theta_\star\}} |\mathcal{I}(\theta_\star)^{-1}\mathcal{I}(\theta) - 1| \lesssim \sup_{\theta \in \text{conv}\{\hat{\theta}_{m,T}^\varepsilon, \theta_\star\}} \frac{|\theta - \theta_\star|}{\sigma_\star^2} = \frac{|\hat{\theta}_{m,T}^\varepsilon - \theta_\star|}{\sigma_\star^2} \lesssim \frac{1}{\sigma_\star^2}\sqrt{\frac{\log(mT/(\mu\delta))}{mT}}.$$

By another application of Proposition A.1, we can ensure the FI radius condition (3.24) holds on $\mathcal{E}_1$ by setting $mT \gtrsim \sigma_\star^{-4}\log(1/(\mu\delta))$, which is already implied by $m \gtrsim \sigma_\star^{-4}\log(T/(\mu\delta))$. Proposition 3.9 now yields via (3.25) that on $\mathcal{E}_1$,

$$\mathcal{I}(\theta_\star)|\hat{\theta}_{m,T}^\varepsilon - \theta_\star|^2 \lesssim d_H^2(\hat{\theta}_{m,T}^\varepsilon, \theta_\star) \lesssim \frac{\log(mT/(\mu\delta))}{m}.$$

Final result. Combining the previous arguments, the final result is that as long as m satisfies:

$$m \gtrsim \frac{1}{\sigma_\star^4} \log\left(\frac{T}{\mu\delta}\right), \tag{3.35}$$

then with probability at least $1 - \delta$ (over $\mathcal{D}_{m,T}$),

$$|\hat{\theta}_{m,T}^\varepsilon - \theta_\star|^2 \lesssim \frac{\sigma_\star^2 \log(mT/(\mu\delta))}{mT}. \tag{3.36}$$

Sharpness of the result. We now evaluate the sharpness of this result by providing an elementary solution based on sub-Exponential tail inequalities for Binomial distributions. We show in Appendix B that there exists an event $\mathcal{E}_2$ with probability at least $1 - \delta$ that satisfies

$$mT \gtrsim \sigma_\star^{-2} \log(1/\delta) \implies |\hat{\theta}_{m,T} - \theta_\star|^2 \lesssim \frac{\sigma_\star^2 \log(1/\delta)}{mT} \text{ on } \mathcal{E}_2. \tag{3.37}$$

Comparing both the requirement on mT in (3.37) to (3.35), in addition to the final error rate to (3.36), we see that the result utilizing our framework is sharp up to log factors in the final error rate, but misses a few factors in the requirement on m. In particular, (3.37) shows that $mT \gtrsim \tilde{O}(\sigma_\star^{-2})$ suffices to enter the CLT rate regime, but (3.35) requires the more conservative bound $m \gtrsim \tilde{O}(\sigma_\star^{-4})$. We note that the source of this conservatism is due to (a) the use of the data processing inequality to lower bound $\mathrm{d}_H(\hat{p}_{m,T}^\varepsilon, p_\star) \geq \mathrm{d}_H(\hat{p}_{m,T}^\varepsilon(z_1, z_2), p_\star(z_1, z_2))$ and (b) further lower bounding $\mathrm{d}_H(\hat{p}_{m,T}^\varepsilon(z_1, z_2), p_\star(z_1, z_2)) \geq \|\hat{p}_{m,T}^\varepsilon(z_1, z_2) - p_\star(z_1, z_2)\|_{\mathrm{TV}}$. The DPI inequality (a) is lossy in this case, since it is possible to prove (at least when $\rho_1 = \mathrm{Unif}(\{1, 2\})$) that the tensorization property $\mathrm{d}_H^2(\theta_0, \theta_1) = 1 - (1 - \mathrm{d}_H^2(\mathrm{Bern}(\theta_0), \mathrm{Bern}(\theta_1)))^{T-1}$ actually holds [see e.g. 57, Lemma 5]. Going from Hellinger to TV distance in (b) is lossy as well since the TV distance between two Bernoulli distributions loses all local curvature information. Nevertheless, we see that our general framework is able to capture the qualitative aspects of this problem which arise from a problem-specific analysis.

3.4 Hellinger Localization Framework

Previously in Section 3.3, we saw a specific example of how Proposition 3.9 was combined with Theorem 3.6 to establish non-asymptotic rates for MLE which exhibit nearly optimal CLT scaling from (3.3). In this section, we will utilize these two results to provide a general recipe, which we call the *Hellinger localization framework*, for establishing rates. Importantly, our framework in addition to being general purpose, does not inherently rely on any mixing, ergodicity, or stationarity properties of the process $z_{1:T}$, but instead relies on the presence of multiple independent trajectories to allow us to learn from possibly non-stationary and/or non-mixing processes. Before we present the main framework, we introduce a key identifiability condition which plays an important role.

Definition 3.11 (Hellinger Identifiability). *We say that the parametric family $\mathcal{P}$ satisfies (γ_1, γ_2)-Hellinger identifiability (or (γ_1, γ_2)-identifiability) about the point $\theta_\star \in \Theta$ if for every $p_\theta \in \mathcal{P}$:*

$$\mathrm{d}_H(p_\theta, p_{\theta_\star}) \leq \gamma_1 \implies \|\theta - \theta_\star\| \leq \gamma_2 \cdot \mathrm{d}_H(p_\theta, p_{\theta_\star}).$$

Definition 3.11 is in essence the minimal set of assumptions needed for parameter recovery; fortunately it is not hard to see that under our stated assumptions Definition 3.11 holds for some

(γ_1, γ_2) under fairly generic conditions (see Proposition A.8 for a precise statement). On the other, obtaining problem specific constants—especially *sharp* constants i.e., $\gamma_2^{-1} \asymp \sqrt{\lambda_{\min}(\mathcal{I}(\theta_\star))}$ as we expect from asymptotic normality (cf. (3.3))—is non-trivial, and one of our main contributions. Indeed, our work can be contextualized as starting from a fairly sub-optimal pair (γ_1, γ_2), and bootstrapping such a pair into a nearly optimal one; see Remark 3.12 for one particular method for obtaining a starting pair (γ_1, γ_2). The specific steps of this recipe, which mirror the steps taken for the example in Section 3.3, are as follows:

Step 1. **Hellinger bound.** If the density class is log-concave (cf. Definition 3.5), or one can prove that $\theta \mapsto d_H^2(\theta, \theta_\star)$ is convex in the *parameter space*, we estimate the covering number $\mathcal{N}_{\mathcal{I}_{\max}} \equiv \mathcal{N}_{\mathcal{I}_{\max}}(\mathcal{P}, \varepsilon)$ at resolution $\varepsilon \asymp \delta/\sqrt{m}$ for $\mathcal{P}$, and apply Theorem 3.6, specifically (3.16), with $\eta \asymp (\mathrm{diam}(\Theta)\sqrt{m})^{-1}$ to obtain the following event $\mathcal{E}_1$ that holds with probability at least $1 - \delta$ for a universal c_0:

$$\sup_{\theta \in \mathrm{conv}\{\hat{\theta}_{m,T}^\varepsilon, \theta_\star\}} d_H^2(\theta, \theta_\star) \lesssim m^{-1} \log(c_0 m \cdot \mathrm{diam}(\Theta)\mathcal{N}_{\mathcal{I}_{\max}}/\delta). \tag{3.38}$$

Otherwise without log-concavity of $\mathcal{P}$, we rely on Proposition 3.3 to derive the bound:

$$\sup_{\theta \in \mathrm{conv}\{\hat{\theta}_{m,T}^\varepsilon, \theta_\star\}} d_H^2(\theta, \theta_\star) \leqslant \sup_{\theta \in \mathrm{conv}\{\hat{\theta}_{m,T}^\varepsilon, \theta_\star\}} \frac{1}{4} d_{\mathrm{FI}}^2(\theta, \theta_\star) \leqslant \frac{\lambda_{\max}(\mathcal{I}_{\max})}{4} \|\hat{\theta}_{m,T}^\varepsilon - \theta_\star\|^2, \tag{3.39}$$

Next, under the assumption of (γ_1, γ_2)-identifiability (cf. Definition 3.11), the previous inequality implies:

$$d_H^2(\hat{\theta}_{m,T}^\varepsilon, \theta_\star) \leqslant \gamma_1^2 \implies \sup_{\theta \in \mathrm{conv}\{\hat{\theta}_{m,T}^\varepsilon, \theta_\star\}} d_H^2(\theta, \theta_\star) \leqslant \frac{\gamma_2^2 \lambda_{\max}(\mathcal{I}_{\max})}{4} \cdot d_H^2(\hat{\theta}_{m,T}^\varepsilon, \theta_\star). \tag{3.40}$$

Applying Theorem 3.6, specifically (3.15), we obtain the event $\mathcal{E}_1$ with probability $1 - \delta$:

$$d_H^2(\hat{\theta}_{m,T}^\varepsilon, \theta_\star) \lesssim m^{-1} \log(c_0 \mathcal{N}_{\mathcal{I}_{\max}}/\delta).$$

Hence as long as

$$m \gtrsim \gamma_1^{-2} \cdot \log(c_0 \mathcal{N}_{\mathcal{I}_{\max}}/\delta), \tag{3.41}$$

then on $\mathcal{E}_1$:

$$\sup_{\theta \in \mathrm{conv}\{\hat{\theta}_{m,T}^\varepsilon, \theta_\star\}} d_H^2(\theta, \theta_\star) \lesssim m^{-1} \gamma_2^2 \lambda_{\max}(\mathcal{I}_{\max}) \log(c_0 \mathcal{N}_{\mathcal{I}_{\max}}/\delta). \tag{3.42}$$

Step 2. **Estimate B_1 and B_2.** We next compute upper bounds for $B_1 \equiv B_1(\hat{\theta}_{m,T}^\varepsilon, \theta_\star)$ and $B_2 \equiv B_2(\hat{\theta}_{m,T}^\varepsilon, \theta_\star)$ defined in (3.20) and (3.21). As these quantities are random variables due to the presence of $\hat{\theta}_{m,T}^\varepsilon$, we often approximate B_1, B_2 by taking a supremum over a larger set $\Theta' \subseteq \Theta$ for which we can ensure that $\mathrm{conv}\{\hat{\theta}_{m,T}^\varepsilon, \theta_\star\} \subseteq \Theta'$ on $\mathcal{E}_1$. For some problems, it suffices to take $\Theta' = \Theta$. However, for other problems, sharper estimates can be derived by more refined Θ'. We will provide examples of both in the sequel.

Control of B_1 relies on the fact that $\nabla_\theta \log p_\theta(z_{1:T})$ forms a martingale in the realizable setting, and utilizes estimates from e.g., Burkholder's martingale Rosenthal inequality (see Theorem A.6 for a precise statement). Control of B_2 is often more straightforward, and a simple triangle inequality often suffices. Regarding scaling of B_1, B_2, in the examples we work through in the sequel both scale at most poly-logarithmically in T.

Step 3. **Parameter error bound.** Once B_1, B_2 are controlled, then the upper bound on $\sup_{\theta \in \mathrm{conv}\{\hat\theta^\varepsilon_{m,T}, \theta_\star\}} \mathrm{d}_H(\theta, \theta_\star)$ (which holds on $\mathcal{E}_1$) derived in Step 1 can be used to establish condition (3.22). Concretely, this is done via a requirement on the minimum number of trajectories m. For the case when $\mathcal{P}$ is log-concave, this requirement scales as

$$m \gtrsim \max\{B_1^4, B_2^2\} \log(c_0 m \cdot \mathrm{diam}(\Theta) \mathcal{N}_{\mathcal{I}_{\max}}/\delta), \tag{3.43}$$

whereas in the general case the trajectory requirement scales as

$$m \gtrsim \max\{B_1^4, B_2^2\} \gamma_2^2 \lambda_{\max}(\mathcal{I}_{\max}) \log(c_0 \mathcal{N}_{\mathcal{I}_{\max}}/\delta). \tag{3.44}$$

Given condition (3.22), Proposition 3.9 yields the following bound on the parameter error:

$$\|\hat\theta^\varepsilon_{m,T} - \theta_\star\|^2_{\mathcal{I}_2(\theta_\star, \hat\theta^\varepsilon_{m,T})} \lesssim m^{-1} \log(c_0 \mathcal{N}_{\mathcal{I}_{\max}}/\delta). \tag{3.45}$$

While this rate is still not quite the CLT rate (3.4) as the dependence on $\mathcal{I}_2(\theta_\star, \hat\theta^\varepsilon_{m,T})$ is not necessarily variance optimal, for many practical applications this rate may be sufficient, especially if it is possible to show that $\mathcal{I}_2(\theta_\star, \hat\theta^\varepsilon_{m,T}) \gtrsim \Omega(T) \cdot I_p$ on $\mathcal{E}_1$.

Step 4. **Verify FI radius.** In order to apply the second part Proposition 3.9 (i.e., obtain the bound (3.25)), the FI radius condition (3.24) remains to be verified for $\theta_0 = \theta_\star$ and $\theta_1 = \hat\theta^\varepsilon_{m,T}$. To do this, we often rely on the following upper bound:

$$\sup_{\theta \in \mathrm{conv}\{\theta_\star, \hat\theta^\varepsilon_{m,T}\}} \|\mathcal{I}(\theta_\star)^{-1/2} \mathcal{I}(\theta) \mathcal{I}(\theta_\star)^{-1/2} - I_p\|_{\mathrm{op}} \leq \sup_{\theta \in \mathrm{conv}\{\theta_\star, \hat\theta^\varepsilon_{m,T}\}} \frac{\|\mathcal{I}(\theta) - \mathcal{I}(\theta_\star)\|_{\mathrm{op}}}{\lambda_{\min}(\mathcal{I}(\theta_\star))}.$$

We then proceed to show a bound on $\|\mathcal{I}(\theta_0) - \mathcal{I}(\theta_1)\|_{\mathrm{op}}$ for $\theta_0, \theta_1 \in \Theta$ of the following form (see Proposition A.4 for a precise statement):

$$\|\mathcal{I}(\theta_0) - \mathcal{I}(\theta_1)\|_{\mathrm{op}} \lesssim T \left[L\|\theta_0 - \theta_1\| + B_{\mathcal{I}} \mathrm{d}_H(\theta_0, \theta_1) \right],$$

for suitable Lipschitz-like constants $L, B_{\mathcal{I}}$. This implies that

$$\sup_{\theta \in \mathrm{conv}\{\theta_\star, \hat\theta^\varepsilon_{m,T}\}} \frac{\|\mathcal{I}(\theta) - \mathcal{I}(\theta_\star)\|_{\mathrm{op}}}{\lambda_{\min}(\mathcal{I}(\theta_\star))} \lesssim \frac{1}{\lambda_{\min}(\bar{\mathcal{I}}(\theta_\star))} \left[L\|\hat\theta^\varepsilon_{m,T} - \theta_\star\| + B_{\mathcal{I}} \sup_{\theta \in \mathrm{conv}\{\theta_\star, \hat\theta^\varepsilon_{m,T}\}} \mathrm{d}_H(\theta, \theta_\star) \right].$$

The RHS of this expression can be bounded by combining either (3.38) or (3.42) (depending on which one holds) with (3.45). Altogether, we have that condition (3.24) holds given a minimum amount of trajectories m when $\mathcal{P}$ is log-concave:

$$mT \gtrsim \frac{L^2 \log(c_0 \mathcal{N}_{\mathcal{I}_{\max}}/\delta)}{\lambda_{\min}^2(\bar{\mathcal{I}}(\theta_\star))\underline{\mu}} \quad \text{and} \quad m \gtrsim \frac{B_{\mathcal{I}}^2 \log(c_0 m \cdot \mathrm{diam}(\Theta) \mathcal{N}_{\mathcal{I}_{\max}}/\delta)}{\lambda_{\min}^2(\bar{\mathcal{I}}(\theta_\star))}, \tag{3.46}$$

24

where $\underline{\mu} := \lambda_{\min}(\mathcal{I}_2(\theta_\star, \hat\theta^\varepsilon_{m,T}))/T$. On the other hand, if $\mathcal{P}$ is not log-concave, we have that the sufficient condition for (3.24) to hold is

$$mT \gtrsim \frac{L^2 \log(c_0 \mathcal{N}_{\mathcal{I}_{\max}}/\delta)}{\lambda^2_{\min}(\bar{\mathcal{I}}(\theta_\star))\underline{\mu}} \quad \text{and} \quad m \gtrsim \frac{B_\mathcal{I}^2 \gamma_2^2 \lambda_{\max}(\mathcal{I}_{\max}) \log(c_0 \mathcal{N}_{\mathcal{I}_{\max}}/\delta)}{\lambda^2_{\min}(\bar{\mathcal{I}}(\theta_\star))}. \tag{3.47}$$

Step 5. **Final result.** Combining all previous steps, we have that as long as:

(i) *For log-concave $\mathcal{P}$:* the conditions (3.43) and (3.46) on m hold,

(ii) *For non-log-concave $\mathcal{P}$:* the conditions (3.41), (3.44), and (3.47) on m hold,

then we obtain the following rate with probability at least $1 - \delta$:

$$\|\hat\theta^\varepsilon_{m,T} - \theta_\star\|^2_{\bar{\mathcal{I}}(\theta_\star)} \lesssim \frac{\log(c_0 \mathcal{N}_{\mathcal{I}_{\max}}/\delta)}{mT}.$$

For the parametric function classes considered in this work, we will generally have $\log(\mathcal{N}_{\mathcal{I}_{\max}}/\delta) \lesssim p \log(mT/\delta)$ due to the standard volumetric estimate, which yields the CLT rate (3.4) up to logarithmic factors. However, depending on the properties of the stochastic process generated by p_θ, in particular the growth rate of the typical realization of $z_{1:T}$, the dependence on T may be worse; an example of this will be given in Section 4.2.

We conclude by making a brief remark regarding the required scaling on m. For the log-concave $\mathcal{P}$ case, the required conditions generically yield the form $m \gtrsim \mathrm{polylog}(T)$ (ignoring all other problem parameters) whereas for the non-log-concave $\mathcal{P}$ case, the scaling requirement increases to $m \gtrsim T \cdot \mathrm{polylog}(T)$. For the latter case, we believe the linear scaling in T to be an artifact of our analysis strategy, in particular the step taken in (3.39). We leave improving this step to future work.

Remark 3.12 (Single-step Hellinger Identifiability). One simple method we utilize to obtain sub-optimal—but *problem specific*—Hellinger identifiability (cf. Definition 3.11) constants is through the use of the *data processing inequality* (DPI) for f-divergences. Suppose that $z_1 \sim \rho_1$ for all $\theta \in \Theta$. Then we have:

$$\mathbb{E}_{z_1 \sim \rho_1}[\mathrm{d}_H^2(p_\theta(z_2 \mid z_1), p_{\theta_\star}(z_2 \mid z_1))] = \mathrm{d}_H^2(p_\theta(z_{1:2}), p_{\theta_\star}(z_{1:2})) \leq \mathrm{d}_H^2(p_\theta, p_{\theta_\star}),$$

where the equality holds from [39, Prop. 7.2] and the inequality is the DPI for f-divergences [cf. 39, Thm. 7.4]. While the inequality above is often lossy, it is in practice often much easier to prove identifiability using the single-step distributions, i.e.,

$$\mathbb{E}_{z_1 \sim \rho_1}[\mathrm{d}_H^2(p_\theta(z_2 \mid z_1), p_{\theta_\star}(z_2 \mid z_1))] \leq \gamma_1^2 \implies \|\theta - \theta_\star\|^2 \leq \gamma_2^2 \cdot \mathbb{E}_{z_1 \sim \rho_1}[\mathrm{d}_H^2(p_\theta(z_2 \mid z_1), p_{\theta_\star}(z_2 \mid z_1))].$$

We will show several examples of this in the sequel.

The remainder of this paper is dedicated to realizing the Hellinger localization framework on a diverse set of estimation problems, which we turn to in Section 4.

4   Case Studies

This section contains the four multi-trajectory parameter recovery with ERM case studies that we consider in this work: (i) a mixture of two-state Markov chains (Section 4.1), (ii) a linear regression from dependent covariates setup with general (i.e., non-Gaussian) product-noises (Section 4.2), (iii) a GLM setup with a non-expansive, non-monotonic activation function (Section 4.3), and (iv) a simple linear-attention sequence model (Section 4.4). The problem setup and analysis for each case study is fairly self-contained, and can be read in any order.

4.1   Mixture of Two-State Markov Chains

We build on the example from Section 3.3 by considering the following mixture formulation. Suppose we have two Markov chains $M^{(0)}$, $M^{(1)}$, and a Bernoulli distribution P on $\{0,1\}$. The generative process we consider proceeds by first sampling $B \sim P$ and $z_1 \sim \rho_1$ independently, and then generating $z_{t+1} \mid z_t, B$ from $M_{z_t,z_{t+1}}^{(B)}$. The goal is to recover the parameters for the two Markov chains $M^{(0)}$, $M^{(1)}$ given m trajectories of length T from this process ($\mathcal{D}_{m,T}$), where $B^{(i)}$ is unobserved for each trajectory i. Such a problem is a special case of learning from mixtures of Markov chains [58, 59]. Our motivation for studying this problem is two-fold: (a) the parameters of both Markov chains clearly cannot be learned in a single-trajectory setting, necessitating a multi-trajectory approach, and (b) the trajectory process $\{z_t\}$ is *not* α-mixing, but we can still apply the Hellinger localization framework to derive sharp rates directly for the MLE.

Let us define $\mathcal{P}$ as instances of this Markov chain mixture with transition matrices:

$$M^{(0)} = \begin{pmatrix} \theta_0 & 1 - \theta_0 \\ 1 - \theta_0 & \theta_0 \end{pmatrix}, \quad M^{(1)} = \begin{pmatrix} \theta_1 & 1 - \theta_1 \\ 1 - \theta_1 & \theta_1 \end{pmatrix}, \tag{4.1}$$

where $\theta = (\theta_0, \theta_1) \in \Theta := [\mu, 1 - \mu]^2$ with $0 < \mu < 1/2$, $P = \mathrm{Bern}(1/2)$, and $\rho_1 = \mathrm{Unif}(\{1,2\})$. In the remainder of the section, we will use $\mathbb{P}_\theta$ to denote a trajectory measure on the space of $z_{1:\infty}$ realized by a fixed parameter $\theta \in \Theta$, and for $i = 0, 1$, $\mathbb{P}_\theta^{(i)}(\cdot) := \mathbb{P}_\theta(\cdot \mid B = i)$. We further use $\mathbb{E}_\theta$ and $\mathbb{E}_\theta^{(i)}$ to denote expectation under $\mathbb{P}_\theta$ and $\mathbb{P}_\theta^{(i)}$ respectively, so that $\mathbb{E}_\theta[X] = \frac{1}{2}\left(\mathbb{E}_\theta^{(0)}[X] + \mathbb{E}_\theta^{(1)}[X]\right)$ for any random variable X.

Mixtures are not α-mixing.   We give a short argument illustrating the lack of α-mixing for the process $\{z_t\}$. We fix a $p_\theta \in \mathcal{P}$ with $\theta_0 \neq \theta_1$. First, we recall the definition of α-mixing and introduce some notation. For $a \leqslant b$, we let $z_{a:b} = (z_a, \ldots, z_b)$, and we let $\sigma(z_{a:b})$ denote the σ-algebra generated by the subsequence $z_{a:b}$. The α-mixing coefficients are defined as (cf. [8, Eq. 2.2], [9, Def. 2.2]):

$$\alpha(k) := \sup_{j \in \mathbb{N}_+} \sup \left\{ |\mathbb{P}_\theta(A_j \cap B_{j,k}) - \mathbb{P}_\theta(A_j)\mathbb{P}_\theta(B_{j,k})| \mid A_j \in \sigma(z_{1:j}), B_{j,k} \in \sigma(z_{j+k:\infty}) \right\}, \quad k \in \mathbb{N}_+. \tag{4.2}$$

The process $\{z_t\}$ is denoted α-mixing if $\alpha(k) \to 0$ as $k \to \infty$. We consider α-mixing in this section, as it is the *weakest* notion of dependency used in the literature; in particular it is known that ψ-mixing $\Rightarrow \phi$-mixing $\Rightarrow \beta$-mixing $\Rightarrow \alpha$-mixing, and furthermore ρ-mixing $\Rightarrow \alpha$-mixing as well [8]. We now proceed as follows. A simple computation shows that for $A_j \in \sigma(z_{1:j}), B_{j,k} \in \sigma(z_{j+k:\infty})$:

$$\mathbb{P}_\theta(B_{j,k} \mid A_j) - \mathbb{P}_\theta(B_{j,k}) = \left(\mathbb{P}_\theta(B = 0 \mid A_j) - \frac{1}{2}\right)\left(\mathbb{P}_\theta^{(0)}(B_{j,k}) - \mathbb{P}_\theta^{(1)}(B_{j,k})\right) + \Delta(A_j, B_{j,k}),$$

where

$$\Delta(A_j, B_{j,k}) := \mathbb{P}_\theta(B = 0 \mid A_j) \left[\mathbb{P}_\theta^{(0)}(B_{j,k} \mid A_j) - \mathbb{P}_\theta^{(0)}(B_{j,k}) \right]$$
$$+ \mathbb{P}_\theta(B = 1 \mid A_j) \left[\mathbb{P}_\theta^{(1)}(B_{j,k} \mid A_j) - \mathbb{P}_\theta^{(1)}(B_{j,k}) \right].$$

Hence, by triangle inequality,

$$|\mathbb{P}_\theta(A_j \cap B_{j,k}) - \mathbb{P}_\theta(A_j)\mathbb{P}_\theta(B_{j,k})|$$
$$\geqslant \mathbb{P}_\theta(A_j) \left| \mathbb{P}_\theta(B = 0 \mid A_j) - \frac{1}{2} \right| \left| \mathbb{P}_\theta^{(0)}(B_{j,k}) - \mathbb{P}_\theta^{(1)}(B_{j,k}) \right| - |\Delta(A_j, B_{j,k})|.$$

We now select $j = 2$, and let $k \in \mathbb{N}_+$. We define the events A_2 and $B_{2,k}$ to be:

$$A_2 := \{z_{1:2} = (1,1)\}, \quad B_{2,k} := \{z_{2+k:2+k+1} = (1,1)\}.$$

We next observe that for $i \in \{0, 1\}$,

$$\mathbb{P}_\theta^{(i)}(B_{2,k}) = \mathbb{P}_\theta^{(i)}(z_{2+k} = 1, z_{2+k+1} = 1) = \frac{\theta_i}{2},$$

and hence $|\mathbb{P}_\theta^{(0)}(B_{j,k}) - \mathbb{P}_\theta^{(1)}(B_{j,k})| = |\theta_0 - \theta_1|/2$. We also note that $\lim_{k\to\infty} |\Delta(A_2, B_{2,k})| = 0$, since we have $\lim_{k\to\infty} |\mathbb{P}_\theta^{(i)}(z_{j+k} = 1 \mid z_j = 1) - 1/2| = 0$ by the ergodicity of the individual Markov chains $M^{(0)}, M^{(1)}$ [see e.g., 60]. On the other hand,

$$\mathbb{P}_\theta(B = 0 \mid A_2) = \frac{\mathbb{P}_\theta^{(0)}(A_2)}{\mathbb{P}_\theta^{(0)}(A_2) + \mathbb{P}_\theta^{(1)}(A_2)} = \frac{\theta_0}{\theta_0 + \theta_1},$$

and hence $|\mathbb{P}_\theta(B = 0 \mid A_2) - 1/2| > 0$ as $\theta_0 \neq \theta_1$. Therefore, we have

$$\liminf_{k\to\infty} \alpha(k) \geqslant \liminf_{k\to\infty} \left(\mathbb{P}_\theta(A_2) \left| \frac{\theta_0}{\theta_0 + \theta_1} - \frac{1}{2} \right| \frac{|\theta_0 - \theta_1|}{2} - |\Delta(A_2, B_{2,k})| \right)$$
$$= \mathbb{P}_\theta(A_2) \left| \frac{\theta_0}{\theta_0 + \theta_1} - \frac{1}{2} \right| \frac{|\theta_0 - \theta_1|}{2} - \limsup_{k\to\infty} |\Delta(A_2, B_{2,k})|$$
$$= \mathbb{P}_\theta(A_2) \left| \frac{\theta_0}{\theta_0 + \theta_1} - \frac{1}{2} \right| \frac{|\theta_0 - \theta_1|}{2} > 0,$$

and hence we have that $\{z_t\}$ is not α-mixing.

Remark 4.1. Although $\{z_t\}_{t=1}^\infty$ is not α-mixing, it is ergodic as $M^{(0)}$ and $M^{(1)}$ admit the same stationary distribution, which implies any time average of single trajectory will converge to the same marginal expectation. In general, the mixture of Markov chain process will be non-ergodic and non-mixing as long as the candidate transition matrices have different stationary distributions.

Remark 4.2. We remark that mixing coefficients are not fully standardized in the literature, and depending on what specific definition is adopted, the trajectory $\{z_t\}$ could be considered mixing. As a specific example, in [12] a weaker definition of β-mixing is considered, defined as $\beta(k) = \sup_{t \geqslant 1} \mathbb{E}_{z_{1:t}} [\|\mathbb{P}_{z_{t+k}}(\cdot \mid z_{1:t}) - \mathbb{P}_{z_{t+k}}(\cdot)\|_{\mathrm{TV}}]$. Under this definition $\{z_t\}$ is actually β-mixing, since the $M^{(i)}$'s admit the same stationary distribution. However, there are many ways to modify

the mixture model (4.1) so that it is not β-mixing under this more relaxed definition. A simple modification is

$$M^{(0)} = \begin{pmatrix} \theta_0 & 1 - \theta_0 \\ 1 - \theta_0' & \theta_0' \end{pmatrix}, \quad M^{(1)} = \begin{pmatrix} \theta_1 & 1 - \theta_1 \\ 1 - \theta_1' & \theta_1' \end{pmatrix},$$

where $\theta_i' = \theta_i + \tau$ (modulating by one if necessary), where τ is a fixed offset. Another option is to consider general two-state Markov chains (so that θ contains four total parameters). For both modifications, the structure of the proof to be presented remains the same, although the detailed calculations may be different, especially for the general two-state parameterization.

Towards stating our main result, we define the following quantities for $\theta \in (0,1)^2$:

$$\mathrm{Gap}(\theta) := |\theta_0 - \theta_1|, \quad \bar{\sigma}^2(\theta) := \max_{i=0,1} \sigma_i^2(\theta), \quad \underline{\sigma}^2(\theta) := \min_{i=0,1} \sigma_i^2(\theta).$$

The following is our main parameter recovery bound for the mixture of Markov chains problem.

Theorem 4.3. *Fix $\delta \in (0,1)$, and define the constants $\rho_\star := \mathrm{Gap}(\theta_\star)$ and $\sigma_{\min}^2 := \mu(1-\mu)$. Suppose that the following conditions hold:*

(a) $\theta_{\star,0} > \theta_{\star,1}$,

(b) $T \gtrsim \max\left\{ \frac{\bar{\sigma}^2(\theta_\star)}{\rho_\star^4}, \frac{1}{\rho_\star^2} \right\} \log^2(1/\mu) \log\left(\frac{\bar{\sigma}^2(\theta_\star)}{\underline{\sigma}^4(\theta_\star)} \cdot \frac{\log(1/\mu)}{\rho_\star} \right)$,

(c) $m \gtrsim \max\left\{ \frac{1}{\rho_\star^4}, \frac{1}{\rho_\star^2 \underline{\sigma}^4(\theta_\star)}, \frac{T}{\rho_\star^2 \underline{\sigma}^2(\theta_\star)} \right\} \log\left(\max\left\{ \frac{1}{\rho_\star^4}, \frac{1}{\rho_\star^2 \underline{\sigma}^4(\theta_\star)}, \frac{T}{\rho_\star^2 \underline{\sigma}^2(\theta_\star)} \right\} \frac{T}{\sigma_{\min}^2 \delta} \right)$,

Let $\hat{\theta}_{m,T}^\varepsilon$ denote the max FI discretized MLE estimator (3.14) at resolution $\varepsilon = \delta/(2\sqrt{2m})$, and suppose the MLE estimator satisfies $\left(\hat{\theta}_{m,T}^\varepsilon \right)_0 \geq \left(\hat{\theta}_{m,T}^\varepsilon \right)_1$. With probability at least $1 - \delta$,

$$\|\hat{\theta}_{m,T}^\varepsilon - \theta_\star\|_{\mathcal{I}(\theta_\star)}^2 \lesssim \frac{1}{mT} \log\left(\frac{mT}{\sigma_{\min}^2 \delta} \right) \quad and \quad \|\hat{\theta}_{m,T}^\varepsilon - \theta_\star\|^2 \lesssim \frac{\bar{\sigma}^2(\theta_\star)}{mT} \log\left(\frac{mT}{\sigma_{\min}^2 \delta} \right). \tag{4.3}$$

Some remarks are in order for Theorem 4.3. First, we note that Assumption (a) in Theorem 4.3 does not change the generality of the result, as the distribution $p_\theta(z_{1:T})$ is invariant under parameter permutation (since B is sampled uniformly over the two choices), and hence we can assume wlog that $\theta_{\star,0} > \theta_{\star,1}$. Furthermore, given an MLE $\hat{\theta}_{m,T}^\varepsilon$, we can always assume wlog $\left(\hat{\theta}_{m,T}^\varepsilon \right)_0 > \left(\hat{\theta}_{m,T}^\varepsilon \right)_1$, otherwise we just permute the estimator. Second, we remark that Theorem 4.3 is nearly fully *instance dependent*, i.e., both the requirements m, T and the final parameter error bound on do not involve the global bound μ outside of poly-logarithmic factors, but instead depend on the problem-specific parameters $\rho_\star, \underline{\sigma}(\theta_\star), \bar{\sigma}(\theta_\star)$ of the true, data-generating distribution. Third, the rate prescribed by (4.3) is in general not improvable, as it matches the two-state Markov chain argument in Section 3.3, specifically the optimal rate in (3.37).

We now discuss the requirements on T via Assumption (b), and m via Assumption (c). Starting with the requirement on T in Assumption (b), the role of this condition is to ensure that there is sufficient information within a trajectory to distinguish which chain most likely generated the data *if the true parameters were known*; hence the scaling of $\mathrm{poly}(1/\mathrm{Gap}(\theta_\star))$ is quite intuitive. On the other hand, the requirement on m in Assumption (c) parallels that of (3.35) in the two-state Markov chain case (cf. Section 3.3). The biggest difference is that in Assumption (c), we have the scaling of $m \gtrsim \tilde{\Omega}(T)$ instead of $m \gtrsim \tilde{\Omega}(1)$ in (3.35). This comes the non-concavity of the MLE for the mixture problem, which required us to use (3.44), compared with the concave MLE for the two-state mixture; as noted in Section 3.4, the requirements on m are worse for non-concave problems.

Comparison to existing results. Learning mixture of Markov chains has been recently studied by a few authors [58, 59, 61]. From this set of works, most related to ours is [59], where the authors develop efficient algorithms for clustering and estimating the family of transition matrices. Adapted to our specific setting, [59, Theorem 4] reads that when:

$$m \gtrsim \frac{\tau_{\mathrm{mix}} \log(1/\delta)}{\rho_\star^3}, \quad T \gtrsim \tau_{\mathrm{mix}} \log(1/\delta), \quad \tau_{\mathrm{mix}} := \max_{i=0,1} \frac{1}{\min\{\theta_{\star,i}, 1-\theta_{\star,i}\}},$$

then their algorithm recovers an estimate $\hat{\theta}_{m,T}$ that satisfies with probability at least $1 - \delta$:

$$\|\hat{\theta}_{m,T} - \theta_\star\|^2 \lesssim \frac{\tau_{\mathrm{mix}}^{2/3}}{T^{2/3}} \frac{\log(1/\delta)}{m}. \tag{4.4}$$

We note that although the result from [59] has less stringent requirements on both the minimum number of trajectories m and trajectory length T compared with Theorem 4.3, the final rate (4.4) has both (i) a $1/(mT^{2/3})$ scaling in comparison to a $1/(mT)$ scaling in (4.3), and (ii) also scales proportion to $\tau_{\mathrm{mix}}^{2/3}$ as opposed to $\bar{\sigma}^2(\theta_\star)$ in (4.3); note that in general τ_{mix} can grow arbitrarily large as $\theta_\star$ approaches the boundary of the positive orthant, whereas $\bar{\sigma}^2(\theta_\star) \leqslant 1/4$ always. On the other hand, as mentioned previously, the work [59] provides an efficient algorithm which can also learn the distribution of the latent variable B, whereas our result Theorem 4.3 uses the MLE estimate which, in this case, requires maximizing a non-concave objective and does not handle the case where the distribution of B must be jointly learned. Extensions of our analysis to more general mixture setups, in addition to practical algorithms such as expectation maximization [62], is left as interesting future work. In Section 4.1.3, we comment in more detail on how our proof techniques may be generalized to other mixture recovery problems.

4.1.1 Preliminary Results for Theorem 4.3

Our analysis resembles that of the two-state Markov chain case (cf. Section 3.3), given that each trajectory can be associated with a particular chain if the trajectory length T is sufficiently long. In the following, we state a few auxiliary results that will be crucial towards enabling our analysis. The following subset $\Theta' \subset \Theta$ plays an important role in localizing the problem-specific parameters:

$$\Theta' := \left\{ \theta \in \Theta \mid \|\theta - \theta_\star\| \leqslant \min\{\mathrm{Gap}(\theta_\star)/(2\sqrt{2}), \underline{\sigma}^2(\theta_\star)/2\} \right\}. \tag{4.5}$$

Proposition 4.4. *Fix* $\theta = (\theta_0, \theta_1) \in (0,1)^2$ *and suppose that* $\theta_0 \neq \theta_1$. *Define for* $i \in \{0,1\}$:

$$\Delta_i(\theta) := \mathrm{KL}(\mathrm{Bern}(\theta_i) \| \mathrm{Bern}(\theta_{1-i})), \quad \sigma_i^2(\theta) := \theta_i(1-\theta_i).$$

Fix $\varepsilon, \delta \in (0,1)$, *and suppose that* T *satisfies:*

$$T \gtrsim \max_{i=0,1} \max \left\{ \frac{\ell^2(\theta)\sigma_i^2(\theta)\log(2/\delta)}{\Delta_i^2(\theta)}, \frac{1}{\Delta_i(\theta)} \left[\log\left(\frac{1}{\varepsilon}\right) + \ell(\theta)\log\left(\frac{2}{\delta}\right) \right] \right\}, \quad \ell(\theta) := \left| \log\left(\frac{\theta_0(1-\theta_1)}{\theta_1(1-\theta_0)}\right) \right|.$$

Denote the posterior density of B *evaluated at* 0:

$$w_\theta(z_{1:T}) := p(B = 0 \mid z_{1:T}) = \frac{p_{\theta_0}(z_{1:T})}{p_{\theta_0}(z_{1:T}) + p_{\theta_1}(z_{1:T})}.$$

We have:

$$\mathbb{P}_\theta^{(0)}\left(w_\theta(z_{1:T}) \geqslant 1 - \varepsilon \right) \geqslant 1 - \delta, \quad \mathbb{P}_\theta^{(1)}\left(w_\theta(z_{1:T}) \leqslant \varepsilon \right) \geqslant 1 - \delta.$$

Proof. To ease notation, we let $T' := T - 1$ for what follows. Given $z_{1:T}$, we have that

$$\log r_\theta(z_{1:T}) := \log \frac{p_{\theta_0}(z_{1:T})}{p_{\theta_1}(z_{1:T})} = \log\left(\frac{\theta_0}{\theta_1}\right) \cdot \sum_{t=1}^{T'} \mathbb{1}\{z_{t+1} = z_t\} + \log\left(\frac{1-\theta_0}{1-\theta_1}\right) \cdot \sum_{t=1}^{T'} \mathbb{1}\{z_{t+1} \neq z_t\}$$

$$=: \log\left(\frac{\theta_0}{\theta_1}\right) \cdot N_{\text{stay}}(z_{1:T}) + \log\left(\frac{1-\theta_0}{1-\theta_1}\right) \cdot (T' - N_{\text{stay}}(z_{1:T})).$$

Next, since conditioned on $B = i$, the random variable $N_{\text{stay}}(z_{1:T}) \sim \text{Bin}(T', \theta_i)$, and therefore the conditional MGF of $N_{\text{stay}}(z_{1:T})$ is:

$$\mathbb{E}_\theta^{(i)}[\exp(\lambda N_{\text{stay}}(z_{1:T}))] = (e^\lambda \theta_i + (1-\theta_i))^{T'}.$$

Hence, we have

$$\log \mathbb{E}_\theta^{(i)}[\exp(\lambda(N_{\text{stay}}(z_{1:T}) - T'\theta_i)] = T' \log(e^\lambda \theta_i + (1-\theta_i)) - T'\theta_i \lambda.$$

By Proposition A.9, we have that with probability at least $1 - \delta$ over $z_{1:T} \sim p_{\theta_i}$,

$$\left|N_{\text{stay}}(z_{1:T}) - T'\theta_i\right| \leq 2\sqrt{2eT'\sigma_i^2(\theta) \log(2/\delta)} + 2\log(2/\delta).$$

Let us temporarily call this event $\mathcal{E}$. On event $\mathcal{E}$ under p_{θ_0},

$$\left|\log r_\theta(z_{1:T}) - T'\text{KL}(\text{Bern}(\theta_0) \parallel \text{Bern}(\theta_1))\right|$$

$$= \left|\log\left(\frac{\theta_0}{\theta_1}\right)\left(N_{\text{stay}}(z_{1:T}) - T'\theta_0\right) + \log\left(\frac{1-\theta_0}{1-\theta_1}\right)\left(T'\theta_0 - N_{\text{stay}}(z_{1:T})\right)\right|$$

$$= \left|\log\left(\frac{\theta_0(1-\theta_1)}{\theta_1(1-\theta_0)}\right)\right| |N_{\text{stay}}(z_{1:T}) - T'\theta_0|$$

$$\lesssim \left|\log\left(\frac{\theta_0(1-\theta_1)}{\theta_1(1-\theta_0)}\right)\right| \left(\sqrt{T\sigma_0^2(\theta) \log(2/\delta)} + \log(2/\delta)\right)$$

$$=: \Psi_1^{(0)}\sqrt{T} + \Psi_2^{(0)},$$

where

$$\Psi_1^{(0)} := \ell(\theta)\sqrt{\sigma_0^2(\theta) \log(2/\delta)}, \quad \Psi_2^{(0)} := \ell(\theta) \log(2/\delta).$$

Therefore we have on event $\mathcal{E}$, there exists a universal $c > 0$ such that:

$$\log r_\theta(z_{1:T}) \geq T'\Delta_0(\theta) - c(\Psi_1^{(0)}\sqrt{T} + \Psi_2^{(0)}). \tag{4.6}$$

Now, we observe that for any $\varepsilon \in (0, 1)$,

$$w_\theta(z_{1:T}) \geq 1 - \varepsilon \iff \log r_\theta(z_{1:T}) \geq \log\left(\frac{1-\varepsilon}{\varepsilon}\right).$$

To achieve the claimed result, it remains to derive sufficient conditions on T so that the RHS of (4.6) is lower bounded by $\log((1-\varepsilon)/\varepsilon)$. First, we require that

$$T'\Delta_0(\theta)/2 \geq c\Psi_1^{(0)}\sqrt{T} \iff T \gtrsim (\Psi_1^{(0)})^2/\Delta_0^2(\theta).$$

We also require that

$$T' \Delta_0(\theta)/2 \geqslant \log\left(\frac{1-\varepsilon}{\varepsilon}\right) + c\Psi_2^{(0)} \impliedby T \gtrsim \frac{1}{\Delta_0(\theta)}\left[\log\left(\frac{1}{\varepsilon}\right) + \Psi_2^{(0)}\right].$$

Hence, we have that as long as:

$$T \gtrsim \max\left\{\frac{(\Psi_1^{(0)})^2}{\Delta_0^2(\theta)}, \frac{1}{\Delta_0(\theta)}\left[\log\left(\frac{1}{\varepsilon}\right) + \Psi_2^{(0)}\right]\right\},$$

then we have

$$\mathbb{P}_\theta^{(0)}\{w_\theta(z_{1:T}) \geqslant 1 - \varepsilon\} \geqslant 1 - \delta.$$

The proof for $\mathbb{P}_\theta^{(1)}$ proceeds exactly the same as above with the roles of θ_0, θ_1 swapped. $\qquad\square$

Remark 4.5. We note that one could get a similar result by applying [63, Theorem 3.9]. However, our requirement on T from the above lemma does not depend linearly on the inverse spectral gap $\asymp [\min\{\theta_i, 1 - \theta_i\}]^{-1}$ of the chain defined by individual θ_i's.

Corollary 4.6. *Fix* $\theta_\star = (\theta_{\star,0}, \theta_{\star,1}) \in \Theta$ *and* $\varepsilon, \delta \in (0,1)$. *Suppose* T *satisfies*

$$T \gtrsim \max\left\{\frac{\bar\sigma^2(\theta_\star)}{\mathrm{Gap}^4(\theta_\star)} \cdot \log^2(1/\mu)\log\left(\frac{2}{\delta}\right), \frac{1}{\mathrm{Gap}^2(\theta_\star)}\left[\log\left(\frac{1}{\varepsilon}\right) + \log(1/\mu)\log\left(\frac{2}{\delta}\right)\right]\right\}.$$

Then for any $\theta \in \Theta'$, *we have:*

$$\mathbb{P}_\theta^{(0)}\left(w_\theta(z_{1:T}) \geqslant 1 - \varepsilon\right) \geqslant 1 - \delta, \quad \mathbb{P}_\theta^{(1)}\left(w_\theta(z_{1:T}) \leqslant \varepsilon\right) \geqslant 1 - \delta.$$

Proof. First by Pinsker's inequality, we have:

$$\Delta_i(\theta) = \mathrm{KL}(\mathrm{Bern}(\theta_i) \parallel \mathrm{Bern}(\theta_{1-i})) \geqslant 2\|\mathrm{Bern}(\theta_i) - \mathrm{Bern}(\theta_{1-i})\|_{\mathrm{TV}}^2 = 2\mathrm{Gap}^2(\theta), \quad i \in \{0,1\}.$$

Next, we have that for $\theta \in \Theta'$:

$$\mathrm{Gap}(\theta) = |\theta_0 - \theta_1| = |(\theta_{\star,0} - \theta_{\star,1}) + (\theta_0 - \theta_{\star,0}) - (\theta_1 - \theta_{\star,1})|$$
$$\geqslant \mathrm{Gap}(\theta_\star) - \|\theta - \theta_\star\|_1 \geqslant \mathrm{Gap}(\theta_\star) - \sqrt{2}\|\theta - \theta_\star\| \geqslant \mathrm{Gap}(\theta_\star)/2.$$

Consequently for $\theta \in \Theta'$ and $i \in \{0,1\}$, $\Delta_i(\theta) \geqslant \mathrm{Gap}^2(\theta_\star)/2$. Next, for any $\theta \in \Theta$, we have:

$$\ell(\theta) = \left|\log\left(\frac{\theta_0(1-\theta_1)}{\theta_1(1-\theta_0)}\right)\right| \leqslant \left|\log\left(\frac{1-\theta_0}{\theta_0}\right)\right| + \left|\log\left(\frac{1-\theta_1}{\theta_1}\right)\right| \leqslant 2\max_{\theta \in [\mu, 1-\mu]}\left|\log\left(\frac{1-\theta}{\theta}\right)\right| \leqslant 2\log(1/\mu),$$

since we assumed $\mu < 1/2$. Finally, for any $\theta \in \Theta'$ and $i \in \{0,1\}$,

$$|\sigma_i^2(\theta) - \sigma_i^2(\theta_\star)| \leqslant |\theta_i - \theta_{\star,i}| \leqslant \|\theta - \theta_\star\| \leqslant \sigma_i^2(\theta_\star)/2 \implies \sigma_i^2(\theta) \leqslant 3\sigma_i^2(\theta_\star)/2$$
$$\implies \bar\sigma^2(\theta) \leqslant 3\bar\sigma^2(\theta_\star)/2.$$

The claim now follows from Proposition 4.4. $\qquad\square$

Corollary 4.7. *Fix $\theta_\star = (\theta_{\star,0}, \theta_{\star,1}) \in \Theta$ and $\varepsilon \in (0,1)$, and suppose T satisfies:*

$$T \gtrsim \max\left\{\frac{\bar{\sigma}^2(\theta_\star)}{\mathrm{Gap}^4(\theta_\star)}, \frac{1}{\mathrm{Gap}^2(\theta_\star)}\right\} \log^2(1/\mu) \log\left(\frac{2}{\varepsilon}\right).$$

Then for any $\theta \in \Theta'$, we have:

$$\mathbb{E}_\theta^{(0)}[|w_\theta(z_{1:T}) - 1|] \leqslant \varepsilon, \quad \mathbb{E}_\theta^{(1)}[w_\theta(z_{1:T})] \leqslant \varepsilon.$$

Furthermore, fix any $k \in \mathbb{N}_+$ and $\eta \in (0,1)$, and suppose that T satisfies:

$$T \gtrsim \max\left\{\frac{\bar{\sigma}^2(\theta_\star)}{\mathrm{Gap}^4(\theta_\star)}, \frac{1}{\mathrm{Gap}^2(\theta_\star)}\right\} \log^2(1/\mu) \log\left(\frac{2}{\eta}\left(\frac{k \log(1/\mu)}{\mathrm{Gap}(\theta_\star)}\right)^k\right).$$

Then for any $\theta \in \Theta'$, we have:

$$\mathbb{E}_\theta^{(0)}[|w_\theta(z_{1:T}) - 1|] \leqslant \frac{\eta}{T^k}, \quad \mathbb{E}_\theta^{(1)}[w_\theta(z_{1:T})] \leqslant \frac{\eta}{T^k}.$$

Proof. For the first part of the statement, we temporarily denote $\mathcal{E} := \{|w_\theta(z_{1:T}) - 1| > \varepsilon/2\}$. Then we have, using $w_\theta(z_{1:T}) \in [0,1]$:

$$\mathbb{E}_\theta^{(1)}[|w_\theta(z_{1:T}) - 1|] = \mathbb{E}_\theta^{(1)}[|w_\theta(z_{1:T}) - 1| \cdot \mathbb{1}\{\mathcal{E}\}] + \mathbb{E}_\theta^{(1)}[|w_\theta(z_{1:T}) - 1| \cdot \mathbb{1}\{\mathcal{E}^c\}]$$

$$\leqslant \mathbb{P}_\theta^{(1)}(|w_\theta(z_{1:T}) - 1| > \varepsilon/2) + \varepsilon/2 \leqslant \varepsilon/2 + \varepsilon/2 = \varepsilon,$$

where the last inequality holds from Corollary 4.6, given our requirement on T. The result for $\mathbb{E}_\theta^{(1)}[w_\theta(z_{1:T})] \leqslant \varepsilon$ follows using same proof.

For the second part of the statement, we invoke the first part with $\varepsilon = \eta/T^k$, and use Proposition A.1 to recover the dependence on T. Specifically, we require

$$T \geqslant A \log(2T^k/\eta), \quad A := c_0 \max\left\{\frac{\bar{\sigma}^2(\theta_\star)}{\mathrm{Gap}^4(\theta_\star)}, \frac{1}{\mathrm{Gap}^2(\theta_\star)}\right\} \log^2(1/\mu).$$

It suffices to require that:

$$T/2 \geqslant A \log(2/\eta), \quad T/2 \geqslant kA \cdot \log T.$$

Applying Proposition A.1 to the second inequality it suffices for $T \gtrsim kA \log(kA)$. Hence, simplifying further by bounding $\bar{\sigma}^2(\theta_\star) \leqslant 1/4$ yields the claim. $\square$

Proposition 4.8. *Fix $\theta_\star = (\theta_{\star,0}, \theta_{\star,1}) \in \Theta$. Suppose T satisfies:*

$$T \gtrsim \max\left\{\frac{\bar{\sigma}^2(\theta_\star)}{\mathrm{Gap}^4(\theta_\star)}, \frac{1}{\mathrm{Gap}^2(\theta_\star)}\right\} \log^2(1/\mu) \log\left(\frac{\bar{\sigma}^2(\theta_\star)}{\underline{\sigma}^4(\theta_\star)} \cdot \frac{\log(1/\mu)}{\mathrm{Gap}(\theta_\star)}\right).$$

We have for any $\theta \in \Theta'$:

$$\frac{1}{4}\begin{pmatrix} \mathcal{I}(\theta_0) & 0 \\ 0 & \mathcal{I}(\theta_1) \end{pmatrix} \preccurlyeq \mathcal{I}(\theta) \preccurlyeq \frac{3}{4}\begin{pmatrix} \mathcal{I}(\theta_0) & 0 \\ 0 & \mathcal{I}(\theta_1) \end{pmatrix},$$

where $\mathcal{I}(\theta_0)$ and $\mathcal{I}(\theta_1)$ are the Fisher information of the individual two-state Markov chains under θ_0 and θ_1 respectively (cf. Section 3.3), i.e., $\mathcal{I}(\theta_i) = \frac{T-1}{\theta_i(1-\theta_i)}$ for $i \in \{0,1\}$.

32

Proof. We start with calculating relevant derivatives for our mixture model $\mathcal{P}$. We recall that the parameter space is $\Theta = [\mu, 1-\mu]^2$. Given $\theta \in \Theta$, the log likelihood ratio is

$$\log p_\theta(z_{1:T}) = \log\left(\frac{1}{2}p_{\theta_0}(z_{1:T}) + \frac{1}{2}p_{\theta_1}(z_{1:T})\right) = \log\left(p_{\theta_0}(z_{1:T}) + p_{\theta_1}(z_{1:T})\right) - \log 2.$$

Therefore the first order information is

$$\nabla_\theta \log p_\theta(z_{1:T}) = \left(w_\theta(z_{1:T})\partial_{\theta_0}\log p_{\theta_0}(z_{1:T}) \quad (1 - w_\theta(z_{1:T}))\partial_{\theta_1}\log p_{\theta_1}(z_{1:T})\right)^\mathsf{T}. \tag{4.7}$$

We now compute the second order information:

$$\partial_{\theta_0}^2 \log p_\theta(z_{1:T}) = w_\theta(z_{1:T})(1 - w_\theta(z_{1:T}))\left(\partial_{\theta_0}\log p_{\theta_0}(z_{1:T})\right)^2 + w_\theta(z_{1:T})\partial_{\theta_0}^2\log p_{\theta_0}(z_{1:T}),$$
$$\partial_{\theta_1}^2 \log p_\theta(z_{1:T}) = w_\theta(z_{1:T})(1 - w_\theta(z_{1:T}))\left(\partial_{\theta_1}\log p_{\theta_1}(z_{1:T})\right)^2 + (1 - w_\theta(z_{1:T}))\partial_{\theta_1}^2\log p_{\theta_1}(z_{1:T}),$$
$$\partial_{\theta_1}\partial_{\theta_0}\log p_\theta(z_{1:T}) = \partial_{\theta_0}\partial_{\theta_1}\log p_\theta(z_{1:T}) = w_\theta(z_{1:T})(w_\theta(z_{1:T}) - 1)\partial_{\theta_0}\log p_{\theta_0}(z_{1:T})\partial_{\theta_1}\log p_{\theta_1}(z_{1:T}). \tag{4.8}$$

Denoting

$$H_\theta(z_{1:T}) := \begin{pmatrix} \partial_{\theta_0}^2\log p_\theta(z_{1:T}) & \partial_{\theta_1}\partial_{\theta_0}\log p_\theta(z_{1:T}) \\ \partial_{\theta_0}\partial_{\theta_1}\log p_\theta(z_{1:T}) & \partial_{\theta_1}^2\log p_\theta(z_{1:T}) \end{pmatrix} = \nabla_\theta^2 \log p_\theta(z_{1:T}),$$

we compute the Fisher information of θ as:

$$\mathcal{I}(\theta) = -\mathbb{E}_\theta\left[H_\theta(z_{1:T})\right] = \frac{1}{2}\mathbb{E}_\theta^{(0)}\left[-H_\theta(z_{1:T})\right] + \frac{1}{2}\mathbb{E}_\theta^{(1)}\left[-H_\theta(z_{1:T})\right],$$

We now further define

$$\mathbb{E}_\theta^{(0)}\left[-H_\theta(z_{1:T})\right] = \begin{pmatrix} \mathbb{E}_\theta^{(0)}\left[-\partial_{\theta_0}^2\log p_{\theta_0}(z_{1:T})\right] & 0 \\ 0 & 0 \end{pmatrix} + \mathbb{E}_\theta^{(0)}\left[-\left(H_\theta(z_{1:T}) - \begin{pmatrix} \partial_{\theta_0}^2\log p_{\theta_0}(z_{1:T}) & 0 \\ 0 & 0 \end{pmatrix}\right)\right],$$
$$:= \begin{pmatrix} \mathcal{I}(\theta_0) & 0 \\ 0 & 0 \end{pmatrix} + \mathbb{E}_\theta^{(0)}\left[E_0(\theta)\right],$$
$$\mathbb{E}_\theta^{(1)}\left[-H_\theta(z_{1:T})\right] = \begin{pmatrix} 0 & 0 \\ 0 & \mathbb{E}_\theta^{(1)}\left[-\partial_{\theta_1}^2\log p_{\theta_1}(z_{1:T})\right] \end{pmatrix} + \mathbb{E}_\theta^{(1)}\left[-\left(H_\theta(z_{1:T}) - \begin{pmatrix} 0 & 0 \\ 0 & \partial_{\theta_1}^2\log p_{\theta_1}(z_{1:T}) \end{pmatrix}\right)\right],$$
$$:= \begin{pmatrix} 0 & 0 \\ 0 & \mathcal{I}(\theta_1) \end{pmatrix} + \mathbb{E}_\theta^{(1)}\left[E_1(\theta)\right].$$

We highlight the following expressions for E_0, E_1:

$$(E_0(\theta))_{11} = (1 - w_\theta(z_{1:T}))\left(\partial_{\theta_0}^2\log p_{\theta_0}(z_{1:T}) - w_\theta(z_{1:T})\left(\partial_{\theta_0}\log p_{\theta_0}(z_{1:T})\right)^2\right),$$
$$(E_1(\theta))_{22} = w_\theta(z_{1:T})\left(\partial_{\theta_1}^2\log p_{\theta_1}(z_{1:T}) - (1 - w_\theta(z_{1:T}))\left(\partial_{\theta_1}\log p_{\theta_1}(z_{1:T})\right)^2\right).$$

With this error decomposition, we have that $\mathcal{I}(\theta)$ can be written as:

$$\mathcal{I}(\theta) = \frac{1}{2}\begin{pmatrix} \mathcal{I}(\theta_0) & 0 \\ 0 & \mathcal{I}(\theta_1) \end{pmatrix} + \frac{1}{2}\left(\mathbb{E}_\theta^{(0)}\left[E_0(\theta)\right] + \mathbb{E}_\theta^{(1)}\left[E_1(\theta)\right]\right). \tag{4.9}$$

We next bound:

$$\frac{1}{2}\|\mathbb{E}_\theta^{(0)}\left[E_0(\theta)\right] + \mathbb{E}_\theta^{(1)}\left[E_1(\theta)\right]\|_{\mathrm{op}} \leqslant \frac{1}{2}\left(\|\mathbb{E}_\theta^{(0)}\left[E_0(\theta)\right]\|_{\mathrm{op}} + \|\mathbb{E}_\theta^{(1)}\left[E_1(\theta)\right]\|_{\mathrm{op}}\right)$$

$$\leqslant \frac{1}{2}\left(\mathbb{E}_\theta^{(0)}\left[\|E_0(\theta)\|_{\mathrm{op}}\right]+\mathbb{E}_\theta^{(1)}\left[\|E_1(\theta)\|_{\mathrm{op}}\right]\right)$$

$$\leqslant \mathbb{E}_\theta^{(0)}\left[\sup_{1\leqslant i,j\leqslant 2}\left|(E_0(\theta))_{ij}\right|\right]+\mathbb{E}_\theta^{(1)}\left[\sup_{1\leqslant i,j\leqslant 2}\left|(E_1(\theta))_{ij}\right|\right], \qquad (4.10)$$

where in the penultimate inequality we applied Jensen's inequality, and in the last step we apply the simple inequality $\|A\|_{\mathrm{op}}\leqslant\|A\|_F\leqslant\sqrt{nm}\cdot\max_{i\in[m],j\in[n]}|A_{ij}|$ for $A\in\mathbb{R}^{m\times n}$. We now recall from the two-state Markov chain example (Section 3.3) the following computation for $i=0,1$,

$$\partial_{\theta_i}\log p_{\theta_i}(z_{1:T})=\sum_{t=1}^{T-1}\left(\frac{1}{\theta_i}\mathbb{1}\{z_{t+1}=z_t\}-\frac{1}{1-\theta_i}\mathbb{1}\{z_{t+1}\neq z_t\}\right),$$

$$\partial_{\theta_i}^2\log p_{\theta_i}(z_{1:T})=-\sum_{t=1}^{T-1}\left(\frac{1}{\theta_i^2}\mathbb{1}\{z_{t+1}=z_t\}+\frac{1}{(1-\theta_i)^2}\mathbb{1}\{z_{t+1}\neq z_t\}\right).$$

It is clear that we have

$$|\partial_{\theta_i}\log p_{\theta_i}(z_{1:T})|\leqslant(T-1)/\underline{\sigma}^2(\theta),\quad -(T-1)/\underline{\sigma}^4(\theta)\leqslant\partial_{\theta_i}^2\log p_{\theta_i}(z_{1:T})\leqslant 0,$$

Hence, it is straightforward to show the following almost surely bounds:

$$|\partial_{\theta_0}^2\log p_\theta(z_{1:T})|\leqslant w_\theta(z_{1:T})(T-1)^2/\underline{\sigma}^4(\theta),$$
$$|\partial_{\theta_1}^2\log p_\theta(z_{1:T})|\leqslant(1-w_\theta(z_{1:T}))(T-1)^2/\underline{\sigma}^4(\theta), \qquad (4.11)$$
$$|\partial_{\theta_1}\partial_{\theta_0}\log p_\theta(z_{1:T})|\leqslant w_\theta(z_{1:T})(1-w_\theta(z_{1:T}))(T-1)^2/\underline{\sigma}^4(\theta).$$

From these bounds we can conclude that:

$$\sup_{1\leqslant i,j\leqslant 2}\left|(E_0(\theta))_{ij}\right|\leqslant 2(1-w_\theta(z_{1:T}))(T-1)^2/\underline{\sigma}^4(\theta),\quad \sup_{1\leqslant i,j\leqslant 2}\left|(E_1(\theta))_{ij}\right|\leqslant 2A_\theta(z_{1:T})(T-1)^2/\underline{\sigma}^4(\theta).$$

Next, we define

$$\bar{\mathcal{I}}_s(\theta):=\frac{1}{2}\begin{pmatrix}\bar{\mathcal{I}}(\theta_0) & 0 \\ 0 & \bar{\mathcal{I}}(\theta_1)\end{pmatrix},\quad E_{\bar{\mathcal{I}}(\theta)}:=\bar{\mathcal{I}}(\theta)-\bar{\mathcal{I}}_s(\theta).$$

From (4.10), we have the following bound:

$$\|E_{\bar{\mathcal{I}}(\theta)}\|_{\mathrm{op}}\leqslant 2(T-1)/\underline{\sigma}^4(\theta)\cdot\underbrace{\left(\mathbb{E}_\theta^{(0)}\left[1-w_\theta(z_{1:T})\right]+\mathbb{E}_\theta^{(1)}\left[w_\theta(z_{1:T})\right]\right)}_{:=\zeta}. \qquad (4.12)$$

Since $E_{\bar{\mathcal{I}}(\theta)}$ is symmetric, this implies:

$$-2(T-1)\zeta/\underline{\sigma}^4(\theta)\cdot I_2\preccurlyeq E_{\bar{\mathcal{I}}(\theta)}\preccurlyeq 2(T-1)\zeta/\underline{\sigma}^4(\theta)\cdot I_2.$$

Therefore,

$$\bar{\mathcal{I}}(\theta)=\bar{\mathcal{I}}_s(\theta)+E_{\bar{\mathcal{I}}(\theta)}$$
$$\succcurlyeq\bar{\mathcal{I}}_s(\theta)-2(T-1)\zeta/\underline{\sigma}^4(\theta)\cdot I_2$$
$$=\bar{\mathcal{I}}_s(\theta)-2(T-1)\zeta/\underline{\sigma}^4(\theta)\cdot\bar{\mathcal{I}}_s^{1/2}(\theta)\bar{\mathcal{I}}_s^{-1}(\theta)\bar{\mathcal{I}}_s^{1/2}(\theta)$$

34

$$\geqslant \bar{\mathcal{I}}_s(\theta) - \frac{2(T-1)\zeta\lambda_{\max}(\bar{\mathcal{I}}_s^{-1}(\theta))}{\underline{\sigma}^4(\theta)} \cdot \bar{\mathcal{I}}_s(\theta)$$

$$= \left(1 - \frac{4(T-1)\zeta\bar{\sigma}^2(\theta)}{\underline{\sigma}^4(\theta)}\right)\bar{\mathcal{I}}_s(\theta).$$

A nearly identical argument shows that $\bar{\mathcal{I}}(\theta) \preccurlyeq \left(1 + \frac{4(T-1)\zeta\bar{\sigma}^2(\theta)}{\underline{\sigma}^4(\theta)}\right)\bar{\mathcal{I}}_s(\theta)$. Hence, if we choose $\zeta \leqslant \frac{\underline{\sigma}^4(\theta)}{8\bar{\sigma}^2(\theta)} \cdot \frac{1}{T-1}$, we will have the desired inequality:

$$\frac{1}{2}\bar{\mathcal{I}}_s(\theta) \preccurlyeq \bar{\mathcal{I}}(\theta) \preccurlyeq \frac{3}{2}\bar{\mathcal{I}}_s(\theta).$$

To conclude, we see that for any $\theta \in \Theta'$ and $i \in \{0,1\}$,

$$|\sigma_i^2(\theta) - \sigma_i^2(\theta_\star)| \leqslant |\theta_i - \theta_{\star,i}| \leqslant \|\theta - \theta_\star\| \leqslant \sigma_i^2(\theta_\star)/2 \implies \sigma_i^2(\theta_\star)/2 \leqslant \sigma_i^2(\theta) \leqslant 3\sigma_i^2(\theta)/2.$$

The RHS above implies that:

$$\underline{\sigma}^4(\theta) \geqslant \underline{\sigma}^4(\theta_\star)/4, \quad \bar{\sigma}^2(\theta) \leqslant 3\bar{\sigma}^2(\theta_\star). \tag{4.13}$$

Consequently, we have for $\theta \in \Theta'$:

$$\zeta \leqslant \frac{\underline{\sigma}^4(\theta_\star)}{48\bar{\sigma}^2(\theta_\star)} \cdot \frac{1}{T-1} \implies \zeta \leqslant \frac{\underline{\sigma}^4(\theta)}{8\bar{\sigma}^2(\theta)} \cdot \frac{1}{T-1}$$

We now apply Corollary 4.7 with $\eta = \frac{\underline{\sigma}^4(\theta_\star)}{48\bar{\sigma}^2(\theta_\star)}$ and $k = 1$, from which the result follows.

$\square$

Proposition 4.9. *Suppose that $T \geqslant 3$. The family of densities defined by (4.1) over $\Theta_+ := \{\theta \in \Theta \mid \theta_0 \geqslant \theta_1\}$ is $\left(\mathrm{Gap}^2(\theta_\star)/44, 13/\mathrm{Gap}(\theta_\star)\right)$-Hellinger identifiable (cf. Definition 3.11) around $\theta_\star \in \Theta_+$.*

Proof. We first generate a table of transition probabilities for any $\theta \in \Theta$ over the first three elements (z_1, z_2, z_3), leading to eight possibilities:

(z_1, z_2, z_3)	$p_\theta(z_1, z_2, z_3)$
$(1,1,1), (2,2,2)$	$\frac{1}{4}\left(\theta_0^2 + \theta_1^2\right)$
$(1,1,2), (1,2,2), (2,2,1), (2,1,1)$	$\frac{1}{4}\left(\theta_0(1-\theta_0) + \theta_1(1-\theta_1)\right)$
$(1,2,1), (2,1,2)$	$\frac{1}{4}\left((1-\theta_0)^2 + (1-\theta_1)^2\right)$

Table 1: A enumeration of the probabilities $p_\theta(z_{1:3})$ over the first three elements (z_1, z_2, z_3) in $z_{1:T}$.

We now use Table 1 to establish a lower bound for the TV distance $\|p_\theta(z_1, z_2, z_3) - p_{\theta_\star}(z_1, z_2, z_3)\|_{\mathrm{TV}}$ in terms of the parameters θ and $\theta_\star$:

$$\|p_\theta(z_{1:3}) - p_{\theta_\star}(z_{1:3})\|_{\mathrm{TV}} = \frac{1}{2}\sum_{(z_{1:3})\in\{1,2\}^3}|p_\theta(z_{1:3}) - p_{\theta_\star}(z_{1:3})|$$

$$= \frac{1}{4}|\theta_0^2 + \theta_1^2 - \theta_{\star,0}^2 - \theta_{\star,1}^2| + \frac{1}{4}|(1-\theta_0)^2 + (1-\theta_1)^2 - (1-\theta_{\star,0})^2 - (1-\theta_{\star,1})^2|$$

$$+ \frac{1}{2}|\theta_0(1-\theta_0) + \theta_1(1-\theta_1) - \theta_{\star,0}(1-\theta_{\star,0}) - \theta_{\star,1}(1-\theta_{\star,1})|$$

35

$$
\begin{aligned}
&= \frac{1}{4}\big|\|\theta\|^2 - \|\theta_\star\|^2\big| + \frac{1}{4}\big|\|\mathbb{1} - \theta\|^2 - \|\mathbb{1} - \theta_\star\|^2\big| + \frac{1}{2}\big|\langle\theta, \mathbb{1} - \theta\rangle - \langle\theta_\star, \mathbb{1} - \theta_\star\rangle\big| \\
&\geqslant \frac{1}{4}\big|\|\theta\|^2 - \|\theta_\star\|^2\big| + \frac{1}{4}\big|\|\mathbb{1} - \theta\|^2 - \|\mathbb{1} - \theta_\star\|^2\big| \\
&= \frac{1}{4}\big|\|\theta\|^2 - \|\theta_\star\|^2\big| + \frac{1}{4}\big|\|\theta\|^2 - \|\theta_\star\|^2 + 2(\|\theta_\star\|_1 - \|\theta\|_1)\big| \\
&\geqslant \max\left\{\frac{1}{4}\big|\|\theta\|^2 - \|\theta_\star\|^2\big|, \frac{1}{2}\big|\|\theta\|_1 - \|\theta_\star\|_1\big|\right\}.
\end{aligned}
$$

Hence by the data processing inequality, we have

$$
\mathrm{d}_H(p_\theta, p_{\theta_\star}) \geqslant \max\left\{\frac{1}{4}\big|\|\theta\|^2 - \|\theta_\star\|^2\big|, \frac{1}{2}\big|\|\theta\|_1 - \|\theta_\star\|_1\big|\right\}.
$$

Suppose we can control $\mathrm{d}_H(p_\theta, p_{\theta_\star}) \leqslant \varepsilon$, this would imply that:

$$
\big|\|\theta\|^2 - \|\theta_\star\|^2\big| \leqslant 4\varepsilon, \quad \big|\|\theta\|_1 - \|\theta_\star\|_1\big| \leqslant 2\varepsilon,
$$

i.e., denoting $a := \theta_0$, and $b := \theta_1$, we have:

$$
\begin{aligned}
S &:= a^2 + b^2 \in \big[\|\theta_\star\|^2 - 4\varepsilon, \|\theta_\star\|^2 + 4\varepsilon\big], &\text{(4.14a)} \\
Q &:= a + b \in \big[\|\theta_\star\|_1 - 2\varepsilon, \|\theta_\star\|_1 + 2\varepsilon\big]. &\text{(4.14b)}
\end{aligned}
$$

Our next step will be to solve for (a, b) in terms of (S, Q). To do this, we first will argue that the discriminant $2S - Q^2 > 0$. We define the following quantities:

$$
\Delta_S := S - \|\theta_\star\|^2, \quad \Delta_Q := Q - \|\theta_\star\|_1, \quad \rho_\star := \mathrm{Gap}(\theta_\star).
$$

With this notation,

$$
\begin{aligned}
2S - Q^2 &= 2(\|\theta_\star\|^2 + \Delta_S) - (\|\theta_\star\|_1^2 + \Delta_Q)^2 \\
&= 2\|\theta_\star\|^2 - \|\theta_\star\|_1^2 + 2\Delta_S - 2\Delta_Q\|\theta_\star\|_1 - \Delta_Q^2 \\
&= \rho_\star^2 + 2\Delta_S - 2\Delta_Q\|\theta_\star\|_1 - \Delta_Q^2 =: \rho_\star^2 + \Delta.
\end{aligned}
$$

Now observe that since $\varepsilon \leqslant \sqrt{2}$, we have $|\Delta| \leqslant 8\varepsilon + 8\varepsilon + 4\varepsilon^2 \leqslant 22\varepsilon$, and hence:

$$
\varepsilon \leqslant \rho_\star^2/44 \implies 2S - Q^2 \in [\rho_\star^2/2, 3\rho_\star^2/2].
$$

This shows that $2S - Q^2 > 0$, and therefore the following is the unique solution with $a > b$ to (4.14):

$$
a = \frac{Q + \sqrt{2S - Q^2}}{2}, \quad b = \frac{Q - \sqrt{2S - Q^2}}{2}.
$$

We next consider how $\sqrt{2S - Q^2}$ scales as a function of the perturbation Δ. Let us define $f(\Delta) := \sqrt{\rho_\star^2 + \Delta}$, which has derivative $f'(\Delta) = \frac{1}{2\sqrt{\rho_\star^2 + \Delta}}$. By concavity of square-root, we have:

$$
f(\Delta) \leqslant f(0) + f'(0)\Delta \leqslant \rho_\star + \frac{1}{2\rho_\star}|\Delta|.
$$

36

On the other hand, by the mean value theorem for some c with $|c| \leqslant |\Delta|$:

$$f(\Delta) = f(0) + f'(c)\Delta \geqslant \rho_\star - \frac{1}{\sqrt{2}\rho_\star}|\Delta|.$$

We can make this interval symmetric:

$$f(\Delta) \in \left[\rho_\star - \frac{1}{\sqrt{2}\rho_\star}|\Delta|, \rho_\star + \frac{1}{\sqrt{2}\rho_\star}|\Delta|\right].$$

Now we can conclude our analysis. We first observe that:

$$\|\theta_\star\|_1 + \rho_\star = 2\theta_{\star,0}, \quad \|\theta_\star\|_1 - \rho_\star = 2\theta_{\star,1}.$$

Next, we start with a. We have:

$$a = \frac{Q + \sqrt{2S - Q^2}}{2} = \frac{\|\theta_\star\|_1 + \Delta_Q + f(\Delta)}{2} \leqslant \frac{\|\theta_\star\|_1 + \rho_\star + \Delta_Q + |\Delta|/(\sqrt{2}\rho_\star)}{2}$$
$$\leqslant \frac{2\theta_{\star,0} + 2\varepsilon + 22\varepsilon/(\sqrt{2}\rho_\star)}{2} \leqslant \theta_{\star,0} + (1 + 11/\sqrt{2})\varepsilon/\rho_\star.$$

A similar argument shows that $a \geqslant \theta_{\star,0} - (1 + 5\sqrt{2})\varepsilon/\rho_\star$, and hence we have:

$$|a - \theta_{\star,0}| \leqslant (1 + 11/\sqrt{2})\varepsilon/\rho_\star.$$

Furthermore, a similar argument also shows that:

$$|b - \theta_{\star,1}| \leqslant (1 + 11/\sqrt{2})\varepsilon/\rho_\star.$$

Hence, we have $\|\theta - \theta_\star\| \leqslant \sqrt{2}\|\theta - \theta_\star\|_\infty \leqslant 13\varepsilon/\rho_\star$. Therefore, we have shown that for all $\theta \in \Theta'$:

$$\|\theta - \theta_\star\| \leqslant \frac{13}{\rho_\star}d_H(p_\theta, p_{\theta_\star}).$$

$\square$

Remark 4.10 (On identifiability up to permutation). We note that picking a uniform distribution for B illustrates one key issue in mixture models, where we can only identify parameters up to a permutation. Therefore we have to assume some additional distinguishability between parameters (e.g., the restricted subset Θ_+) to guarantee unique identifiability. We discuss this issue in more detail in Section 4.1.3.

4.1.2   Proof of Theorem 4.3

For compactness of notation, in the proof we will use

$$w_\theta^{(0)}(z_{1:T}) := w_\theta(z_{1:T}), \quad w_\theta^{(1)}(z_{1:T}) := 1 - w_\theta^{(0)}(z_{1:T}).$$

Covering Number Bound. We first compute an upper bound $\mathcal{I}_{\max}$ such that $\mathcal{I}(\theta) \preccurlyeq \mathcal{I}_{\max}$ for all $\theta \in \Theta$. Applying Corollary C.2 and recall the related computation in Proposition 4.8, specifically (4.11), we have:

$$\mathcal{I}(\theta) \preccurlyeq \begin{pmatrix} \mathbb{E}_\theta \left[\left| \partial_{\theta_0}^2 \log p_\theta(z_{1:T}) \right| + \left| \partial_{\theta_1} \partial_{\theta_0} \log p_\theta(z_{1:T}) \right| \right] & 0 \\ 0 & \mathbb{E}_\theta \left[\left| \partial_{\theta_1}^2 \log p_\theta(z_{1:T}) \right| + \left| \partial_{\theta_1} \partial_{\theta_0} \log p_\theta(z_{1:T}) \right| \right] \end{pmatrix}$$

$$\preccurlyeq \frac{5}{4} \frac{(T-1)^2}{\underline{\sigma}^4(\theta)} I_2.$$

Hence, we have for all $\theta \in \Theta$:

$$\mathcal{I}(\theta) \preccurlyeq \mathcal{I}_{\max} := \frac{5(T-1)^2}{4\sigma_{\min}^4} I_2.$$

Consequently, for any $\theta, \theta' \in \Theta$,

$$\mathrm{d}_{\mathcal{I}_{\max}} (p_\theta, p_{\theta'}) = \left\| \theta - \theta' \right\|_{\mathcal{I}_{\max}} = \sqrt{\frac{5}{4}} \frac{T-1}{\sigma_{\min}^2} \left\| \theta - \theta' \right\|.$$

Therefore we can upper bound the metric entropy

$$\log \mathcal{N}_{\mathcal{I}_{\max}} (\mathcal{P}, \varepsilon) \leqslant \log \mathcal{N}_{\|\cdot\|} \left([\mu, 1-\mu]^2, \sqrt{\frac{4}{5}} \frac{\sigma_{\min}^2 \varepsilon}{T-1} \right) \leqslant 2 \log \left(\sqrt{\frac{5}{4}} \frac{T-1}{\sigma_{\min}^2 \varepsilon} \right) + \log 2.$$

We now first apply Theorem 3.6 (a) with resolution $\varepsilon = \delta/(2\sqrt{2m})$, which yields an event $\mathcal{E}_1$ with probability at least $1-\delta$, where on $\mathcal{E}_1$:

$$\mathrm{d}_H^2(\hat{\theta}_{m,T}^\varepsilon, \theta_\star) \leqslant \frac{8 \log \left(\sqrt{10} \frac{\sqrt{m}(T-1)}{\sigma_{\min}^2 \delta} \right) + 8 \log 2 + 4 \log(1/\delta) + \delta^2/4}{m} \leqslant \frac{21 \log \left(\frac{4mT}{\sigma_{\min}^2 \delta} \right)}{m}, \tag{4.15}$$

where the last inequality follows from by assumption, $m \geqslant 4$, and the observation

$$\log \left(\sqrt{10} \frac{\sqrt{m}(T-1)}{\sigma_{\min}^2 \delta} \right) \geqslant \log \left(8\sqrt{10} \frac{1}{\delta} \right) \geqslant \max \left\{ \log 2, \log(1/\delta), \delta^2 \right\}, \quad \forall \, \delta \in (0, 1).$$

For the remainder of the proof, we will assume we are on the event $\mathcal{E}_1$. Invoking the Hellinger identifiability (Proposition 4.9) and Proposition A.1:

$$m \geqslant \frac{42}{\gamma_1^2} \log \left(\frac{168T}{\gamma_1^2 \sigma_{\min}^2 \delta} \right) \implies m \geqslant \frac{21}{\gamma_1^2} \log \left(\frac{4mT}{\sigma_{\min}^2 \delta} \right)$$

$$\implies \mathrm{d}_H(\hat{\theta}_{m,T}^\varepsilon, \theta_\star) \leqslant \gamma_1$$

$$\implies \left\| \hat{\theta}_{m,T}^\varepsilon - \theta_\star \right\| \leqslant \gamma_2 \mathrm{d}_H(\hat{\theta}_{m,T}^\varepsilon, \theta_\star) \leqslant 5\gamma_2 \sqrt{\frac{\log \left(\frac{4mT}{\sigma_{\min}^2 \delta} \right)}{m}}.$$

We can now additionally impose (cf. Proposition A.1)

$$m \geqslant \frac{50\gamma_2^2}{c^2} \log \left(\frac{200\gamma_2^2 T}{c^2 \sigma_{\min}^2 \delta} \right) \implies m \geqslant \frac{25\gamma_2^2 \log \left(\frac{4mT}{\sigma_{\min}^2 \delta} \right)}{c^2},$$

where c is some constant radius. This results in the following key identity, which will be used in the sequel:

$$m \geqslant \max\left\{\frac{42}{\gamma_1^2}\log\left(\frac{168T}{\gamma_1^2\sigma_{\min}^2\delta}\right),\frac{50\gamma_2^2}{c^2}\log\left(\frac{200\gamma_2^2 T}{c^2\sigma_{\min}^2\delta}\right)\right\} \implies \|\hat\theta_{m,T}^\varepsilon - \theta\| \leqslant c. \tag{4.16}$$

Estimate B_1 and B_2. We now estimate $B_1(\hat\theta_{m,T}^\varepsilon,\theta_\star)$ and $B_2(\hat\theta_{m,T}^\varepsilon,\theta_\star)$. Our first step is to guarantee that $\hat\theta_{m,T}^\varepsilon \in \Theta'$ (cf. (4.5)), i.e.,

$$\|\hat\theta_{m,T}^\varepsilon - \theta_\star\| \leqslant \min\left\{\frac{\rho_\star}{2\sqrt{2}},\frac{\underline\sigma^2(\theta_\star)}{2}\right\}.$$

In view of (4.16), it suffices to require

$$m \geqslant \max\left\{\frac{42}{\gamma_1^2}\log\left(\frac{168T}{\gamma_1^2\sigma_{\min}^2\delta}\right),\frac{50\gamma_2^2}{\min\left\{\frac{\rho_\star}{2\sqrt{2}},\frac{\underline\sigma^2(\theta_\star)}{2}\right\}^2}\log\left(\frac{200\gamma_2^2 T}{\min\left\{\frac{\rho_\star}{2\sqrt{2}},\frac{\underline\sigma^2(\theta_\star)}{2}\right\}^2\sigma_{\min}^2\delta}\right)\right\}. \tag{4.17}$$

We abbreviate the above to be written as:

$$m \gtrsim \max\left\{\frac{1}{\rho_\star^4},\frac{1}{\rho_\star^2\underline\sigma^4(\theta_\star)}\right\}\log\left(\max\left\{\frac{1}{\rho_\star^4},\frac{1}{\rho_\star^2\underline\sigma^4(\theta_\star)}\right\}\frac{T}{\sigma_{\min}^2\delta}\right). \tag{4.18}$$

In addition to (4.18), further requiring

$$T \gtrsim \max\left\{\frac{\bar\sigma^2(\theta_\star)}{\rho_\star^4},\frac{1}{\rho_\star^2}\right\}\log^2(1/\mu)\log\left(\frac{\bar\sigma^2(\theta_\star)}{\underline\sigma^4(\theta_\star)}\cdot\frac{\log(1/\mu)}{\rho_\star}\right), \tag{4.19}$$

implies by Proposition 4.8 that for any $\theta \in \operatorname{conv}\{\hat\theta_{m,T}^\varepsilon,\theta_\star\}$:

$$\frac{1}{4}\mathcal{I}_{\mathsf{diag}}(\theta) \preccurlyeq \mathcal{I}(\theta) \preccurlyeq \frac{3}{4}\mathcal{I}_{\mathsf{diag}}(\theta),\quad \mathcal{I}_{\mathsf{diag}}(\theta) := \mathsf{diag}\left\{\mathcal{I}(\theta_0),\mathcal{I}(\theta_1)\right\} = \mathsf{diag}\left\{\frac{T-1}{\sigma_0^2(\theta)},\frac{T-1}{\sigma_1^2(\theta)}\right\}. \tag{4.20}$$

And therefore applying Proposition C.3, we have for any $\theta \in \operatorname{conv}\{\hat\theta_{m,T}^\varepsilon,\theta_\star\}$:

$$\sup_{\|v\|=1}\left\|\left\langle v\mathcal{I}(\theta)^{-1/2}\nabla_\theta\log p_\theta(z_{1:T})\right\rangle\right\|_{\mathcal{L}^4(p_\theta)}$$

$$\leqslant \sup_{\|v\|=1}\left\|\left\langle v\left(\tfrac{1}{4}\right)^{-1/2}\mathsf{diag}\left\{\mathcal{I}(\theta_0)^{-1/2},\mathcal{I}(\theta_1)^{-1/2}\right\}\nabla_\theta\log p_\theta(z_{1:T})\right\rangle\right\|_{\mathcal{L}^4(p_\theta)}$$

$$= 2\sup_{\|v\|=1}\left\|\sum_{i=0,1}\left(v_i\sqrt{\tfrac{\theta_i(1-\theta_i)}{T-1}}w_\theta^{(i)}(z_{1:T})\partial_{\theta_i}\log p_{\theta_i}(z_{1:T})\right)\right\|_{\mathcal{L}^4(p_\theta)}$$

$$\lesssim \underbrace{\left\|\sqrt{\tfrac{\theta_0(1-\theta_0)}{T-1}}w_\theta^{(0)}(z_{1:T})\partial_{\theta_0}\log p_{\theta_0}(z_{1:T})\right\|_{\mathcal{L}^4(p_\theta)}}_{:=(I)} + \underbrace{\left\|\sqrt{\tfrac{\theta_1(1-\theta_1)}{T-1}}w_\theta^{(1)}(z_{1:T})\partial_{\theta_1}\log p_{\theta_1}(z_{1:T})\right\|_{\mathcal{L}^4(p_\theta)}}_{:=(II)}.$$

39

Notice each term *without* the density ratio $w_\theta^{(i)}(z_{1:T})$ resembles that what we have seen in Section 3.3. Hence we upper bound (I) as:

$$
\begin{aligned}
(I) &= \left(\frac{\theta_0^2(1-\theta_0)^2}{(T-1)^2} \mathbb{E}_\theta \left[\left(w_\theta^{(0)}(z_{1:T}) \partial_{\theta_0} \log p_{\theta_0}(z_{1:T}) \right)^4 \right] \right)^{1/4} \\
&\leqslant \left(\frac{\theta_0^2(1-\theta_0)^2}{(T-1)^2} \mathbb{E}_\theta^{(0)} \left[(\partial_{\theta_0} \log p_{\theta_0}(z_{1:T}))^4 \right] + \frac{\theta_0^2(1-\theta_0)^2}{(T-1)^2} \mathbb{E}_\theta^{(1)} \left[\left(w_\theta^{(0)}(z_{1:T}) \partial_{\theta_0} \log p_{\theta_0}(z_{1:T}) \right)^4 \right] \right)^{1/4} \\
&\lesssim \left(\frac{1}{T\underline{\sigma}^2(\theta)} + O(1) + \frac{(T-1)^2}{\underline{\sigma}^4(\theta)} \mathbb{E}_\theta^{(1)} \left[\left| w_\theta^{(0)}(z_{1:T}) \right| \right] \right)^{1/4} \\
&\lesssim \left(\frac{1}{T\underline{\sigma}^2(\theta)} + O(1) \right)^{1/4} \lesssim \left(\frac{1}{T\underline{\sigma}^2(\theta_\star)} + O(1) \right)^{1/4},
\end{aligned}
$$
$$(4.21)$$

where the third to last step follows from (3.30), and the last two steps we first set a large enough T such that the last term is also $O(1)$, for which our assumption (4.19) would suffice together with (4.13). A similar argument shows that

$$
(II) \lesssim \left(\frac{1}{T\underline{\sigma}^2(\theta_\star)} + O(1) \right)^{1/4},
$$

and therefore we can conclude

$$
B_1^4 \lesssim \max \left\{ \frac{1}{T\underline{\sigma}^2(\theta_\star)}, 1 \right\}.
$$

Now for B_2, we proceed as we did for B_1. For any $\theta \in \mathrm{conv}\left\{ \hat{\theta}_{m,T}^\varepsilon, \theta_\star \right\}$:

$$
\begin{aligned}
&\sup_{\|v\|=1} \left\| \left\langle v \mathcal{I}(\theta)^{-1/2} \nabla_\theta^2 \log p_\theta(z_{1:T}) \mathcal{I}(\theta)^{-1/2} v \right\rangle \right\|_{\mathcal{L}^2(p_\theta)} \\
&\leqslant \left(\tfrac{1}{4} \right)^{-1} \Bigg(\sup_{\|v\|=1} \left\| v_0^2 \tfrac{\theta_0(1-\theta_0)}{(T-1)} w_\theta^{(0)}(z_{1:T}) w_\theta^{(1)}(z_{1:T}) \left(\partial_{\theta_0} \log p_{\theta_0}(z_{1:T}) \right)^2 \right\|_{\mathcal{L}^2(p_\theta)} \\
&\quad + \sup_{\|v\|=1} \left\| v_0^2 \tfrac{\theta_0(1-\theta_0)}{(T-1)} w_\theta^{(0)}(z_{1:T}) \partial_{\theta_0}^2 \log p_{\theta_0}(z_{1:T}) \right\|_{\mathcal{L}^2(p_\theta)} \\
&\quad + \sup_{\|v\|=1} \left\| v_1^2 \tfrac{\theta_1(1-\theta_1)}{(T-1)} w_\theta^{(0)}(z_{1:T}) w_\theta^{(1)}(z_{1:T}) \left(\partial_{\theta_1} \log p_{\theta_1}(z_{1:T}) \right)^2 \right\|_{\mathcal{L}^2(p_\theta)} \\
&\quad + \sup_{\|v\|=1} \left\| v_1^2 \tfrac{\theta_1(1-\theta_1)}{(T-1)} w_\theta^{(1)}(z_{1:T}) \partial_{\theta_1}^2 \log p_{\theta_1}(z_{1:T}) \right\|_{\mathcal{L}^2(p_\theta)} \\
&\quad + \sup_{\|v\|=1} \left\| v_0 v_1 \left(\sum_{i=0,1} \tfrac{\theta_i(1-\theta_i)}{(T-1)} \right) w_\theta^{(0)}(z_{1:T}) w_\theta^{(1)}(z_{1:T}) \partial_{\theta_0} \log p_{\theta_0}(z_{1:T}) \partial_{\theta_1} \log p_{\theta_1}(z_{1:T}) \right\|_{\mathcal{L}^2(p_\theta)} \Bigg) \\
&\lesssim \underbrace{\left\| w_\theta^{(0)}(z_{1:T}) \left(\sqrt{\tfrac{\theta_0(1-\theta_0)}{(T-1)}} \partial_{\theta_0} \log p_{\theta_0}(z_{1:T}) \right)^2 \right\|_{\mathcal{L}^2(p_\theta)}}_{:=(III)} + \underbrace{\left\| \tfrac{\theta_0(1-\theta_0)}{(T-1)} w_\theta^{(0)}(z_{1:T}) \partial_{\theta_0}^2 \log p_{\theta_0}(z_{1:T}) \right\|_{\mathcal{L}^2(p_\theta)}}_{:=(IV)}
\end{aligned}
$$

$$
+ \left\| w_\theta^{(1)}(z_{1:T}) \left(\sqrt{\tfrac{\theta_1(1-\theta_1)}{(T-1)}} \partial_{\theta_1} \log p_{\theta_1}(z_{1:T}) \right)^2 \right\|_{\mathcal{L}^2(p_\theta)} + \underbrace{\left\| \tfrac{\theta_1(1-\theta_1)}{(T-1)} w_\theta^{(1)}(z_{1:T}) \partial_{\theta_1}^2 \log p_{\theta_1}(z_{1:T}) \right\|_{\mathcal{L}^2(p_\theta)}}_{:=(VI)}
$$

$$
\underbrace{}_{:=(V)}
$$

$$
+ \underbrace{\tfrac{\bar{\sigma}^2(\theta)}{T-1} \left\| w_\theta^{(0)}(z_{1:T}) w_\theta^{(1)}(z_{1:T}) \partial_{\theta_0} \log p_{\theta_0}(z_{1:T}) \partial_{\theta_1} \log p_{\theta_1}(z_{1:T}) \right\|_{\mathcal{L}^2(p_\theta)}}_{:=(VII)}.
$$

Now using similar arguments for the upper bound of (I) in (4.21), we can upper bound (III):

$$
(III) = \left(\frac{\theta_0^2(1-\theta_0)^2}{(T-1)^2} \mathbb{E}_\theta \left[\left(\sqrt{w_\theta^{(0)}(z_{1:T})} \partial_{\theta_0} \log p_{\theta_0}(z_{1:T}) \right)^4 \right] \right)^{1/2}
$$

$$
\leqslant \left(\frac{\theta_0^2(1-\theta_0)^2}{(T-1)^2} \mathbb{E}_\theta^{(0)} \left[(\partial_{\theta_0} \log p_{\theta_0}(z_{1:T}))^4 \right] + \frac{\theta_0^2(1-\theta_0)^2}{(T-1)^2} \mathbb{E}_\theta^{(1)} \left[\left(\sqrt{w_\theta^{(0)}(z_{1:T})} \partial_{\theta_0} \log p_{\theta_0}(z_{1:T}) \right)^4 \right] \right)^{1/2}
$$

$$
\lesssim \left(\frac{1}{T\underline{\sigma}^2(\theta)} + O(1) + \frac{(T-1)^2}{\underline{\sigma}^4(\theta)} \mathbb{E}_\theta^{(1)} \left[\left| w_\theta^{(0)}(z_{1:T}) \right| \right] \right)^{1/2}
$$

$$
\lesssim \left(\frac{1}{T\underline{\sigma}^2(\theta_\star)} + O(1) \right)^{1/2},
$$

and (IV):

$$
(IV) = \left(\frac{\theta_0^2(1-\theta_0)^2}{(T-1)^2} \mathbb{E}_\theta \left[\left(w_\theta^{(0)}(z_{1:T}) \partial_{\theta_0}^2 \log p_{\theta_0}(z_{1:T}) \right)^2 \right] \right)^{1/2}
$$

$$
\leqslant \left(\frac{\theta_0^2(1-\theta_0)^2}{(T-1)^2} \mathbb{E}_\theta^{(0)} \left[(\partial_{\theta_0}^2 \log p_{\theta_0}(z_{1:T}))^2 \right] + \frac{\theta_0^2(1-\theta_0)^2}{(T-1)^2} \mathbb{E}_\theta^{(1)} \left[\left(w_\theta^{(0)}(z_{1:T}) \partial_{\theta_0}^2 \log p_{\theta_0}(z_{1:T}) \right)^2 \right] \right)^{1/2}
$$

$$
\lesssim \left(\frac{1}{T\underline{\sigma}^2(\theta)} + O(1) + \frac{1}{\underline{\sigma}^4(\theta)} \mathbb{E}_\theta^{(1)} \left[\left| w_\theta^{(0)}(z_{1:T}) \right| \right] \right)^{1/2}
$$

$$
\lesssim \left(\frac{1}{T\underline{\sigma}^2(\theta_\star)} + O(1) \right)^{1/2}.
$$

By symmetry between θ_0 and θ_1, we have

$$
(V) \lesssim \left(\frac{1}{T\underline{\sigma}^2(\theta_\star)} + O(1) \right)^{1/2}, \quad (VI) \lesssim \left(\frac{1}{T\underline{\sigma}^2(\theta_\star)} + O(1) \right)^{1/2}.
$$

Finally, we can upperbound (VII):

$$
(VII) \leqslant \frac{\bar{\sigma}^2(\theta)(T-1)}{\underline{\sigma}^4(\theta)} \left(\mathbb{E}_\theta^{(0)} \left[\left(w_\theta^{(1)}(z_{1:T}) \right)^2 \right] + \mathbb{E}_\theta^{(1)} \left[\left(w_\theta^{(0)}(z_{1:T}) \right)^2 \right] \right)^{1/2}
$$

$$
\lesssim \frac{\bar{\sigma}^2(\theta_\star)(T-1)}{\underline{\sigma}^4(\theta_\star)} \left(\mathbb{E}_\theta^{(0)} \left[\left| w_\theta^{(1)}(z_{1:T}) \right| \right] + \mathbb{E}_\theta^{(1)} \left[\left| w_\theta^{(0)}(z_{1:T}) \right| \right] \right)^{1/2}
$$

$$
\lesssim O(1),
$$

where at the second step we again applied (4.13). Therefore $B_2^2 \lesssim \max\left\{\frac{1}{T\underline{\sigma}^2(\theta_\star)}, 1\right\}$, and hence:

$$\max\{B_1^4, B_2^2\} \lesssim \max\left\{\frac{1}{T\underline{\sigma}^2(\theta_\star)}, 1\right\}. \tag{4.22}$$

Parameter error bound. Our first step is to define a more refined $\mathcal{I}'_{\max}$ for $\theta \in \Theta'$. From Proposition 4.8, for any $\theta \in \Theta'$:

$$\lambda_{\max}(\mathcal{I}(\theta)) \leqslant \frac{T}{\underline{\sigma}^2(\theta)} \leqslant \frac{4T}{\underline{\sigma}^2(\theta_\star)}.$$

Therefore, we have:

$$\theta \in \Theta' \Longrightarrow \mathcal{I}(\theta) \preccurlyeq \frac{4T}{\underline{\sigma}^2(\theta_\star)} I_2 =: \mathcal{I}'_{\max}.$$

Hence, following (3.40) using $\mathcal{I}'_{\max}$ instead of $\mathcal{I}_{\max}$, we obtain that:

$$d_H(\hat\theta^\varepsilon_{m,T}, \theta_\star) \leqslant \gamma_1 \Longrightarrow \sup_{\theta \in \mathrm{conv}\{\hat\theta^\varepsilon_{m,T}, \theta_\star\}} d_H^2(\theta, \theta_\star) \leqslant \frac{\gamma_2^2 T}{\underline{\sigma}^2(\theta_\star)} d_H^2(\hat\theta^\varepsilon_{m,T}, \theta_\star) \lesssim \frac{\gamma_2^2 T \log(mT/(\sigma_{\min}^2 \delta))}{\underline{\sigma}^2(\theta_\star) m},$$

where recall that the last inequality is from (4.15). Therefore, combining the above inequality with (4.22) and Proposition A.1, condition (3.22) holds if in addition to (4.18) and (4.19) we also have:

$$m \gtrsim \max\left\{\frac{1}{\rho_\star^2 \underline{\sigma}^4(\theta_\star)}, \frac{T}{\rho_\star^2 \underline{\sigma}^2(\theta_\star)}\right\} \log\left(\max\left\{\frac{1}{\rho_\star^2 \underline{\sigma}^4(\theta_\star)}, \frac{T}{\rho_\star^2 \underline{\sigma}^2(\theta_\star)}\right\} \frac{T}{\sigma_{\min}^2 \delta}\right). \tag{4.23}$$

Applying Proposition 3.9 (a), we obtain from (3.23):

$$\|\hat\theta^\varepsilon_{m,T} - \theta_\star\|_{\mathcal{I}_2(\theta_\star, \hat\theta^\varepsilon_{m,T})}^2 \lesssim m^{-1} \log\left(\frac{mT}{\sigma_{\min}^2 \delta}\right).$$

Furthermore, from (4.20), we have that $\mathcal{I}_2(\theta_\star, \hat\theta^\varepsilon_{m,T}) \succcurlyeq c_0 T \cdot I_2$ for some $c_0 > 0$, and hence the following parameter error bound holds as well:

$$\|\hat\theta^\varepsilon_{m,T} - \theta_\star\|^2 \lesssim \frac{1}{mT} \log\left(\frac{mT}{\sigma_{\min}^2 \delta}\right). \tag{4.24}$$

Verify FI radius. For the variance-weighted CLT rate, we want to show the FI radius condition (3.24). By combining Proposition C.4 and (4.20), we have for any $\theta \in \Theta'$,

$$\|\mathcal{I}(\theta_\star)^{-1/2} \mathcal{I}(\theta) \mathcal{I}(\theta_\star)^{-1/2} - I\|_{\mathrm{op}} = \|\mathcal{I}(\theta_\star)^{-1/2}\left(\mathcal{I}(\theta) - \mathcal{I}(\theta_\star)\right) \mathcal{I}(\theta_\star)^{-1/2}\|_{\mathrm{op}}$$
$$\leqslant 4\|\mathcal{I}_{\mathsf{diag}}(\theta_\star)^{-1/2}\left(\mathcal{I}(\theta) - \mathcal{I}(\theta_\star)\right) \mathcal{I}_{\mathsf{diag}}(\theta_\star)^{-1/2}\|_{\mathrm{op}}.$$

Hence condition (3.24) is implied by:

$$\|\mathcal{I}_{\mathsf{diag}}(\theta_\star)^{-1/2}\left(\mathcal{I}(\theta) - \mathcal{I}(\theta_\star)\right) \mathcal{I}_{\mathsf{diag}}(\theta_\star)^{-1/2}\|_{\mathrm{op}} \leqslant 1/8. \tag{4.25}$$

Now recall from the error decomposition in (4.9), we have

$$\mathcal{I}(\theta) = \frac{1}{2}\mathcal{I}_{\mathsf{diag}}(\theta) + \frac{1}{2}\left(\mathbb{E}_{\theta}^{(0)}\left[E_0(\theta)\right] + \mathbb{E}_{\theta}^{(1)}\left[E_1(\theta)\right]\right),$$

$$\mathcal{I}(\theta_\star) = \frac{1}{2}\mathcal{I}_{\mathsf{diag}}(\theta_\star) + \frac{1}{2}\left(\mathbb{E}_{\theta_\star}^{(0)}\left[E_0(\theta_\star)\right] + \mathbb{E}_{\theta_\star}^{(1)}\left[E_1(\theta_\star)\right]\right).$$

Hence denoting

$$E(\theta, \theta_\star) := \left(\mathcal{I}(\theta) - \mathcal{I}(\theta_\star)\right) - \left(\frac{1}{2}\mathcal{I}_{\mathsf{diag}}(\theta) - \frac{1}{2}\mathcal{I}_{\mathsf{diag}}(\theta_\star)\right)$$

$$= \frac{1}{2}\left(\mathbb{E}_{\theta}^{(0)}\left[E_0(\theta)\right] + \mathbb{E}_{\theta}^{(1)}\left[E_1(\theta)\right]\right) - \frac{1}{2}\left(\mathbb{E}_{\theta_\star}^{(0)}\left[E_0(\theta_\star)\right] + \mathbb{E}_{\theta_\star}^{(1)}\left[E_1(\theta_\star)\right]\right),$$

we have (4.25) is equivalent to

$$\left\|\mathcal{I}_{\mathsf{diag}}(\theta_\star)^{-1/2}\left(\frac{1}{2}\mathcal{I}_{\mathsf{diag}}(\theta) - \frac{1}{2}\mathcal{I}_{\mathsf{diag}}(\theta_\star) + E(\theta, \theta_\star)\right)\mathcal{I}_{\mathsf{diag}}(\theta_\star)^{-1/2}\right\|_{\mathsf{op}} \leqslant 1/8.$$

By triangle inequality, a further sufficient condition is:

$$\left\|\mathsf{diag}\left\{\frac{\mathcal{I}(\theta_0)}{\mathcal{I}(\theta_{\star,0})} - 1, \frac{\mathcal{I}(\theta_1)}{\mathcal{I}(\theta_{\star,1})} - 1\right\}\right\|_{\mathsf{op}} = \max\left\{\left|\frac{\mathcal{I}(\theta_0)}{\mathcal{I}(\theta_{\star,0})} - 1\right|, \left|\frac{\mathcal{I}(\theta_1)}{\mathcal{I}(\theta_{\star,1})} - 1\right|\right\} \leqslant \frac{1}{8}, \qquad (4.26)$$

$$\left\|\mathcal{I}_{\mathsf{diag}}^{-1/2}(\theta_\star)E(\theta, \theta_\star)\mathcal{I}_{\mathsf{diag}}^{-1/2}(\theta_\star)\right\|_{\mathsf{op}} \leqslant \frac{1}{16}. \qquad (4.27)$$

From Section 3.3, specifically, (3.34), we have that (4.26) can be satisfied by requiring

$$\frac{\sqrt{2}}{\underline{\sigma}^2(\theta_\star)}\|\theta - \theta_\star\| \leqslant \frac{2}{\underline{\sigma}^2(\theta_\star)}\max\left\{|\theta_0 - \theta_{0,\star}|, |\theta_1 - \theta_{1,\star}|\right\} \leqslant \frac{1}{8}.$$

So it suffices to require

$$\frac{\|\hat{\theta}_{m,T}^\varepsilon - \theta_\star\|^2}{\underline{\sigma}^4(\theta_\star)} \leqslant \frac{1}{128},$$

and in view of (4.24) and Proposition A.1, this holds if:

$$mT \gtrsim \frac{1}{\underline{\sigma}^4(\theta_\star)}\log\left(\frac{1}{\underline{\sigma}^4(\theta_\star)\sigma_{\min}^2\delta}\right).$$

This condition is however already implied by (4.23) (up to adjusting constant factors). We now address condition (4.27). Recalling $\mathcal{I}_{\mathsf{diag}}(\theta_\star) \succcurlyeq (T-1)\cdot I_2$, we have:

$$\left\|\mathcal{I}_{\mathsf{diag}}^{-1/2}(\theta_\star)E(\theta, \theta_\star)\mathcal{I}_{\mathsf{diag}}^{-1/2}(\theta_\star)\right\|_{\mathsf{op}} \overset{(a)}{\leqslant} \frac{1}{T-1}\|E(\theta, \theta_\star)\|_{\mathsf{op}}$$

$$\overset{(b)}{\lesssim} \frac{T-1}{\underline{\sigma}^4(\theta)}\left(\mathbb{E}_{\theta}^{(0)}\left[w_\theta^{(1)}(z_{1:T})\right] + \mathbb{E}_{\theta}^{(1)}\left[w_\theta^{(0)}(z_{1:T})\right]\right) + \frac{T-1}{\underline{\sigma}^4(\theta_\star)}\left(\mathbb{E}_{\theta_\star}^{(0)}\left[w_{\theta_\star}^{(1)}(z_{1:T})\right] + \mathbb{E}_{\theta_\star}^{(1)}\left[w_{\theta_\star}^{(0)}(z_{1:T})\right]\right)$$

$$\overset{(c)}{\lesssim} \frac{T-1}{\underline{\sigma}^4(\theta_\star)}\left(\mathbb{E}_{\theta}^{(0)}\left[w_\theta^{(1)}(z_{1:T})\right] + \mathbb{E}_{\theta}^{(1)}\left[w_\theta^{(0)}(z_{1:T})\right] + \mathbb{E}_{\theta_\star}^{(0)}\left[w_{\theta_\star}^{(1)}(z_{1:T})\right] + \mathbb{E}_{\theta_\star}^{(1)}\left[w_{\theta_\star}^{(0)}(z_{1:T})\right]\right) \overset{(d)}{\lesssim} O(1),$$

where (a) uses Proposition C.4, (b) uses (4.12) from the proof of Proposition 4.8, and (c) uses (4.13), and (d) uses the requirement on T from (4.19) (possibly adjusting constant factors as necessary), and follows the arguments bounding ζ in (4.12) from Proposition 4.8.

43

Final result. We now have the necessary requirements to invoke Proposition 3.9 (b) to conclude:

$$\|\hat{\theta}^{\varepsilon}_{m,T} - \theta_\star\|^2_{\mathcal{I}(\theta_\star)} \lesssim \frac{1}{m} \log\left(\frac{mT}{\sigma^2_{\min}\delta}\right).$$

In particular, combined with Proposition 4.8, this also implies

$$\|\hat{\theta}^{\varepsilon}_{m,T} - \theta_\star\|^2 \lesssim \frac{\bar{\sigma}^2(\theta_\star)}{mT} \log\left(\frac{mT}{\sigma^2_{\min}\delta}\right).$$

We conclude by summarizing the requirements on m, T. In total, we require conditions (4.18), (4.19), and (4.23) to hold. These are readily simplified into assumptions (b) and (c) in the theorem statement, from which the result follows.

4.1.3 Extensions of Proof Techniques to More General Mixture Problems

We now discuss the extent to which the proof strategy for Section 4.1 extends to general mixture distributions. We consider a density class $\mathcal{P} := \left\{p_\theta \mid \theta \in \Theta^k\right\}$ over k parameters $\{\theta_1, \ldots, \theta_k\} \subset \Theta$ where for all $i \neq j$, $\theta_i \neq \theta_j$, and a multinomial distribution on the index set $[k]$ parameterized by $f := (f_1, \ldots, f_{k-1}) \in \Delta^{k-1}_\mu$, where $f_k := 1 - \sum_{i=1}^{k-1} f_i$ and

$$\Delta^{k-1}_\mu := \left\{f \in \mathbb{R}^{k-1} \,\Big|\, \sum_{i=1}^{k-1} f_i \leqslant 1 - \mu, \ f_i \geqslant \mu \ \forall i \in [k-1]\right\}$$

denotes the μ-strict interior of k-dimensional probability simplex, for $0 < \mu \leqslant 1/k$. We require the weights f_k to be in the strict interior for regularity (e.g., differentiability and exchanging derivatives with integrals) reasons. The data-generating process we consider proceeds similarly to what we considered in Section 4.1. Specifically, a latent variable $B \in [k]$ is first drawn according to $\{f_i\}_{i=1}^k$, which is then used condition the data $z_{1:T} \sim p_{\theta_B}$.

Mixture with known weights. We first suppose f is known. Hence the densities $p_\theta \in \mathcal{P}$ are:

$$p_\theta(z_{1:T}) = \sum_{i=1}^k f_i p_{\theta_i}(z_{1:T}),$$

where

$$\theta := \begin{pmatrix}\theta_1^\mathsf{T} & \ldots & \theta_k^\mathsf{T}\end{pmatrix}^\mathsf{T} = \begin{pmatrix}\theta_{1,1} & \ldots & \theta_{1,d} & \ldots & \theta_{k,1} & \ldots & \theta_{k,d}\end{pmatrix}^\mathsf{T} \in \Theta^k$$

is the joint parameter to be learned. A simple computation of the first order information yields:

$$\nabla_\theta \log p_\theta(z_{1:T}) = \left(w^{(1)}_\theta(z_{1:T})\nabla_{\theta_1} \log p_{\theta_1}(z_{1:T})^\mathsf{T} \ \ \ldots \ \ w^{(k)}_\theta(z_{1:T})\nabla_{\theta_k} \log p_{\theta_k}(z_{1:T})^\mathsf{T}\right)^\mathsf{T}, \qquad (4.28)$$

where $w^{(B)}_\theta(z_{1:T})$ is the posterior density of B given $z_{1:T}$:

$$w^{(i)}_\theta(z_{1:T}) := p_\theta(B = i \mid z_{1:T}) = \frac{p_\theta(B = i)p_\theta(z_{1:T} \mid B = i)}{\sum_{i=1}^k p_\theta(B = i, z_{1:T})} = \frac{f_i p_{\theta_i}(z_{1:T})}{\sum_{i=1}^k f_i p_{\theta_i}(z_{1:T})}.$$

44

We now compute the second order information in blocks: for $1 \leqslant i \neq j \leqslant k$,

$$
\begin{aligned}
H_\theta^{(ii)}(z_{1:T}) &:= \nabla_{\theta_i}^2 \log p_\theta(z_{1:T}) \\
&= w_\theta^{(i)}(z_{1:T})\left(1 - w_\theta^{(i)}(z_{1:T})\right)\left(\nabla_{\theta_i} \log p_{\theta_i}(z_{1:T})\right)^{\otimes 2} + w_\theta^{(i)}(z_{1:T})\nabla_{\theta_i}^2 \log p_{\theta_i}(z_{1:T}), \quad (4.29)
\end{aligned}
$$

$$
H_\theta^{(ij)}(z_{1:T}) := \nabla_{\theta_j}\nabla_{\theta_i} \log p_\theta(z_{1:T}) = w_\theta^{(i)}(z_{1:T})w_\theta^{(j)}(z_{1:T})\nabla_{\theta_i} \log p_{\theta_i}(z_{1:T})\nabla_{\theta_j} \log p_{\theta_j}(z_{1:T})^{\mathsf{T}},
$$

and the Hessian matrix $\nabla_\theta^2 \log p_\theta(z_{1:T})$ is

$$
H_\theta(z_{1:T}) := \begin{pmatrix}
H_\theta^{(11)}(z_{1:T}) & H_\theta^{(12)}(z_{1:T}) & \cdots & H_\theta^{(1k)}(z_{1:T}) \\
H_\theta^{(21)}(z_{1:T}) & H_\theta^{(22)}(z_{1:T}) & \cdots & H_\theta^{(2k)}(z_{1:T}) \\
\vdots & \vdots & \ddots & \vdots \\
H_\theta^{(k1)}(z_{1:T}) & H_\theta^{(k2)}(z_{1:T}) & \cdots & H_\theta^{(kk)}(z_{1:T})
\end{pmatrix}. \tag{4.30}
$$

We note the similarity between (4.28–4.30) and (4.7–4.8). We first consider the covering number bound. We assume for each mixture component $1 \leqslant i \leqslant k$, we have the almost sure bounds:

$$
\|\nabla_{\theta_i} \log p_{\theta_i}(z_{1:T})\| \lesssim c_i T, \quad \|\nabla_{\theta_i}^2 \log p_{\theta_i}(z_{1:T})\|_{\mathrm{op}} \lesssim c_i' T,
$$

where c_i, c_i' are some constants depending on system parameters such as state dimension, etc. As we will see in later sections, this is possible to show for many systems. We then have for $1 \leqslant i, j \leqslant k$,

$$
\|H_\theta^{(ij)}(z_{1:T})\|_{\mathrm{op}} \lesssim \left(\max_{1 \leqslant i \leqslant k} c_i\right)^2 T^2.
$$

Therefore we can apply Proposition C.1 to get

$$
H_\theta(z_{1:T}) \preceq \mathsf{blk\text{-}diag}\left\{ \left(\sum_{j=1}^k \|H_\theta^{(1j)}(z_{1:T})\|_{\mathrm{op}}\right) I_d, \ \ldots, \ \left(\sum_{j=1}^k \|H_\theta^{(kj)}(z_{1:T})\|_{\mathrm{op}}\right) I_d \right\} \preceq k \left(\max_{1 \leqslant i \leqslant k} c_i\right)^2 T^2 I_{kd},
$$

which implies $\mathcal{I}(\theta) = -\mathbb{E}_\theta\left[H_\theta(z_{1:T})\right] \preceq k \left(\max_{1 \leqslant i \leqslant k} c_i\right)^2 T^2 I_{kd}$. That is, we have

$$
\mathcal{I}_{\max} := k \left(\max_{1 \leqslant i \leqslant k} c_i\right)^2 T^2 I_{kd},
$$

which gives us a bound on the metric entropy under FI-norm:

$$
\log \mathcal{N}_{\mathrm{FI}}\left(\mathcal{P}, \varepsilon\right) \lesssim kd \log\left(k \left(\max_{1 \leqslant i \leqslant k} c_i\right) T \Big/ \varepsilon\right).
$$

This allows us to carry out the analysis done in Step 1.

For Step 2 onward, we inspect $w_\theta^{(i)}(z_{1:T})$:

$$
w_\theta^{(i)}(z_{1:T}) = \frac{1}{1 + \frac{\sum_{j \neq i} f_j p_{\theta_j}(z_{1:T})}{f_i p_{\theta_i}(z_{1:T})}} = \frac{1}{1 + \sum_{j \neq i} \frac{f_j}{f_i} \frac{p_{\theta_j}(z_{1:T})}{p_{\theta_i}(z_{1:T})}}.
$$

Ideally, each component of the mixture should be identifiable from the others given long enough trajectories. Hence, a natural assumption is that when $B = i$ (i.e., $z_{1:T} \sim p_{\theta_i}$), for any $j \neq i$ the following ergodic condition holds: there exists some constant $\Delta_{ij} > 0$ such that

$$
\frac{1}{T} \log\left(\frac{p_{\theta_j}(z_{1:T})}{p_{\theta_i}(z_{1:T})}\right) \xrightarrow{T \to \infty} -\Delta_{ij} \quad \text{a.s.} \tag{4.31}
$$

45

The immediate implication of (4.31) is that when $B = i$:

$$w_\theta^{(i)}(z_{1:T}) = \frac{1}{1 + \sum_{j \neq i} \frac{f_j}{f_i} \exp\left\{\log\left(\frac{p_{\theta_j}(z_{1:T})}{p_{\theta_i}(z_{1:T})}\right)\right\}} \xrightarrow{T \to \infty} \frac{1}{1 + \sum_{j \neq i} \frac{f_j}{f_i} \exp\left\{-\infty \cdot \Delta_{ij}\right\}} = 1 \quad \text{a.s.}$$

On a flip side, when $B = k \neq i$, we have:

$$w_\theta^{(i)}(z_{1:T}) = \frac{1}{1 + \sum_{j \neq i} \frac{f_j}{f_i} \exp\left\{\log\left(\frac{p_{\theta_j}(z_{1:T})}{p_{\theta_i}(z_{1:T})}\right)\right\}} \leqslant \frac{1}{1 + \frac{f_k}{f_i} \exp\left\{\log\left(\frac{p_{\theta_k}(z_{1:T})}{p_{\theta_i}(z_{1:T})}\right)\right\}}$$

$$\xrightarrow{T \to \infty} \frac{1}{1 + \frac{f_k}{f_i} \exp\left\{\infty \cdot \Delta_{ki}\right\}} = 0 \quad \text{a.s.}$$

Since $w_\theta^{(i)}(z_{1:T}) \geqslant 0$, we have $w_\theta^{(i)}(z_{1:T}) \xrightarrow{T \to \infty} 0$ a.s. when $B = k \neq i$. Therefore we conclude

$$w_\theta^{(i)}(z_{1:T}) \xrightarrow{T \to \infty} \mathbb{1}\{B = i\} \quad \text{a.s.} \tag{4.32}$$

For a concrete example of (4.31), suppose that the p_{θ_i}'s are Markovian for each $1 \leqslant i \leqslant k$, i.e., $p_{\theta_i}(z_{1:T}) = p(z_1) \prod_{t=1}^{T-1} p_{\theta_i}(z_{t+1} \mid z_t)$. The natural ergodicity assumption in this case is the following:

$$\frac{1}{T} \sum_{t=1}^{T-1} h_{ij}(z_t, z_{t+1}) \xrightarrow{T \to \infty} \mathbb{E}_{(z_t, z_{t+1}) \sim \pi_{\theta_i} \otimes p_{\theta_i}}\left[h_{ij}(z_t, z_{t+1})\right] \quad \text{a.s.,} \tag{4.33}$$

where $h_{ij}(z_t, z_{t+1}) := \log\left(\frac{p_{\theta_j}(z_{t+1}|z_t)}{p_{\theta_i}(z_{t+1}|z_t)}\right)$ and π_{θ_i} is the density of ergodic measure of the Markov process $\{z_t\}_{t=1}^\infty$ under p_{θ_i}. Here, the $\otimes$ notation denotes the following operation between a density π and a transition density p: $\pi \otimes p(z_t, z_{t+1}) := \pi(z_t) p(z_{t+1} \mid z_t)$. To see why this implies (4.31), observe that $\pi_{\theta_i} \otimes p_{\theta_i}$ is the density of the ergodic measure of the augmented process $\{(z_t, z_{t+1})\}_{t=1}^\infty$ under p_{θ_i}. Hence the right hand side of (4.33) reads:

$$\mathbb{E}_{(z_t, z_{t+1}) \sim \pi_{\theta_i} \otimes p_{\theta_i}}\left[h_{ij}(z_t, z_{t+1})\right] = \int_{\mathsf{Z}^2} \pi_{\theta_i} \otimes p_{\theta_i}(z_t, z_{t+1}) \log\left(\frac{p_{\theta_j}(z_{t+1}|z_t)}{p_{\theta_i}(z_{t+1}|z_t)}\right) \mathrm{d}z_t \, \mathrm{d}z_{t+1}$$

$$= -\int_{\mathsf{Z}^2} \pi_{\theta_i} \otimes p_{\theta_i}(z_t, z_{t+1}) \log\left(\frac{\pi_{\theta_i} \otimes p_{\theta_i}(z_t, z_{t+1})}{\pi_{\theta_i} \otimes p_{\theta_j}(z_t, z_{t+1})}\right) \mathrm{d}z_t \, \mathrm{d}z_{t+1}$$

$$= -\mathrm{KL}\left(\pi_{\theta_i} \otimes p_{\theta_i} \,\|\, \pi_{\theta_i} \otimes p_{\theta_j}\right) =: -\Delta_{ij}.$$

Following (4.32), we can again argue that for large T, the Hessian (and therefore the Fisher information) can be controlled by a block-diagonal matrix of the mixture components' Fisher information matrices

$$\mathcal{I}_{\text{blk-diag}}(\theta) := \begin{pmatrix} -\mathbb{E}_\theta^{(1)}\left[\nabla_{\theta_1}^2 \log p_{\theta_1}(z_{1:T})\right] & 0 & \cdots & 0 \\ 0 & -\mathbb{E}_\theta^{(2)}\left[\nabla_{\theta_2}^2 \log p_{\theta_2}(z_{1:T})\right] & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\mathbb{E}_\theta^{(k)}\left[\nabla_{\theta_k}^2 \log p_{\theta_k}(z_{1:T})\right] \end{pmatrix} \tag{4.34}$$

under the Loewner order.

Some remarks are in order. First, for an exact analogue of Proposition 4.8, we need to prove the Loewner order equivalence holds uniformly within a ball around $\theta_\star$. Our previous proof strategy requires characterizing non-asymptotic mixing behavior instead of the asymptotic convergence as in (4.31). In particular, if we are working with a Markov process and consider the augmented process $\{(z_t, z_{t+1})\}_{t=1}^\infty$ with an ergodic measure denoted by π, we expect the following Bernstein-type inequality to hold:[10]

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T-1} h(z_t, z_{t+1}) - \mathbb{E}_\pi\left[h(z_1, z_2)\right]\right| \geqslant s\right) \lesssim \exp\left(\frac{-Ts^2}{\mathrm{Var}_\pi(h) + C(\mathcal{H})s}\right), \quad h \in \mathcal{H},$$

where $C(\mathcal{H})$ typically quantifies boundedness of the function class $\mathcal{H}$. Such results are available for various classes of mixing processes (see e.g., [64–69]). In particular, stable LDS are geometrically β-mixing [70], and therefore [66, Theorem 1 and 2] are readily applicable. Specialized bounds for Markov chains are also available [see e.g., 71, Theorem 2.4]. This would allow us to obtain instance-optimal rates for the more general mixture dynamics, including but not limited to those considered in prior art (e.g., [59, 72]).

Second, we face the technical challenge that Hellinger identifiability (analogue of Proposition 4.9) does not in general hold when a subset of weights are equal. In the scalar case, as in Proposition 4.9, we worked around this issue by assuming a monotone order on the corresponding parameters to guarantee unique identifiability of parameters. In the general case, we need to redefine the notion of Hellinger identifiability to be symmetry aware. In particular, let $\mathcal{J} = (\mathcal{J}_1, \ldots, \mathcal{J}_\ell)$ partition the indices $[k]$ into $\ell \leqslant k$ equivalence classes, where $f_{i_1} = f_{i_2}$ for all $i_1, i_2 \in \mathcal{J}_i$, and $f_{i_1} \neq f_{j_1}$ for all $i_1 \in \mathcal{J}_i, j_1 \in \mathcal{J}_j$ with $i \neq j$. Next, let $\mathsf{Sym}_{\mathcal{J}}(\theta)$ denote the set of cardinality $\prod_{i=1}^\ell (|\mathcal{J}_i|!)$ which given a parameter vector $\theta \in \Theta^k$ enumerates all possible permutations within each equivalence class for θ. We then consider the following modified definition of Hellinger identifiability (cf. Definition 3.11), which states that there exists (γ_1, γ_2) such that:

$$\mathrm{d}_H(p_\theta, p_\star) \leqslant \gamma_1 \implies \min_{\theta^\pi \in \mathsf{Sym}_{\mathcal{J}}(\theta)} \|\theta^\pi - \theta_\star\| \leqslant \gamma_2 \cdot \mathrm{d}_H(p_\theta, p_\star).$$

We note the above symmetrized condition needs to be shown on a problem-specific basis.

We conclude by remarking that under the particular outline above, it seems necessary to require ergodicity of mixture *components* to obtain non-trivial results, as otherwise the behavior of Fisher information is difficult to analyze. However, we do know that, for example, LDS with non-mixing behavior can still be identified with parametric rate from a single trajectory [24] and so we postulate that e.g., identifying mixtures of LDS may also be possible without mixing assumptions. This reflects a limitation for our current instantiation of Hellinger localization for mixture recovery, which we leave addressing to future work.

Mixture with unknown weights. We now extend the above calculation to the fully general setting where the joint parameter of interest is:

$$(\theta, f) := \begin{pmatrix} \theta_1^\mathsf{T} & \cdots & \theta_k^\mathsf{T} & f^\mathsf{T} \end{pmatrix}^\mathsf{T} = \begin{pmatrix} \theta_{1,1} & \cdots & \theta_{1,d} & \cdots & \theta_{k,1} & \cdots & \theta_{k,d} & f_1 & \cdots & f_{k-1} \end{pmatrix}^\mathsf{T} \in \Theta^k \times \Delta_\mu^{k-1}.$$

---

[10]Here we only stated a schematic form. More precisely, under various mixing conditions, the right hand side might include additional polylog(T) factors.

The density now takes the form

$$p_{\theta,f}(z_{1:T}) = \sum_{i=1}^{k-1} f_i p_{\theta_i}(z_{1:T}) + \left(1 - \sum_{i=1}^{k-1} f_i\right) p_{\theta_k}(z_{1:T}).$$

The cross-terms of the information matrix between the mixture weights f and parameters θ are:

$$\partial_{f_i} \nabla_{\theta_j} \log p_{\theta,f}(z_{1:T}) = \left(\frac{w_{\theta,f}^{(j)}(z_{1:T}) w_{\theta,f}^{(k)}(z_{1:T})}{f_k} - \frac{w_{\theta,f}^{(j)}(z_{1:T}) w_{\theta,f}^{(i)}(z_{1:T})}{f_i}\right) \nabla_{\theta_j} \log p_{\theta_j}(z_{1:T}),$$

$$\partial_{f_i} \nabla_{\theta_i} \log p_{\theta,f}(z_{1:T}) = \left(\frac{1}{f_i} w_{\theta,f}^{(i)}(z_{1:T}) \left(1 - w_{\theta,f}^{(i)}(z_{1:T})\right) + \frac{w_{\theta,f}^{(i)}(z_{1:T}) w_{\theta,f}^{(k)}(z_{1:T})}{f_k}\right) \nabla_{\theta_i} \log p_{\theta_i}(z_{1:T}),$$

$$(4.35)$$

where the posterior weight

$$w_{\theta,f}^{(i)}(z_{1:T}) := \frac{f_i p_{\theta_i}(z_{1:T})}{\sum_{i=1}^{k} f_i p_{\theta_i}(z_{1:T})}.$$

is defined similarly as before. Under the same ergodicity assumptions (4.31) from the previous section, both terms of (4.35) will converge to zero. We now calculate the Hessian with respect to the mixture weights:

$$\partial_{f_j} \partial_{f_i} \log p_\theta(z_{1:T}) = \frac{w_{\theta,f}^{(i)}(z_{1:T}) w_{\theta,f}^{(k)}(z_{1:T})}{f_i f_k} - \frac{w_{\theta,f}^{(i)}(z_{1:T}) w_{\theta,f}^{(j)}(z_{1:T})}{f_i f_j} + \frac{w_{\theta,f}^{(j)}(z_{1:T}) w_{\theta,f}^{(k)}(z_{1:T})}{f_j f_k} - \left(\frac{w_{\theta,f}^{(k)}(z_{1:T})}{f_k}\right)^2,$$

$$\partial_{f_i}^2 \log p_\theta(z_{1:T}) = \frac{2 w_{\theta,f}^{(i)}(z_{1:T}) w_{\theta,f}^{(k)}(z_{1:T})}{f_i f_k} - \left(\frac{w_{\theta,f}^{(i)}(z_{1:T})}{f_i}\right)^2 - \left(\frac{w_{\theta,f}^{(k)}(z_{1:T})}{f_k}\right)^2.$$

Now under the ergodicity assumption (4.31), we can write the Hessian block $H_f(z_{1:T}) = \nabla_f^2 \log p_{\theta,f}(z_{1:T})$ and FI matrix block $\mathcal{I}(f) = -\mathbb{E}_{\theta,f}[H_f(z_{1:T})]$ for T large as follows:

$$H_f(z_{1:T}) \xrightarrow{T \to \infty} - \mathrm{diag}\left\{\left(\frac{w_{\theta,f}^{(1)}(z_{1:T})}{f_1}\right)^2, \ldots, \left(\frac{w_{\theta,f}^{(k-1)}(z_{1:T})}{f_{k-1}}\right)^2\right\} - \left(\frac{w_{\theta,f}^{(k)}(z_{1:T})}{f_k}\right)^2 \mathbb{1}\mathbb{1}^\mathsf{T},$$

$$\mathcal{I}(f) \xrightarrow{T \to \infty} \underbrace{\mathrm{diag}\left\{\mathbb{E}_{\theta,f}\left[\left(\frac{w_{\theta,f}^{(1)}(z_{1:T})}{f_1}\right)^2\right], \ldots, \mathbb{E}_{\theta,f}\left[\left(\frac{w_{\theta,f}^{(k-1)}(z_{1:T})}{f_{k-1}}\right)^2\right]\right\}}_{:=\mathcal{I}_{\mathsf{diag}}(f)} + \mathbb{E}_{\theta,f}\left[\left(\frac{w_{\theta,f}^{(k)}(z_{1:T})}{f_k}\right)^2\right]\mathbb{1}\mathbb{1}^\mathsf{T}$$

$$\preceq \mathcal{I}_{\mathsf{diag}}(f) + \frac{1}{\left(1 - \sum_{i=1}^{k-1} f_i\right)^2} \mathbb{1}\mathbb{1}^\mathsf{T}.$$

$$(4.36)$$

Now combining (4.34), (4.35) and (4.36) together, we obtain

$$
\mathcal{I}(\theta, f) \xrightarrow{T \to \infty}
\begin{pmatrix}
\mathcal{I}_{\text{blk-diag}}(\theta) & 0 \\
0 & \mathcal{I}_{\text{diag}}(f) + \mathbb{E}_{\theta, f}\left[\left(\frac{w_{\theta, f}^{(k)}(z_{1:T})}{f_k}\right)^2\right] \mathbb{1}\mathbb{1}^{\top}
\end{pmatrix}
$$
$$
\preceq
\begin{pmatrix}
\mathcal{I}_{\text{blk-diag}}(\theta) & 0 \\
0 & \mathcal{I}_{\text{diag}}(f) + \frac{1}{\left(1 - \sum_{i=1}^{k-1} f_i\right)^2} \mathbb{1}\mathbb{1}^{\top}
\end{pmatrix}.
$$
(4.37)

This sheds light on how we can generalize the proof in Section 4.1.2 to the case where the mixture weights f also need to be estimated, in addition to the mixture parameters θ. Indeed, the proof strategy would be similar to that of Section 4.1.2; however, one additional challenge is that one would need to verify the Hellinger identifiability of the mixture weights.

4.2 Dependent Regression under General Product-Noise Distributions

We next consider the following family of trajectory distributions $p_\theta(z_{1:T})$ over $\mathsf{Z} = \mathbb{R}^d$ parameterized by $\theta \in \Theta$ of the following form:

$$
z_{t+1} = M(z_t)\theta + w_t, \quad z_1 \sim \rho_1.
$$
(4.38)

Here, the matrix-valued map $M : \mathbb{R}^d \mapsto \mathbb{R}^{d \times p}$ is allowed to be non-linear, and assumed to be known. This setup generalizes the linear system identification problem detailed in Section 3.1 and has received considerable attention recently as a tractable form of non-linear system identification, especially when a control input is added to the matrix M, i.e., $z_{t+1} = M(z_t, u_t)\theta + w_t$ (see e.g., [73–76]); a more detailed literature review is given in Section 4.2.1. The noise variable w_t is drawn independently across time t from a distribution which has the following product density w.r.t. the Lebesgue measure on $\mathbb{R}^d$:

$$
p_{\boldsymbol{\phi}}(w) = \prod_{j=1}^d p_\phi(w_j), \quad p_\phi(w_1) = \exp\left\{-\phi(w_1)\right\}/Z(\phi), \quad Z(\phi) := \int \exp\left\{-\phi(w_1)\right\} dw_1,
$$
(4.39)

where $\phi : \mathbb{R} \mapsto \mathbb{R}$ is a known scalar function parameterizing the noise distribution. Hence, our setup differs from more standard settings in the following way: we do not need to assume the noise is either Gaussian or sub-Gaussian, but the functional form of the density is needed to solve the MLE. In what follows, given a vector $w \in \mathbb{R}^d$, we let $\boldsymbol{\phi}$ denote the function mapping $\mathbb{R}^d \mapsto \mathbb{R}^d$ defined as $\boldsymbol{\phi}(w) := (\phi(w_1), \ldots, \phi(w_d))$. With this notation, we can write the MLE (3.1) for (4.38) as:

$$
\hat{\theta}_{m,T} \in \arg\min_{\theta \in \Theta} \sum_{i=1}^m \sum_{t=1}^{T-1} \langle \mathbb{1}, \boldsymbol{\phi}(z_{t+1}^{(i)} - M(z_t^{(i)})\theta) \rangle.
$$
(4.40)

We assume the following regularity conditions on ϕ.

Definition 4.11 (Regularity conditions on ϕ). *We say that $\phi : \mathbb{R} \mapsto \mathbb{R}$ is (β_1, β_2)-regular for constants $1 \leqslant \beta_1, \beta_2 < \infty$, if the following conditions hold:*

(a) $\phi \in C^2(\mathbb{R})$ and $Z(\phi) < \infty$,

(b) Both $\lim_{|x| \to \infty} \phi(x) = \infty$ and $\lim_{|x| \to \infty} |\phi'(x)| \exp(-\phi(x)) = 0$,

49

(c) $\sigma_\phi^2 := (\mathbb{E}_{w\sim p_\phi}[(\phi'(w))^2])^{-1} < \infty,$

(d) $\mathbb{E}_{w\sim p_\phi}[(\phi''(w))^2] \leqslant \beta_1/\sigma_\phi^4,$ and

(e) $\mathbb{E}_{w\sim p_\phi}[(\phi'(w))^8] \leqslant \beta_2/\sigma_\phi^8.$

Before we look at a few examples of distributions satisfying Definition 4.11, we briefly describe the role of each condition. Condition (a) simply ensures that p_ϕ is a well-defined $C^2(\mathbb{R})$ density; having two derivatives is crucial in our framework, which relies on second order expansions. Condition (b) controls the growth of p_ϕ and states that p_ϕ must tend to zero in both directions which implies that $\mathbb{E}_{w\sim p_\phi}[\phi'(w)] = 0$, and that the following integration by parts (IBP) identity $\mathbb{E}_{w\sim p_\phi}[(\phi'(w))^2] = \mathbb{E}_{w\sim p_\phi}[\phi''(w)]$ holds (see Proposition 4.16); both identities play a key role in our analysis. Condition (c) ensures (via the IBP identity) that the amount of integrated curvature $\mathbb{E}_{w\sim p_\phi}[\phi''(w)]$ is bounded away from zero, and is necessary for non-degenerate Fisher Information matrices $\mathcal{I}(\theta)$; note that in the case when ϕ is convex, then this condition is equivalent to $\phi''(w)$ cannot equal zero almost everywhere. Conditions (d) and (e) are *hyper-contractivity* conditions; indeed by the IBP identity (d) is equivalent to $\mathbb{E}_{w\sim p_\phi}[(\phi''(w))^2] \leqslant \beta_1(\mathbb{E}_{w\sim p_\phi}[\phi''(w)])^2$, and similarly (e) states that $\mathbb{E}_{w\sim p_\phi}[(\phi'(w))^8] \leqslant \beta_2(\mathbb{E}_{w\sim p_\phi}[(\phi'(w))^2])^4$; by Jensen's inequality both β_1, β_2 must be $\geqslant 1$.

We next build some intuition for the generality of Definition 4.11, by giving a few examples below with explicit (β_1, β_2) constants; proofs for the examples are given in Section 4.2.3.

Example 4.12 (Multivariate normal distribution). *For any $\nu > 0$,*

$$\phi_\nu(x) = \frac{x^2}{2\nu^2}$$

satisfies the conditions in Definition 4.11 with $\sigma_{\phi_\nu}^2 = \nu^2$ and $(\beta_1, \beta_2) = (1, 105)$.

Example 4.13 (Smoothed "Bang-Bang" noise). *For $\nu > 0$, consider*

$$\phi_\nu(x) = \frac{x^2 + 1}{2\nu^2} - \log\cosh(x/\nu^2).$$

This corresponds to $p_{\phi_\nu} = \frac{1}{2}\mathsf{N}(1, \nu^2) + \frac{1}{2}\mathsf{N}(-1, \nu^2)$, a Gaussian mixture model with two ν^2 variance mixtures centered as ± 1. If $\nu \in (0, 1)$, then ϕ_ν satisfies the conditions in Definition 4.11 for $(\beta_1, \beta_2) = (c'/\nu^6, c''/\nu^{24})$, where c', c'' are universal positive constants. Note that when $\nu \to 0$, p_{ϕ_ν} approaches "bang-bang" noise $\frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1}$.

Example 4.14 (Smoothed Laplace distribution). *Let us define ϕ as:*

$$\phi_{c,\nu}(x) = \frac{1}{c}\log\cosh(cx/\nu), \quad c, \nu \in \mathbb{R}_{>0}.$$

As $c \to \infty$, we have that $\phi_{c,\nu}(x) \to |x/\nu|$ pointwise, so $p_{\phi_{c,\nu}}$ is a smoothed Laplace distribution with second-order curvature. Define $Z(c) := \int \cosh(cx)^{-1/c}\mathrm{d}x$. We have that $\phi_{c,\nu}$ satisfies the conditions of Definition 4.11 for $(\beta_1, \beta_2) = \left(\frac{2c}{\tanh^2(cZ(c)/4)}, \frac{16}{\tanh^8(cZ(c)/4)}\right)$.

With the data generating process $p_\theta(z_{1:T})$ in place, we now turn to the analysis of the MLE estimator in this model. We remark that the MLE estimator (4.40) in general for this problem is *not* the solution to a least-squares regression problem (unless p_ϕ is Gaussian), nor is it generally the

solution to convex optimization problem (unless ϕ is convex). Furthermore, as seen in Example 4.14, the noise does not necessarily have sub-Gaussian tails as well, as is the standard assumption in many dependent learning works (e.g., [1, 19, 24, 25, 38, 76]), although we note that some works have also considered heavier-tailed noise in various settings [27, 77–79]. The following is our main result regarding parameter recovery for the model (4.38).

Theorem 4.15. *Fix $\delta \in (0, 1)$, and suppose the following assumptions hold:*

(a) ϕ *is* (β_1, β_2)*-regular per Definition 4.11,*

(b) $M_1 := \sup_{\theta \in \Theta} \left(\mathbb{E}_{p_\theta} \left[\frac{1}{T-1} \sum_{t=1}^{T-1} \| M(z_t) \|_{\mathrm{op}}^4 \right] \right)^{1/4} < \infty,$

(c) $M_2 := \sup_{\theta \in \Theta} \left(\mathbb{E}_{p_\theta} \left[\frac{1}{T-1} \sum_{t=1}^{T-1} \| M(z_t) \|_{\mathrm{op}}^8 \right] \right)^{1/8} < \infty,$

(d) $\bar{\mu} := \sup_{\theta \in \Theta} \lambda_{\max} \left(\bar{\mathcal{I}}(\theta) \right) < \infty,$

(e) $\underline{\mu} := \inf_{\theta \in \Theta} \lambda_{\min} \left(\bar{\mathcal{I}}(\theta) \right) > 0,$ *and*

(f) $T \geqslant \beta_2^{1/2} d^2 (M_2/M_1)^4.$

Let $\Theta = \{ \theta \in \mathbb{R}^p \mid \| \theta \| \leqslant R \}$*, and let* $\hat{\theta}_{m,T}^\varepsilon$ *denote the max FI discretized MLE estimator (3.14) at resolution* $\varepsilon = \delta/(2\sqrt{2m})$*. Assume wlog that* $R, M_1, \bar{\mu} \geqslant 1$*, and define* $\kappa := \bar{\mu}/\underline{\mu}$*. Then:*

(a). *If* $\mathcal{P}$ *is* (γ_1, γ_2)*-identifiable (cf. Definition 3.11) and the number of trajectories* m *satisfies:*

$$m \gtrsim \max \left\{ p/\gamma_1^2 \cdot \log(c_1' p/\delta \cdot R\bar{\mu}T/\gamma_1), \ pT\gamma_2^2 \cdot \beta_1 M_1^4 \kappa/\underline{\mu} \cdot \log(c_1'' p/\delta \cdot \beta_1 RM_1 T\kappa \max\{\gamma_2, 1\}), \right.$$

$$\left. pT\gamma_2^2 \cdot M_1^4 \kappa/(\underline{\mu}\sigma_\phi^4) \cdot \log(c_1''' p/\delta \cdot RM_1 T\kappa \max\{\sigma_\phi^{-1}, 1\} \max\{\gamma_2, 1\}) \right\},$$

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

$$\| \hat{\theta}_{m,T}^\varepsilon - \theta_\star \|_{\bar{\mathcal{I}}(\theta_\star)}^2 \lesssim \frac{p \log(c_1 R\bar{\mu} \cdot mT/\delta)}{mT}. \tag{4.41}$$

Here, c_1, c_1', c_1'', c_1''' *are universal positive constants.*

(b). *On the other hand if* ϕ *is convex, then as long as the number of trajectories* m *satisfies*

$$m \gtrsim \max \left\{ p \cdot \beta_1 M_1^4/\underline{\mu}^2 \cdot \log(c_2' p/\delta \cdot \beta_1 RM_1 T\kappa), \right.$$

$$\left. p \cdot M_1^4/(\underline{\mu}^2 \sigma_\phi^4) \cdot \log(c_2'' p/\delta \cdot RM_1 T\kappa \max\{\sigma_\phi^{-1}, 1\}) \right\},$$

then with probability at least $1 - \delta$ *the rate (4.41) also holds, with* c_1 *replaced by* c_2*. Here,* c_2, c_2', c_2'' *are universal positive constants.*

Before turning to the proof of Theorem 4.15 (cf. Section 4.2.2), some remarks are in order. For the present discussion, we focus only on the characteristics of Theorem 4.15, deferring a detailed account and comparison to related work to Section 4.2.1. Focusing only on the parameters m, T, p, Theorem 4.15 states that (a) in general, if $m \gtrsim \tilde{\Omega}(pT)$, then the nearly (up to logarithmic factors) instance-optimal rate $\|\hat{\theta}_{m,T}^{\varepsilon} - \theta_\star\|_{\mathcal{I}(\theta_\star)}^2 \lesssim \tilde{O}(p/(mT))$ from (4.41) holds, and (b) if the scalar function ϕ parameterizing the noise distribution (4.39) is convex then the requirement on m improves to $m \gtrsim \tilde{\Omega}(p)$, with the final rate (4.41) remaining the same. In the ϕ convex case (b), the requirement on m is in general not improvable, as is shown by the lower bounds in [1, Section 6] for linear regression. For case (a), the worse dependence comes from the non-concavity of the log-likelihood when ϕ is not convex, which requires us to use (3.44), compared with the concave log-likelihood when ϕ is convex; see the discussion in Section 3.4.

We next comment on the stated assumptions. The regularity Assumption (a) was previously discussed in the remarks following Definition 4.11. Assumptions (b) and (c) control the growth of the feature matrix $M(z_t)$ over the trajectory $z_{1:T}$. By two applications of Jensen's inequality, with $T' := T - 1$,

$$\left(\frac{1}{T'} \sum_{t=1}^{T'} \mathbb{E}_{p_\theta} \|M(z_t)\|_{\mathrm{op}}^4\right)^{1/4} \leqslant \left(\frac{1}{T'} \sum_{t=1}^{T'} \sqrt{\mathbb{E}_{p_\theta} \|M(z_t)\|_{\mathrm{op}}^8}\right)^{1/4} \leqslant \left(\frac{1}{T'} \sum_{t=1}^{T'} \mathbb{E}_{p_\theta} \|M(z_t)\|_{\mathrm{op}}^8\right)^{1/8} \leqslant M_2,$$

and therefore $M_1 \leqslant M_2$, so assumption (c) actually implies (b). The growth of both M_1 and M_2 as a function of T governs the dependence on T for the minimum number of trajectories m; if M is almost surely bounded, then M_1, M_2 are trivially $O(1)$. Assumption (d) is implied by Assumption (b), since by another application of Jensen's inequality we have $\bar{\mu} \leqslant (M_1/\sigma_\phi)^2$. Assumption (e) is states that the FI matrix $\mathcal{I}(\theta)$ is not degenerate over Θ, and is necessary for parameter recovery. Assumption (f) is made to simplify the resulting expressions for the minimum number of trajectories m required, and can be easily removed.

4.2.1   Comparison to System Identification Literature

Linear Dynamical System Identification.   The LDS system identification problem reviewed in (3.5) is a special case of the model (4.38), with $p = d^2$, $\theta = \mathrm{vec}(A)$, and $M(z) = (z^\mathsf{T} \otimes I_d)$. Hence Theorem 4.15 can be thought of as a generalization of the results from [1] for multi-trajectory learning in LDS. However, there are some caveats/limitations to the extent that Theorem 4.15 truly generalizes the result. Focusing on Assumption (c), we have that $\|M(z)\|_{\mathrm{op}} = \|(z^\mathsf{T} \otimes I_d)\|_{\mathrm{op}} = \|z\|$, and hence Assumption (c) posits a uniform bound on the quantity $\chi(\theta) := \frac{1}{T'} \sum_{t=1}^{T'} \mathbb{E}_{p_\theta}[\|z_t\|^8]$ as θ varies over Θ. However, the quantity $\chi(\theta)$ exhibits two phase-transitions depending on the operator norm of $\mathrm{mat}(\theta)$. When $\|\mathrm{mat}(\theta)\|_{\mathrm{op}} < 1$, then $\chi(\theta) = O(1)$ (ignoring all constants other than T) by the ergodic theorem. On the other hand, when $\|\mathrm{mat}(\theta)\|_{\mathrm{op}} = 1$, then $\chi(\theta) = \mathrm{poly}(T)$. Finally, when $\|\mathrm{mat}(\theta)\|_{\mathrm{op}} = \rho > 1$, we have $\chi(\theta) = \rho^{O(T)}$. In the last regime, the bound (4.41) becomes sub-optimal compared with (3.6), and ends up scaling as $1/m$ instead of the optimal $1/(mT)$. Furthermore, in the $\|\mathrm{mat}(\theta)\|_{\mathrm{op}} = 1$ regime, the requirement on m becomes $m \gtrsim \mathrm{poly}(T)$, which is also not sharp. Thus, for LDS system identification, (4.38) is only sharp in the case when $R < 1$.

It is important to clarify that the main issue is not that the Hellinger framework requires stability/mixing of the process $z_{1:T}$, but instead the issue is that for LTI systems, the states z_t can easily grow exponential in T depending on the parameter A, which makes both covering and localization arguments extremely sensitive to minor perturbations in the parameters. The situation

for LDS can be somewhat reconciled by utilizing a closed-form lower bound for the trajectory-level Hellinger distance [56, Section 4], which would address the sub-optimal $m \gtrsim \mathrm{poly}(T)$ requirement when $R = 1$. However, when $R > 1$, this strategy would still not yield the correct rates, as we would still need to perform a covering argument under the hood.

In the least-squares analysis from [1], the issue of uniform convergence when A is not stable is handled elegantly via the special structure of the square loss. In particular, the square loss lends itself to an *offset basic inequality* [52] which allows the uniform convergence to be *self-normalized*, preventing unstable A's from adversely affecting the resulting covering numbers. We leave to future work a generalized form of self-normalization that can also be applied to log losses and the Hellinger framework. One possible starting point for this extension is the work of [80], which provides techniques to define and analyze offset empirical processes for logarithmic, and more generally exp-concave losses.

Non-linear System Identification. In its general form, problem (4.38) is typically studied in its controlled variant, i.e., $z_{t+1} = M(z_t, u_t)\theta + w_t$, where z_t is interpreted as the state of a discrete-time dynamical system, and u_t the control input at time t. We note that Theorem 4.15 for identifying the model (4.38) can be readily translated into this control setting with some minor modifications to incorporate the expectation over the control sequence u_t in the Fisher information matrix, and also to include the full map $M(z_t, u_t)$ in the definitions for M_1, M_2 in Assumptions (b), (c); we omit the exact result in the interest of space. Learning in the controlled formulation of (4.38) is studied mostly as an active learning problem, with a focus on designing optimal algorithms for selecting inputs [73–75, 81]; the necessity of active learning in the single-trajectory setting, absent smoothness conditions on $M(z, u)$, was demonstrated by [73]. The line of work from [75, 81] considers task-guided exploration, proposing an algorithm that quantifies which system parameters are most relevant to solving the task, and actively explores to minimize uncertainty in these parameters, achieving a near instance-optimal rate for the downstream task; this was later extended by [82] to general parametric dynamics models. Extending our Hellinger localization framework for active exploration, especially for downstream control tasks, is exciting future work.

Perhaps the most directly related work is that of [76], which shows that the feature map M being real-analytic is sufficient to allow *non-active* i.i.d. random control signals to suffice for parameter recovery. Their arguments proceed by showing that since the zeros of the real-analytic function have measure zero, this implies that the standard martingale small-ball conditions (cf. [24]) used to show a lower bound on the empirical covariance matrix hold generically. This idea is also applicable to our framework, and can be used to certify non-degeneracy of the Fisher information matrix as required by Assumption (e) in Theorem 4.15 for real-analytic feature maps.

4.2.2 Proof of Theorem 4.15

We first state a simple result regarding the noise distribution p_ϕ which will be useful in our analysis.

Proposition 4.16. *Given (a) and (b) of Definition 4.11, the following identities are valid:*

(a) $\mathbb{E}_{w \sim p_\phi}[\phi'(w)] = 0,$

(b) $\mathbb{E}_{w \sim p_\phi}[(\phi'(w))^2] = \mathbb{E}_{w \sim p_\phi}[\phi''(w)].$

Proof. We first note that:

$$p'_\phi(w) = -\phi'(w)\exp(-\phi(w))/Z(w) = -\phi'(w)p_\phi(w).$$

For (a), we see that:

$$\mathbb{E}_{w\sim p_\phi}[\phi'(w)] = \int \phi'(w)p_\phi(w)\mathrm{d}w = -\int p'_\phi(w)\mathrm{d}w = -\big[p_\phi(w)\big]_{-\infty}^\infty = 0,$$

where the last equality holds from the assumption that $\phi(w) \to \infty$ as $|w| \to \infty$, and hence $p_\phi(\pm\infty) = 0$. For (b) using integration by parts,

$$\mathbb{E}_{w\sim p_\phi}[(\phi'(w))^2] = \int (\phi'(w))^2 p_\phi(w)\mathrm{d}w = -\int \phi'(w)p'_\phi(w)\mathrm{d}w$$
$$= \int \phi''(w)p_\phi(w)\mathrm{d}w - \big[p_\phi\phi'\big]_{-\infty}^\infty = \mathbb{E}_{w\sim p_\phi}[\phi''(w)],$$

where the last equality holds by the limiting behavior $\lim_{|x|\to\infty}|\phi'(x)|\exp(-\phi(x)) = 0$. $\qquad\square$

Covering number bound. Let $\boldsymbol{\phi}'(x) := (\phi'(x_1),\ldots\phi'(x_d))$ and $\boldsymbol{\phi}''(x) := (\phi''(x_1),\ldots,\phi''(x_d))$. Using this notation, we compute the gradient and Hessian of the log probability as:

$$\nabla_\theta \log p_\theta(z_{t+1}\mid z_t) = M(z_t)^\mathsf{T}\boldsymbol{\phi}'(z_{t+1} - M(z_t)\theta),$$
$$\nabla_\theta^2 \log p_\theta(z_{t+1}\mid z_t) = -M(z_t)^\mathsf{T}\mathsf{diag}(\boldsymbol{\phi}''(z_{t+1} - M(z_t)\theta))M(z_t).$$

Consequently, we see that

$$\mathcal{I}(\theta) = -\sum_{t=1}^{T-1}\mathbb{E}_{p_\theta}[\nabla_\theta^2\log p_\theta(z_{t+1}\mid z_t)]$$
$$= \sum_{t=1}^{T-1}\mathbb{E}_{p_\theta}[M(z_t)^\mathsf{T}\mathsf{diag}(\boldsymbol{\phi}''(w_t))M(z_t)] = \frac{1}{\sigma_\phi^2}\sum_{t=1}^{T-1}\mathbb{E}_{p_\theta}[M(z_t)^\mathsf{T}M(z_t)],$$

where the last equality utilizes the IBP identity $\mathbb{E}_{w\sim p_\phi}[\phi''(w)] = \mathbb{E}_{w\sim p_\phi}[(\phi'(w))^2]$ from Proposition 4.16. Furthermore, we can construct a uniform bound $\mathcal{I}_{\max}$ using the definition of $\bar{\mu}$:

$$\mathcal{I}(\theta) \preccurlyeq \mathcal{I}_{\max} := T\bar{\mu}\cdot I_p, \quad \theta \in \Theta.$$

Therefore we have an upper bound for the metric entropy under the max-FI divergence:

$$\log\mathcal{N}_{\mathcal{I}_{\max}}(\mathcal{P},\varepsilon) \leqslant \log\mathcal{N}_{\|\cdot\|}\left(\Theta, \varepsilon\sqrt{\frac{1}{T\bar{\mu}}}\right) \leqslant p\log\left(\frac{3R}{\varepsilon}\sqrt{T\bar{\mu}}\right).$$

From Theorem 3.6, with probability at least $1 - \delta$,

$$\mathrm{d}_H^2(\hat{\theta}_{m,T}^\varepsilon, \theta_\star) \lesssim \frac{p\log(c_1R\bar{\mu}\cdot mT/\delta)}{m}. \tag{4.42}$$

Hence by the arguments outlined in (3.42) combined with Proposition A.1, as long as

$$m \gtrsim \gamma_1^{-2}p\log(c_1'pR\bar{\mu}\cdot T/(\gamma_1\delta)), \tag{4.43}$$

54

then whenever (4.42) holds, we have

$$\sup_{s \in [0,1]} \mathrm{d}_H^2((1-s)\theta_\star + s\hat{\theta}_{m,T}^\varepsilon, \theta_\star) \lesssim \frac{\gamma_2^2 T \bar{\mu} p \log(c_1 R \bar{\mu} \cdot mT/\delta)}{m}. \tag{4.44}$$

Call this event $\mathcal{E}_1$. Furthermore, since the log-likelihood of $z_{1:T}$ for $\theta \in \Theta$ can be written as:

$$\begin{aligned}
\log p_\theta(z_{1:T}) &= \sum_{t=1}^{T-1} \log p_\phi(z_{t+1} - M(z_t)\theta) + \text{const} \\
&= \sum_{t=1}^{T-1} \sum_{j=1}^{d} \log p_\phi(\langle z_{t+1} - M(z_t)\theta, e_j \rangle) + \text{const} \\
&= -\sum_{t=1}^{T-1} \sum_{j=1}^{d} \phi(\langle z_{t+1} - M(z_t)\theta, e_j \rangle) + \text{const},
\end{aligned}$$

where const does not depend on θ, when ϕ is convex, then $\mathcal{P}$ is log-concave (cf. Definition 3.5). It is not hard to see that $\mathrm{diam}(\Theta) = 2RT\bar{\mu}$. Hence from Theorem 3.6, with probability at least $1 - \delta$,

$$\sup_{s \in [0,1]} \mathrm{d}_H^2((1-s)\theta_\star + s\hat{\theta}_{m,T}^\varepsilon, \theta_\star) \lesssim \frac{p \log(c_2 R \bar{\mu} \cdot mT/\delta)}{m}. \tag{4.45}$$

Call this event $\mathcal{E}_{1,\mathrm{cvx}}$.

Estimate B_1 and B_2. We first focus on B_1. Let us fix a test vector $v \in \mathbb{R}^p$, and define $d_t := v^\mathsf{T} M^\mathsf{T}(z_t)\phi'(w_t)$, so that for $z_{1:T} \sim p_\theta$,

$$\langle v, \nabla_\theta \log p_\theta(z_{1:T}) \rangle = \sum_{t=1}^{T-1} d_t.$$

Our first observation is that, with $\mathcal{F}_t := \sigma(z_{1:t+1})$, we have $\mathbb{E}[d_t \mid \mathcal{F}_{t-1}] = \mathbb{E}[v^\mathsf{T} M^\mathsf{T}(z_t)\phi'(w_t) \mid \mathcal{F}_{t-1}] = 0$ by Proposition 4.16, and hence $(d_t)_{t \geq 1}$ is a MDS adapted to the filtration $(\mathcal{F}_t)_{t \geq 1}$. Next, we compute:

$$\mathbb{E}[d_t^2 \mid \mathcal{F}_{t-1}] = \mathbb{E}[v^\mathsf{T} M^\mathsf{T}(z_t)\phi'(w_t)\phi'(w_t)^\mathsf{T} M(z_t) v \mid \mathcal{F}_{t-1}] = v^\mathsf{T} M^\mathsf{T}(z_t)\mathbb{E}_{w \sim p_\phi}[\phi'(w)\phi'(w)^\mathsf{T}]M(z_t)v.$$

Now since $\mathbb{E}_{w \sim p_\phi}[\phi'(w_j)\phi'(w_k)] = (\mathbb{E}_{w \sim p_\phi}[\phi'(w)])^2 = 0$ for $j, k \in [d]$ with $j \neq k$ by coordinate-wise independence of p_ϕ and Proposition 4.16, we have that

$$\mathbb{E}_{w \sim p_\phi}[\phi'(w)\phi'(w)^\mathsf{T}] = \mathbb{E}_{w_1 \sim p_\phi}[(\phi'(w_1))^2] \cdot I_d = \sigma_\phi^{-2} \cdot I_d.$$

Hence,

$$\begin{aligned}
\mathbb{E}\left(\sum_{t=1}^{T-1} \mathbb{E}[d_t^2 \mid \mathcal{F}_{t-1}]\right)^2 &= \sigma_\phi^{-4} \mathbb{E}\left(\sum_{t=1}^{T-1} \|M(z_t)v\|^2\right)^2 \leq (T-1)\sigma_\phi^{-4} \sum_{t=1}^{T-1} \mathbb{E}\|M(z_t)v\|^4 \\
&\leq (T-1)\sigma_\phi^{-4} \|v\|^4 \sum_{t=1}^{T-1} \mathbb{E}\|M(z_t)\|_{\mathrm{op}}^4 \leq (T-1)^2 \sigma_\phi^{-4} \|v\|^4 M_1^4,
\end{aligned}$$

Now let us focus on $\mathbb{E}[d_t^4]$:

$$\mathbb{E}[d_t^4] = \mathbb{E}[(v^\mathsf{T} M^\mathsf{T}(z_t)\phi'(w_t)\phi'(w_t)^\mathsf{T} M(z_t)v)^2]$$
$$\leqslant \mathbb{E}[\|\phi'(w_t)\|^4 (v^\mathsf{T} M^\mathsf{T}(z_t)M(z_t)v)^2]$$
$$\leqslant \sqrt{\mathbb{E}_{w\sim p_\phi}[\|\phi'(w)\|^8]}\sqrt{\mathbb{E}[(v^\mathsf{T} M^\mathsf{T}(z_t)M(z_t)v)^4]}.$$

Next, we control

$$\mathbb{E}_{w\sim p_\phi}[\|\phi'(w)\|^8] = \mathbb{E}_{w\sim p_\phi}\left(\sum_{j=1}^{d}(\phi'(w_j))^2\right)^4 \leqslant d^3\mathbb{E}_{w\sim p_\phi}\left(\sum_{j=1}^{d}(\phi'(w_j))^8\right)$$
$$= d^4\mathbb{E}_{w_1\sim p_\phi}[(\phi'(w_1))^8] \leqslant d^4\beta_2\sigma_\phi^{-8}.$$

where the penultimate inequality follows from Hölder's inequality, and the last inequality follows from Definition 4.11. Hence,

$$\sum_{t=1}^{T-1}\mathbb{E}[d_t^4] \leqslant d^2\beta_2^{1/2}\sigma_\phi^{-4}\sum_{t=1}^{T-1}\sqrt{\mathbb{E}[(v^\mathsf{T} M^\mathsf{T}(z_t)M(z_t)v)^4]}$$
$$\leqslant d^2\beta_2^{1/2}\sigma_\phi^{-4}\|v\|^4\sum_{t=1}^{T-1}\sqrt{\mathbb{E}\|M(z_t)\|_{\mathrm{op}}^8}$$
$$\leqslant (T-1)d^2\beta_2^{1/2}\sigma_\phi^{-4}\|v\|^4\sqrt{\frac{1}{T-1}\sum_{t=1}^{T-1}\mathbb{E}\|M(z_t)\|_{\mathrm{op}}^8}$$
$$\leqslant (T-1)d^2\beta_2^{1/2}\sigma_\phi^{-4}\|v\|^4 M_2^4.$$

Now we will set $v = \mathcal{I}(\theta)^{-1/2}\bar{v}$ where $\bar{v} \in \mathbb{S}^{p-1}$ is a unit test vector; hence $\|v\| \leqslant \sqrt{\sigma_\phi^2/(\underline{\mu}T)}$. By Rosenthal's inequality for MDS (Theorem A.6), we have:

$$\left(\mathbb{E}\left(\sum_{t=1}^{T-1}d_t\right)^4\right)^{1/4} \lesssim \left(\mathbb{E}\left(\sum_{t=1}^{T-1}\mathbb{E}[d_t^2\mid\mathcal{F}_{t-1}]\right)^2\right)^{1/4} + \left(\sum_{t=1}^{T-1}\mathbb{E}[d_t^4]\right)^{1/4}$$
$$\lesssim \sqrt{T}\sigma_\phi^{-1}\|v\|M_1 + T^{1/4}d^{1/2}\beta_2^{1/8}\sigma_\phi^{-1}\|v\|M_2$$
$$\lesssim M_1/\sqrt{\underline{\mu}} + T^{-1/4}d^{1/2}\beta_2^{1/8}M_2/\sqrt{\underline{\mu}}$$
$$\lesssim M_1/\sqrt{\underline{\mu}},$$

where the last inequality holds from Assumption (f). Hence we have

$$\sup_{\theta_1,\theta_2\in\Theta} B_1(\theta_1,\theta_2) \lesssim M_1/\sqrt{\underline{\mu}}.$$

We next focus on B_2. We first fix a vector $q \in \mathbb{R}^d$, and observe that

$$\mathbb{E}_{w\sim p_\phi}[(q^\mathsf{T}\mathsf{diag}(\phi''(w))q)^2] = \sum_{i,j=1}^{d}q_i^2 q_j^2\mathbb{E}_{w\sim p_\phi}[\phi''(w_i)\phi''(w_j)]$$
$$= \sum_{i=1}^{d}q_i^4\mathbb{E}_{w\sim p_\phi}[(\phi''(w))^2] + \sum_{i\neq j}^{d}q_i^2 q_j^2(\mathbb{E}_{w\sim p_\phi}[\phi''(w)])^2 \leqslant \beta_1\sigma_\phi^{-4}\|q\|^4,$$

where the last inequality holds from Definition 4.11 and the IBP identity (cf. Proposition 4.16) $\mathbb{E}_{w \sim p_\theta}[\phi''(w)] = \mathbb{E}_{w \sim p_\theta}[(\phi'(w))^2] = \sigma_\phi^{-2}$. Hence fixing a test vector $v \in \mathbb{R}^p$, we have

$$\begin{aligned}
\mathbb{E}[(v^\mathsf{T} \nabla_\theta^2 \log p_\theta(z_{t+1} \mid z_t)v)^2] &= \mathbb{E}[(v^\mathsf{T} M^\mathsf{T}(z_t)\,\mathsf{diag}(\phi''(w_t))M(z_t)v)^2] \\
&= \mathbb{E}[\mathbb{E}[(v^\mathsf{T} M^\mathsf{T}(z_t)\,\mathsf{diag}(\phi''(w_t))M(z_t)v)^2 \mid z_t]] \\
&\leqslant \beta_1 \sigma_\phi^{-4} \mathbb{E}\|M(z_t)v\|^4 \leqslant \beta_1 \sigma_\phi^{-4}\|v\|^4 \mathbb{E}\|M(z_t)\|_{\mathrm{op}}^4.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}\left(\sum_{t=1}^{T-1} v^\mathsf{T} \nabla_\theta^2 \log p_\theta(z_{t+1} \mid z_t)v\right)^2 &\leqslant (T-1)^2\left[\frac{1}{T-1}\sum_{t=1}^{T-1}\mathbb{E}(v^\mathsf{T} \nabla_\theta^2 \log p_\theta(z_{t+1} \mid z_t)v)^2\right] \\
&\leqslant (T-1)^2 \beta_1 \sigma_\phi^{-4}\|v\|^4\left[\frac{1}{T-1}\sum_{t=1}^{T-1}\mathbb{E}\|M(z_t)\|_{\mathrm{op}}^4\right] \\
&\leqslant (T-1)^2 \beta_1 \sigma_\phi^{-4}\|v\|^4 M_1^4.
\end{aligned}$$

Now again we choose $v = \mathcal{I}(\theta)^{-1/2}\bar{v}$ for a unit norm $\bar{v} \in \mathbb{R}^p$. We then have

$$\left(\mathbb{E}\left(\sum_{t=1}^{T-1} v^\mathsf{T} \nabla_\theta^2 \log p_\theta(z_{t+1} \mid z_t)v\right)^2\right)^{1/2} \lesssim T\beta_1^{1/2}\sigma_\phi^{-2}\|v\|^2 M_1^2 \leqslant \beta_1^{1/2}M_1^2/\underline{\mu}.$$

Hence we have

$$\sup_{\theta_1,\theta_2 \in \Theta} B_2(\theta_1,\theta_2) \lesssim \beta_1^{1/2}\frac{M_1^2}{\underline{\mu}}.$$

Altogether, we can bound

$$\sup_{\theta_1,\theta_2 \in \Theta} \max\{B_1^2(\theta_1,\theta_2), B_2(\theta_1,\theta_2)\} \lesssim \frac{\beta_1^{1/2}M_1^2}{\underline{\mu}}. \tag{4.46}$$

Parameter error bound. We first cover the case where ϕ is not convex. Combining (4.44), (4.46), and Proposition A.1, as long as m satisfies (4.43) and also

$$m \gtrsim pT\gamma_2^2 \cdot \beta_1 M_1^4 \kappa/\underline{\mu} \cdot \log(c_1'' p/\delta \cdot \beta_1 R M_1 T \max\{\gamma_2,1\}\kappa), \tag{4.47}$$

then condition (3.22) holds on $\mathcal{E}_1$. Hence from Proposition 3.9, combining (3.23) with (4.44), we have on $\mathcal{E}_1$:

$$\|\hat{\theta}_{m,T}^\varepsilon - \theta_\star\|^2 \lesssim \frac{p\log(c_1 R\bar{\mu} \cdot mT/\delta)}{\underline{\mu}mT}.$$

We now turn to the case where ϕ is convex. Combining (4.45), (4.46), and Proposition A.1, we see that if m satisfies:

$$m \gtrsim p \cdot \beta_1 M_1^4/\underline{\mu}^2 \cdot \log(c_2' p/\delta \cdot \beta_1 R M_1 T \kappa), \tag{4.48}$$

Hence from Proposition 3.9, combining (3.23) with (4.45), we have on $\mathcal{E}_{1,\mathrm{cvx}}$,

$$\|\hat{\theta}_{m,T}^\varepsilon - \theta_\star\|^2 \lesssim \frac{p\log(c_2 R\bar{\mu} \cdot mT/\delta)}{\underline{\mu}mT}.$$

Verify FI radius. In order to verify the FI radius condition (3.24), we will make use of Proposition A.3. Fix $\theta_1, \theta_2 \in \Theta$ and a unit norm $v \in \mathbb{S}^{p-1}$. We have the following:

$$
\begin{aligned}
|v^\mathsf{T}(\mathcal{I}(\theta_1) - \mathcal{I}(\theta_2))v| &= \frac{1}{\sigma_\phi^2}\left| \mathbb{E}_{p_{\theta_1}}\left[\sum_{t=1}^{T-1} \|M(z_t)v\|^2\right] - \mathbb{E}_{p_{\theta_2}}\left[\sum_{t=1}^{T-1} \|M(z_t)v\|^2\right]\right| \\
&\overset{(a)}{\leqslant} \frac{\sqrt{2}}{\sigma_\phi^2}\left(\left\|\sum_{t=1}^{T-1}\|M(z_t)v\|^2\right\|_{\mathcal{L}^2(p_{\theta_1})} + \left\|\sum_{t=1}^{T-1}\|M(z_t)v\|^2\right\|_{\mathcal{L}^2(p_{\theta_2})}\right) \mathrm{d}_H(p_{\theta_1}, p_{\theta_2}) \\
&\overset{(b)}{\leqslant} \frac{2\sqrt{2}(T-1)M_1^2}{\sigma_\phi^2}\mathrm{d}_H(p_{\theta_1}, p_{\theta_2}),
\end{aligned}
$$

where (a) follows from Proposition A.3 and (b) follows from Jensen's inequality. Hence by the variational characterization of operator norm, for any $\theta \in \Theta$:

$$
\frac{\|\mathcal{I}(\theta) - \mathcal{I}(\theta_\star)\|_{\mathrm{op}}}{\lambda_{\min}(\mathcal{I}(\theta_\star))} \lesssim \frac{M_1^2}{\underline{\mu}\sigma_\phi^2}\mathrm{d}_H(\theta, \theta_\star).
$$

Hence from (4.44) and Proposition A.1, as long as m satisfies (4.43), (4.47), and

$$
m \gtrsim pT\gamma_2^2 \cdot M_1^4\kappa/(\underline{\mu}\sigma_\phi^4) \cdot \log(c_1'''p/\delta \cdot RM_1 T\kappa \max\{\sigma_\phi^{-1}, 1\}\max\{\gamma_2, 1\}),
$$

then the FI radius condition (3.24) holds on $\mathcal{E}_1$. On the other hand when ϕ is convex, from (4.45) and Proposition A.1, as long as m satisfies (4.48) and:

$$
m \gtrsim p \cdot M_1^4/(\underline{\mu}^2\sigma_\phi^4) \cdot \log(c_2''p/\delta \cdot RM_1 T\kappa \max\{\sigma_\phi^{-1}, 1\}), \tag{4.49}
$$

then the FI radius condition (3.24) holds on $\mathcal{E}_{1,\mathrm{cvx}}$. The result for both cases now follows from Proposition 3.9, specifically (3.25).

4.2.3 Proof of Regularity Conditions for Example Distributions

Proof for smoothed bang-bang noise (Example 4.13). We abbreviate $p_\nu = p_{\phi_\nu}$. We have that $\phi_\nu'(x) = \frac{x}{\nu^2} - \frac{1}{\nu^2}\tanh(x/\nu^2)$ and $\mathbb{E}_{x\sim p_\nu}[x^2] = 1 + \nu^2$. For any $\varepsilon \in (0,1)$, we have by Young's inequality

$$
(x - \tanh(x/\nu^2))^2 \geqslant (1-\varepsilon)x^2 + (1-1/\varepsilon)\tanh^2(x/\nu^2) \geqslant (1-\varepsilon)x^2 - (1/\varepsilon - 1).
$$

Hence

$$
\mathbb{E}_{x\sim p_\nu}[(x - \tanh(x/\nu^2))^2] \geqslant (1-\varepsilon)(1+\nu^2) - (1/\varepsilon - 1) =: \varphi_\nu(\varepsilon).
$$

Basic calculus yields

$$
\max_{\varepsilon\in(0,1)} \varphi_\nu(\varepsilon) = (\sqrt{1+\nu^2} - 1)^2 \text{ at } \varepsilon = 1/\sqrt{1+\nu^2}.
$$

Hence we have show that

$$
\sigma^{-2} = \mathbb{E}_{x\sim p_\nu}[(\phi_\nu'(x))^2] = \frac{1}{\nu^2}\mathbb{E}_{x\sim p_\nu}[(x - \tanh(x/\nu^2))^2] \geqslant \frac{(\sqrt{1+\nu^2} - 1)^2}{\nu^2}.
$$

Furthermore, imposing the restriction that $\nu \in (0, 1)$, a second order Taylor expansion around $\nu = 0$ yields that $\sqrt{1 + \nu^2} - 1 \geqslant \frac{\nu^2}{2^{3/2}}$ and hence $\sigma^{-2} \geqslant \nu^2/8$. Next, we compute $\phi_\nu''(x) = \frac{1}{\nu^2} - \frac{1}{\nu^4} \operatorname{sech}^2(x/\nu^2)$, and hence

$$|\phi_\nu''(x)| \leqslant \max\{1/\nu^2, 1/\nu^4\} = 1/\nu^4 = (\sigma^2/\nu^4)/\sigma^2 \leqslant (8/\nu^6)/\sigma^2.$$

Hence we can set $\beta_1 = 8/\nu^6$. Next, we have

$$\mathbb{E}_{x \sim p_\nu}[(\phi_\nu'(x))^8] \leqslant 128 \mathbb{E}_{x \sim p_\nu}\left[\frac{x^8}{\nu^{16}} + \frac{1}{\nu^{16}}\right].$$

For each mixture index $i \in \{1, 2\}$, we have

$$\mathbb{E}[x^8 \mid i] \leqslant 7^4 (\mathbb{E}[x^2 \mid i])^4 = 7^4(1 + \nu^2)^4 \leqslant 8 \cdot 7^4(1 + \nu^8).$$

Consequently,

$$\mathbb{E}_{x \sim p_\nu}[(\phi_\nu'(x))^8] \lesssim \frac{1}{\nu^{16}} = \frac{\sigma^8}{\nu^{16}} \cdot \frac{1}{\sigma^8} \lesssim \frac{1}{\nu^{24}}.$$

Hence we can set $\beta_2 \lesssim 1/\nu^{24}$. $\qquad\square$

Proof for smoothed Laplace noise (Example 4.14). The first and second derivatives of $\phi_{c,\nu}$ are:

$$\phi_{c,\nu}'(x) = \frac{1}{\nu} \tanh(cx/\nu) \in [-1/\nu, 1/\nu], \quad \phi_{c,\nu}''(x) = \frac{c}{\nu^2} \operatorname{sech}^2(cx/\nu) \in [0, c/\nu^2].$$

Define $Z(c, \nu) := \int \cosh(cx/\nu)^{-1/c} \mathrm{d}x$. By a change of variables, $Z(c, \nu) = \nu Z(c)$. Also for $t \in (0, 1/\nu)$,

$$\mathbb{P}_{x \sim p_\phi}(|\phi_{c,\nu}'(x)| \leqslant t) = \mathbb{P}_{x \sim p_\phi}(x \in [-\nu \tanh^{-1}(t\nu)/c, \nu \tanh^{-1}(t\nu)/c]) =: \mathbb{P}_{x \sim p_\phi}(x \in I_{c,\nu}(t)).$$

We control the RHS probability by:

$$\mathbb{P}_{x \sim p_\phi}(x \in I_{c,\nu}(t)) = \frac{1}{Z(c, \nu)} \int_{I_{c,\nu}(t)} \exp(-c^{-1} \log \cosh(cx/\nu)) \mathrm{d}x \leqslant \frac{|I_{c,\nu}(t)|}{Z(c, \nu)} = \frac{2 \tanh^{-1}(t\nu)}{c Z(c)}.$$

Choosing t' such that $\frac{2 \tanh^{-1}(t'\nu)}{c Z(c)} = 1/2$, i.e., $t' = \nu^{-1} \tanh(c Z(c)/4)$:

$$\sigma^{-2} = \mathbb{E}_{x \sim p_\phi}[(\phi_{c,\nu}'(x))^2] \geqslant (t')^2 \mathbb{P}_{x \sim p_\phi}(|\phi_{c,\nu}'(x)| \geqslant t') \geqslant \nu^{-2} \tanh^2(c Z(c)/4)/2.$$

Hence we have

$$\sigma^2 \in \left[\nu^2, \frac{2\nu^2}{\tanh^2(c Z(c)/4)}\right].$$

Now let us focus on controlling β_1. We have:

$$|\phi_{c,\nu}''(x)| \leqslant \frac{c}{\nu^2} = \frac{c\sigma^2}{\nu^2} \cdot \frac{1}{\sigma^2} \leqslant \frac{2c}{\tanh^2(c Z(c)/4)} \cdot \frac{1}{\sigma^2}.$$

Hence we can take $\beta_1 = \frac{2c}{\tanh^2(c Z(c)/4)}$. We next focus on controlling β_2. We have:

$$\mathbb{E}_{w \sim p_\phi}[(\phi'(w))^8] \leqslant \frac{1}{\nu^8} = \frac{\sigma^8}{\nu^8} \cdot \frac{1}{\sigma^8} \leqslant \frac{16}{\tanh^8(c Z(c)/4)} \cdot \frac{1}{\sigma^8}.$$

Hence we can take $\beta_2 = \frac{16}{\tanh^8(c Z(c)/4)}$. $\qquad\square$

4.3 Non-Monotonic Sinusoidal GLM Dynamics

For our next setup, we consider the following generalized linear model (GLM) of dynamics, given parameters $A \in \mathbb{R}^{d \times d}$,

$$z_{t+1} = \sin(Az_t) + w_t, \quad z_0 = 0, \quad w_t \sim \mathsf{N}(0, \sigma^2 I_d), \tag{4.50}$$

where $\sin(\cdot)$ is overloaded to apply component-wise given a vector input, and w_t is drawn independently across time. While more general GLM dynamics $z_{t+1} = \phi(Az_t) + w_t$ have been studied in the literature in the context of system identification [19, 27, 38, 83, 84], the specific sinusoidal GLM we consider is more challenging as it is an instance of a *non-monotonic, non-expansive*[11] activation function. Furthermore, we do not impose any stability assumptions on the A matrix in (4.50), as is done in prior works. The following theorem is our main result for parameter recovery in this model. In the following result, we let $\hat{\theta}^\varepsilon_{m,T} = \text{vec}(\hat{A}^\varepsilon_{m,T})$ and $\theta_\star = \text{vec}(A_\star)$.

Theorem 4.17. *Fix $\delta \in (0,1)$. Consider the max FI discretized MLE at resolution $\varepsilon = \delta/(2\sqrt{2m})$ over the set $\Theta = \{A \in \mathbb{R}^{d \times d} \mid \|A\|_F \leqslant R\}$ for $R \geqslant 1$. Put $A_{\star,\min} := \min_{j \in [d]} \|A_\star[j]\|$, where $A_\star[j] \in \mathbb{R}^d$ denotes the j-th row of $A_\star$, and suppose $A_{\star,\min} > 0$. Suppose also that $T \gtrsim d^2$. There exists constants Φ_i, $i \in \{1,2,3\}$, which scale as $\text{poly}(\sigma, 1/\sigma, 1/A_{\star,\min}, 1/d)$, such that if m satisfies for universal positive constants c_0, c_1, c_2:*

$$m \gtrsim \max\left\{\Phi_1 d^2 \log\left(\frac{c_1 \Phi_1 \Xi}{\delta}\right), \Phi_2 d^{10} T \log\left(\frac{c_2 \Phi_2 \Xi}{\delta}\right), \Phi_3 d^{11} T \log\left(\frac{c_3 \Phi_3 \Xi}{\delta}\right)\right\}, \quad \Xi := \frac{dRT(d + \sigma^2)}{\sigma^2},$$

then with probability at least $1 - \delta$ over $\mathcal{D}_{m,T}$,

$$\|\hat{\theta}^\varepsilon_{m,T} - \theta_\star\|^2_{\bar{\mathcal{I}}(\theta_\star)} \lesssim \frac{d^2}{mT} \log\left(\frac{c_0 RmT(d + \sigma^2)}{\sigma^2 \delta}\right).$$

The precise expressions for Φ_i are given in the proof.

Note that

$$\|\hat{\theta}^\varepsilon_{m,T} - \theta_\star\|^2_{\bar{\mathcal{I}}(\theta_\star)} = \frac{1}{\sigma^2(T-1)} \sum_{t=1}^{T-1} \mathbb{E}_{p_{\theta_\star}}[\text{diag}(\cos^2(A_\star z_t))(\hat{A}^\varepsilon_{m,T} - A_\star)z_t z_t^\mathsf{T}(\hat{A}^\varepsilon_{m,T} - A_\star)^\mathsf{T}],$$

where we emphasize that the expression on the RHS is over a *fresh* trajectory $z_{1:T} \sim p_{\theta_\star}$ that is independent of $\mathcal{D}_{m,T}$. Nevertheless, there is not a simple closed-form reduction and hence we leave it in its present form. If we treat σ, R, and $A_{\star,\min}$ as constants, then Theorem 4.17 states that whenever both $m \gtrsim \tilde{\Omega}(d^{11}T)$ and $T \gtrsim d^2$, then the nearly (up to logarithmic factors) instance-optimal rate $\|\hat{\theta}^\varepsilon_{m,T} - \theta_\star\|^2_{\bar{\mathcal{I}}(\theta_\star)} \lesssim \tilde{O}(d^2/(mT))$ holds with high probability. We suspect that the $m \gtrsim \tilde{\Omega}(d^{11})$ requirement is sub-optimal and can be further improved with a more refined analysis. On the other hand, the requirement on $T \gtrsim d^2$ is made to simplify the expressions in the proof and can be removed. Furthermore, in our proof we show that $\lambda_{\min}(\bar{\mathcal{I}}(\theta_\star)) \gtrsim 1/d^2$, which implies the parameter bound $\|\hat{\theta}^\varepsilon_{m,T} - \theta_\star\|^2 \lesssim \tilde{O}(d^4/(mT))$. We leave to future work a sharp analysis of $\lambda_{\min}(\bar{\mathcal{I}}(\theta_\star))$ to determine the optimal *un-weighted* parameter error bound.

---

[11]An activation function $\phi(x)$ is *expansive* if there exists a $\zeta > 0$ such that $|\phi(x) - \phi(y)| \geqslant \zeta|x - y|$ for all $x, y \in \mathbb{R}$.

Comparison to existing results. To the best of our knowledge, this is the first rate for parameter recovery in any GLM dynamics model in the multi-trajectory setting, which obtains a nearly instance-optimal rate of $\tilde{O}(d^2/(mT))$. The previous sharpest rate for this problem utilizes the fact that the MLE $\hat{\theta}_{m,T}$ for this problem involves solving a realizable, parametric least-squares ERM problem, and hence the reduction described in Section 3.1 to the result of [2] would provide a similar bound on the *excess risk*, but not the weighted parameter error directly; as described in Section 3.1, however, without verification of the weakly sub-Gaussian condition specifically for the function class $\{\sin(A_1 x) - \sin(A_2 x) \mid A_1, A_2 \in \Theta\}$ (which to the best of our knowledge has not been shown in the literature), the best burn-in requirement on m, T that this reduction can provide depends exponentially on the process dimension d. On the other hand, Theorem 4.17 requires that both $m \gtrsim \text{poly}(d) \cdot T$ and $T \gtrsim \text{poly}(d)$ requirement; while it is likely that our exact polynomial dependence is not optimal, we are able to break the exponential in d barrier of existing results.

In the single-trajectory setting, several recent works have explicitly studied parameter recovery for GLM dynamics [19, 27, 38, 83, 84]. However, the specific setting (4.50) we consider does not satisfy the requisite assumptions for any of these works, and hence these works cannot be used as a basis for reduction. To start, the works [19, 27, 38] all assume a model class with a 1-Lipschitz monotonic activation function ϕ, with fast rates further requiring ϕ to be expansive. The sin function only satisfies the 1-Lipschitz requirement; the bounded and oscillatory nature of sin violates the other assumptions. The work [83] requires a one-point convexity assumption on the population loss, which is challenging to verify; they are only able to verify their condition assuming uniformly monotonic activations (i.e., $\phi' \geq \zeta > 0$). Regarding stability of $A_\star$, [19, 38] additionally assume Lyapunov stability conditions, specifically that there exists a diagonal positive definite K and scalar $\rho < 1$ such that $A_\star^\top K A_\star \preceq \rho \cdot K$; however, it is immediately obvious that this does not hold in our setting as we permit $A_\star = c \cdot I_d$ for $c > 1$, which would require $\rho \geq c^2 > 1$. This also violates the explicit assumption made in some works that $\|A_\star\|_{\text{op}} < 1$ [84]. Other works such as [27, 83] made explicit exponential regularity assumptions on the trajectories that given a noise sequence $\{w_t\}_{t=1}^{T-1}$, the difference in states from two initial states will expand at most by a factor of $\rho = 1 + O(1/T)$ every timestep; however, this is also violated in our setting.[12]

Hellinger identifiability for sinusoidal GLMs. Before we turn to applying the Hellinger localization framework to this problem, we discuss the main technical challenge: absent the strict monotonically increasing activation function assumption $\phi'(x) \geq \gamma > 0$, establishing both that (a) $\mathcal{I}(\theta) \succeq \Omega(T) \cdot I_{d^2}$ and that (b) $d_H^2(\hat{p}_{m,T}^\varepsilon, p_\star) \leq \gamma^2$ implies $\|\hat{\theta}_{m,T}^\varepsilon - \theta_\star\|^2 \lesssim \gamma^2$ (i.e., Hellinger identifiability (Definition 3.11)) becomes substantially more challenging. A key step towards establishing identifiability is to show the bound $\gamma^2 := \mathbb{E}_{z \sim \mathsf{N}(0, \sigma^2 I_d)}[(\sin(\langle u_1, z\rangle) - \sin(\langle u_2, z\rangle))^2] \gtrsim \sigma^2 \|u_1 - u_2\|^2$ when γ^2 is sufficiently small. We believe this result, which is detailed in Appendix D, to be of independent interest, and may be helpful in e.g., analyzing neural networks with sinusoidal activation functions [85]. We remark that a similar bound is shown for ReLU activations in [38, Lemma 11], in particular for $z \sim \mathsf{N}(\mu, \sigma^2 I_d)$, $\mathbb{E}_z[(\text{ReLU}(\langle u_1, z\rangle) - \text{ReLU}(\langle u_2, z\rangle))^2] \geq \frac{\sigma^2}{4} e^{-\|\mu\|^2/\sigma^2} \|u_1 - u_2\|^2$. A key difference is in how this style of result is used in our analysis versus in [38]. In our

---

[12]Concretely, our setup does not satisfy [27, Assumption 4] for any $\rho \leq 1 + O(1/T)$, as we now show. Let $\Phi_t(z)$ denote the value of z_t following the dynamics $z_{t+1} = \sin(2z_t)$ starting at $z_1 = z$. Suppose there exists positive $c_1, \rho = 1 + c_2/T$ such $|\Phi_t(z) - \Phi_t(z')| \leq c_1 \rho^t |z - z'|$ for all $z, z' \in \mathbb{R}$ and $t \in \mathbb{N}$. Clearly we have $\Phi_t(0) = 0$ for all t. Furthermore, one can show that $\lim_{t\to\infty} \Phi_t(z) = r^\star$, where $r^\star \approx 0.94775$ is the unique solution to $r = \sin(2r)$ in $(0, 1]$, for all $z \in (0, 1]$. Now, let T_0 be such that $|\Phi_t(\bar{z}) - r^\star| \leq r^\star/2$ for all $t \geq T_0$, where $\bar{z} := r^\star/(4c_1 e^{c_2})$ (we can always take c_1, c_2 large enough so that $\bar{z} \in (0, 1)$). Hence, for any $T \geq T_0$, we have $r^\star/2 \leq |\Phi_T(\bar{z})| \leq c_1 \rho^T |\bar{z}| \leq c_1 e^{c_2} |\bar{z}| = r^\star/4$, a contradiction.

61

analysis, the Hellinger identifiability is only used for the first two timestep as noted in Remark 3.12, and hence we only need to consider taking expectation over $z \sim \mathsf{N}(0, \sigma^2 I_d)$. On the other hand, [38] proves identifiability at every timestep in order to relate parameter recovery error to average prediction error; hence their analysis is required to consider non-zero means μ's, leading to parameter error rates that have an e^d dependence on the dimension in general (cf. [38, Theorem 3]).

4.3.1 Proof of Theorem 4.17

We will let $\theta = \text{vec}(A) \in \mathbb{R}^{d^2}$, $\Theta = \{\theta \in \mathbb{R}^{d^2} \mid \|\theta\| \leqslant R\}$, and define the map $M(z) := (z^\mathsf{T} \otimes I_d)$ so that $Az = M(z)\text{vec}(A) = M(z)\theta$. For what follows, we will often use A and θ interchangeably, and similarly for $A_\star$ and $\theta_\star$, choosing whichever notation is more convenient.

Step 1: Covering number bound. Define $h_\theta(z) := \sin(M(z)\theta)$ and its Jacobian w.r.t. θ $D_\theta h_\theta(z) = \text{diag}(\cos(M(z)\theta))M(z)$ (similar to $\sin(\cdot)$, $\cos(\cdot)$ is also overloaded to apply component-wise given a vector input). We observe that

$$\nabla_\theta \log p_\theta(z_{t+1} \mid z_t) = -\frac{1}{\sigma^2}(D_\theta h_\theta(z_t))^\mathsf{T}(h_\theta(z_t) - z_{t+1}),$$

and hence

$$\mathcal{I}(\theta) = \frac{1}{\sigma^2} \sum_{t=1}^{T-1} \mathbb{E}_{z_t \sim p_\theta(\cdot \mid z_{t-1})}[(D_\theta h_\theta(z_t))^\mathsf{T}(D_\theta h_\theta(z_t))].$$

We evaluate:

$$\mathbb{E}[(D_\theta h_\theta(z_t))^\mathsf{T}(D_\theta h_\theta(z_t))] = \mathbb{E}[M(z_t)^\mathsf{T} \text{diag}(\cos^2(M(z_t)\theta))M(z_t)].$$

Let us start with an upper bound. Since $\cos(x)^2 \in [0, 1]$, we can see that $\text{diag}(\cos^2(M(z_t)\theta)) \preccurlyeq I_d$. This allows us to simplify the expression:

$$\mathcal{I}(\theta) \preccurlyeq \frac{1}{\sigma^2} \sum_{t=1}^{T-1} \mathbb{E}_{z_t \sim p_\theta(\cdot \mid z_{t-1})}[M(z_t)^\mathsf{T} M(z_t)]$$

$$= \frac{1}{\sigma^2} \sum_{t=1}^{T-1} \mathbb{E}_{z_t \sim p_\theta(\cdot \mid z_{t-1})}[(z_t^\mathsf{T} \otimes I_d)^\mathsf{T}(z_t^\mathsf{T} \otimes I_d)]$$

$$= \frac{1}{\sigma^2} \sum_{t=1}^{T-1} \mathbb{E}_{z_t \sim p_\theta(\cdot \mid z_{t-1})}[z_t z_t^\mathsf{T}] \otimes I_d.$$

We now turn to analyzing $\mathbb{E}_{z_t \sim p_\theta(\cdot \mid z_{t-1})}[z_t z_t^\mathsf{T}]$. Expanding $z_t = \mu_{t-1} + w_{t-1}$ for $\mu_{t-1} = h_\theta(z_{t-1})$ and observing that $\mathbb{E}[w_{t-1}] = 0$, we can see:

$$\mathbb{E}_{z_t \sim p_\theta(\cdot \mid z_{t-1})}[z_t z_t^\mathsf{T}] = \mathbb{E}_{w_{t-1}}[(\mu_{t-1} + w_{t-1})(\mu_{t-1} + w_{t-1})^\mathsf{T}]$$

$$= \mathbb{E}_{w_{t-1}}[\mu_{t-1}\mu_{t-1}^\mathsf{T} + \mu_{t-1}w_{t-1}^\mathsf{T} + w_{t-1}\mu_{t-1}^\mathsf{T} + w_{t-1}w_{t-1}^\mathsf{T}]$$

$$= \mu_{t-1}\mu_{t-1}^\mathsf{T} + \sigma^2 I_d.$$

We fix a $v \in \mathbb{R}^d$ with unit norm, observe $\sin(x)^2 \in [0, 1]$, and bound the outer product of μ_{t-1}:

$$v^\mathsf{T}\mu_{t-1}\mu_{t-1}^\mathsf{T}v = (v^\mathsf{T}\mu_{t-1})^2 \leqslant \|v\|\|\mu_{t-1}\| \leqslant d.$$

Putting this all together gives the maximum eigenvalue of the Fisher information:

$$\lambda_{\max}(\mathcal{I}(\theta)) \leqslant \frac{T}{\sigma^2}(d + \sigma^2).$$

We note that this is a parameter agnostic upper bound, and as such we can set $\mathcal{I}_{\max} = \frac{T}{\sigma^2}(d+\sigma^2) \cdot I_p$. By Proposition D.5 and the data processing inequality, for any $\theta \in \Theta$,

$$\mathrm{d}_H^2(\theta, \theta_\star) \lesssim \min\left\{1, \frac{1}{\sigma^2}\right\} \Longrightarrow \|\theta - \theta_\star\|^2 \lesssim d \max\left\{1, \frac{1}{\sigma^2}\right\} \mathrm{d}_H^2(\theta, \theta_\star).$$

This shows that $\mathcal{P}$ is (γ_1, γ_2)-identifiable (cf. Definition 3.11) for constants:

$$\gamma_1 \asymp \min\{1, 1/\sigma\}, \quad \gamma_2 \asymp \sqrt{d}\max\{1, 1/\sigma\}.$$

Hence from (3.40),

$$\mathrm{d}_H^2(\hat{\theta}_{m,T}^\varepsilon, \theta_\star) \lesssim \min\left\{1, \frac{1}{\sigma^2}\right\} \Longrightarrow \sup_{\theta \in \mathrm{conv}\{\hat{\theta}_{m,T}^\varepsilon, \theta_\star\}} \mathrm{d}_H^2(\theta, \theta_\star) \lesssim \frac{dT(d+\sigma^2)}{\sigma^2}\max\left\{1, \frac{1}{\sigma^2}\right\}\mathrm{d}_H^2(\hat{\theta}_{m,T}^\varepsilon, \theta_\star).$$

$$(4.51)$$

Our next step is to apply Theorem 3.6 to ensure that the LHS condition in (4.51) holds. To do this, we shall estimate the covering number of $\mathcal{P}$ in the max FI divergence through the ℓ_2 covering number. If we fix some $\theta \in \Theta$ and let θ' denote its closest element in an ε-covering of Θ in ℓ_2,

$$\mathrm{d}_{\mathcal{I}_{\max}}(\theta, \theta') = \|\theta - \theta'\|_{\mathcal{I}_{\max}} \leqslant \sqrt{\lambda_{\max}(\mathcal{I}_{\max})}\|\theta - \theta'\| \leqslant \frac{\sqrt{T}}{\sigma}\sqrt{d+\sigma^2}\varepsilon.$$

This allows the relation

$$\mathcal{N}_{\mathcal{I}_{\max}}(\mathcal{P}, \varepsilon) \leqslant \mathcal{N}_{\|\cdot\|}\left(\Theta, \frac{\varepsilon\sigma}{\sqrt{T(d+\sigma^2)}}\right) \leqslant \left(3R\frac{\sqrt{T(d+\sigma^2)}}{\varepsilon\sigma}\right)^{d^2}.$$

We now apply Theorem 3.6 with $\varepsilon = \delta/(2\sqrt{2m})$ to conclude that with probability at least $1 - \delta$:

$$\mathrm{d}_H^2(\hat{\theta}_{m,T}^\varepsilon, \theta_\star) \lesssim \frac{d^2}{m}\log\left(\frac{c_0 RmT(d+\sigma^2)}{\sigma^2\delta}\right),$$

$$(4.52)$$

where c_0 is a universal positive constant. Call this event $\mathcal{E}_1$. If we define $1/\Phi_0 := \min\{1, 1/\sigma^2\}$, plugging this bound into (4.51) and applying Proposition A.1 yields that if m satisfies

$$m \gtrsim \Phi_0 d^2 \log\left(\frac{c_0' \Phi_0 dRT(d+\sigma^2)}{\sigma^2\delta}\right),$$

$$(4.53)$$

where c_0' is a universal constant, then the following also holds on $\mathcal{E}_1$:

$$\sup_{\theta \in \mathrm{conv}\{\hat{\theta}_{m,T}^\varepsilon, \theta_\star\}} \mathrm{d}_H^2(\theta, \theta_\star) \lesssim \frac{d^4 T}{m}\max\left\{\frac{1}{d}, \frac{1}{\sigma^2}, \frac{1}{\sigma^4}\right\}\log\left(\frac{c_0 RmT(d+\sigma^2)}{\sigma^2\delta}\right).$$

$$(4.54)$$

For what follows, we define the set

$$\Theta' := \{\theta \in \Theta \mid \forall j \in [d], \|\mathrm{mat}(\theta)[j] - \mathrm{mat}(\theta_\star)[j]\| \leqslant \|\mathrm{mat}(\theta_\star)[j]\|/2\}.$$

A key property of Θ' that we will utilize is that $\theta \in \Theta'$ implies $\|\mathrm{mat}(\theta)[j]\| \geqslant \|\mathrm{mat}(\theta_\star)[j]\|/2$ for all $j \in [d]$ by the triangle inequality.

Ensuring that $\hat{\theta}^\varepsilon_{m,T} \in \Theta'$ on $\mathcal{E}_1$. Using Proposition D.5 and (4.52), we have that if m satisfies,

$$\frac{d^2}{m} \log\left(\frac{c_0 RmT(d+\sigma^2)}{\sigma^2\delta}\right) \lesssim \min\left\{(A^2_{\star,\min}\min\{1,\sigma^2\}), \min\{1, 1/\sigma^2\}\right\} \triangleq 1/\Phi_1.$$

then we have that $\hat{\theta}^\varepsilon_{m,T} \in \Theta'$ on $\mathcal{E}_1$. Using Proposition A.1, this holds whenever m satisfies:

$$m \gtrsim \Phi_1 d^2 \log\left(\frac{c_1\Phi_1 dRT(d+\sigma^2)}{\sigma^2\delta}\right), \tag{4.55}$$

where c_1 is another univeral constant. We can note that since $\Phi_1 \geqslant \Phi_0$, this constraint automatically satisfies (4.53) (possibly after adjusting the value of c_1).

Lower bound FI matrix. Our first task is to estimate a lower bound on the FI matrix for $\theta \in \Theta'$. Recall the Fisher information from earlier:

$$\mathcal{I}(\theta) = \frac{1}{\sigma^2}\sum_{t=1}^{T-1}\mathbb{E}_{z_t\sim p_\theta}[(D_\theta h_\theta(z_t))^\mathsf{T}(D_\theta h_\theta(z_t))]$$

We can expand $z_t = \mu_{t-1} + w_{t-1}$ for $\mu_{t-1} = h_\theta(x_{t-1})$ to expand the dependency of the expectation on the previous timestep.

$$\begin{aligned}
\mathbb{E}[(D_\theta h_\theta(z_t))^\mathsf{T}(D_\theta h_\theta(z_t))] \\
&= \mathbb{E}[M(z_t)^\mathsf{T}\,\mathsf{diag}(\cos^2(M(z_t)\theta))M(z_t)] \\
&= \mathbb{E}[\mathbb{E}[M(\mu_{t-1}+w_{t-1})^\mathsf{T}\,\mathsf{diag}(\cos^2(M(\mu_{t-1}+w_{t-1})\theta))M(\mu_{t-1}+w_{t-1}) \mid \mu_{t-1}]],
\end{aligned}$$

We choose a ν_0 to be specified later, and observe that by Proposition D.1 and a union bound, with $A = \mathrm{mat}(\theta)$ and $A[j] \in \mathbb{R}^d$ denoting the j-th row of A,

$$\mathbb{P}\left(\bigcup_{j\in[d]}\left\{\cos^2(\langle A[j],\mu_{t-1}\rangle + \langle A[j],w_{t-1}\rangle) \leqslant \frac{\nu_0}{c_0 d}\min\{1,\sigma\|A[j]\|\}\right\}\,\Big|\,\mu_{t-1}\right) \leqslant \nu_0.$$

Call this event, which depends on μ_{t-1}, $\mathcal{E}_{\mu_{t-1}}$. Letting $A_{\min} := \min_{j\in[d]}\|A[j]\|$, we now write:

$$\begin{aligned}
\mathbb{E}[(D_\theta h_\theta(z_t))^\mathsf{T}(D_\theta h_\theta(z_t))] \\
&= \mathbb{E}[\mathbb{E}[M(\mu_{t-1}+w_{t-1})^\mathsf{T}\,\mathsf{diag}(\cos^2(M(\mu_{t-1}+w_{t-1})\theta))M(\mu_{t-1}+w_{t-1}) \mid \mu_{t-1}]] \\
&\succcurlyeq \mathbb{E}[\mathbb{E}[M(\mu_{t-1}+w_{t-1})^\mathsf{T}\,\mathsf{diag}(\cos^2(M(\mu_{t-1}+w_{t-1})\theta))M(\mu_{t-1}+w_{t-1})\mathbb{1}_{\mathcal{E}_{\mu_{t-1}}} \mid \mu_{t-1}]] \\
&\succcurlyeq \frac{\nu_0^2}{c_0^2 d^2}\min\{1,\sigma^2 A^2_{\min}\}\cdot\mathbb{E}[\mathbb{E}[((\mu_{t-1}+w_{t-1})^{\otimes 2}\otimes I_d)\mathbb{1}_{\mathcal{E}_{\mu_{t-1}}} \mid \mu_{t-1}]] \\
&= \frac{\nu_0^2}{c_0^2 d^2}\min\{1,\sigma^2 A^2_{\min}\}\cdot(\mathbb{E}[\mathbb{E}[(\mu_{t-1}+w_{t-1})^{\otimes 2}\mathbb{1}_{\mathcal{E}_{\mu_{t-1}}} \mid \mu_{t-1}]])\otimes I_d,
\end{aligned}$$

where the last equality follows from the bi-linearity of the Kronecker product. We now fix a $v \in \mathbb{R}^d$ with unit-norm, and observe, dropping subscripts on t for clarity,

$$\mathbb{E}_w[\langle v,\mu+w\rangle^2\mathbb{1}_{\mathcal{E}_\mu}] = \mathbb{E}_w[\langle v,\mu+w\rangle^2] - \mathbb{E}_w[\langle v,\mu+w\rangle^2\mathbb{1}_{\mathcal{E}_\mu^c}]$$

64

$$\geqslant \mathbb{E}_w[\langle v, \mu + w \rangle^2] - \sqrt{\mathbb{E}_w[\langle v, \mu + w \rangle^4]}\sqrt{\mathbb{P}(\mathbb{1}_{\mathcal{E}_\mu^c})}$$

$$\geqslant \mathbb{E}_w[\langle v, \mu + w \rangle^2] - \sqrt{\mathbb{E}_w[\langle v, \mu + w \rangle^4]}\sqrt{\nu_0}$$

$$\geqslant (1 - 9\sqrt{\nu_0})\mathbb{E}_w[\langle v, \mu + w \rangle^2]$$

$$= (1 - 9\sqrt{\nu_0})v^\mathsf{T}(\mu\mu^\mathsf{T} + \sigma^2 I_d)v$$

$$\geqslant (1 - 9\sqrt{\nu_0})\sigma^2,$$

where the last inequality holds by hyper-contractivity of Gaussian polynomials [see e.g., 86, Ch. 5]. Hence we set $\nu_0 = 1/324$, from which we conclude

$$\mathbb{E}\big[(\mu_{t-1} + w_{t-1})^{\otimes 2}\mathbb{1}_{\mathcal{E}_{\mu_{t-1}}}\big] \geqslant \sigma^2/2 \cdot I_d.$$

Consequently, we have that

$$\mathbb{E}[(D_\theta h_\theta(z_t))^\mathsf{T}(D_\theta h_\theta(z_t))] \geqslant \frac{\nu_0^2}{c_0^2 d^2}\min\{1, \sigma^2\}\sigma^2/2 \cdot I_{d^2} = \frac{c_1^2}{d^2}\min\{1, \sigma^2 A_{\min}^2\}(T-1)\sigma^2 \cdot I_{d^2}.$$

Therefore, we conclude that:

$$\mathcal{I}(\theta) = \frac{1}{\sigma^2}\sum_{t=1}^{T-1}\mathbb{E}_{z_t \sim p_\theta}[(D_\theta h_\theta(z_t))^\mathsf{T}(D_\theta h_\theta(z_t))] \geqslant \frac{c_1^2}{d^2}\min\{1, \sigma^2 A_{\min}^2\}(T-1) \cdot I_{d^2}.$$

Up until this point, we have not used the assumption that $\theta \in \Theta'$. Observe that $A_{\min} \geqslant A_{\star,\min}/2$ whenever $\theta \in \Theta'$, we finally conclude that for $\theta \in \Theta'$,

$$\mathcal{I}(\theta) = \frac{1}{\sigma^2}\sum_{t=1}^{T-1}\mathbb{E}_{z_t \sim p_\theta}[(D_\theta h_\theta(z_t))^\mathsf{T}(D_\theta h_\theta(z_t))] \geqslant \frac{c_1^2}{4d^2}\min\{1, \sigma^2 A_{\star,\min}^2\}(T-1) \cdot I_{d^2}. \qquad (4.56)$$

Step 2: Estimating B_1 and B_2. We will estimate B_1 and B_2 over Θ'. We start with B_1. First, given a trajectory $z_{1:T} \sim p_\theta$, we have that

$$\nabla_\theta \log p_\theta(z_{1:T}) = \sum_{t=1}^{T-1}\nabla_\theta \log p_\theta(z_{t+1} \mid z_t) = \frac{1}{\sigma^2}\sum_{t=1}^{T-1}(D_\theta h_\theta(z_t))^\mathsf{T}w_t.$$

Now fix a test vector $v \in \mathbb{R}^{d^2}$, and consider:

$$v^\mathsf{T}\nabla_\theta \log p_\theta(z_{1:T}) = \frac{1}{\sigma^2}\sum_{t=1}^{T-1}\langle D_\theta h_\theta(z_t)v, w_t \rangle =: \sum_{t=1}^{T-1}d_t.$$

A useful inequality is the following.

$$\|D_\theta h_\theta(z)\|_{\mathrm{op}} = \|\mathsf{diag}(\cos(M(z)\theta))M(z)\|_{\mathrm{op}} \leqslant \|M(z)\|_{\mathrm{op}} = \|z\|.$$

We first compute:

$$\mathbb{E}[d_t^2 \mid \mathcal{F}_{t-1}] = \mathbb{E}[\langle D_\theta h_\theta(z_t)v, w_t \rangle^2 \mid \mathcal{F}_{t-1}] = \sigma^2\|D_\theta h_\theta(z_t)v\|^2 \leqslant \sigma^2\|z_t\|^2\|v\|^2.$$

65

We next bound $\mathbb{E}[\|z_t\|^4]$ as:

$$\begin{aligned}
\mathbb{E}[\|z_t\|^4] &= \mathbb{E}[\|h_\theta(x_{t-1}) + w_{t-1}\|^4] \\
&\leqslant 8\mathbb{E}[\|h_\theta(x_{t-1})\|^4 + \|w_{t-1}\|^4] \leqslant 8(1 + \mathbb{E}[\|w_{t-1}\|^4]) \leqslant 8(1 + 3d^2\sigma^4).
\end{aligned}$$

Using this bound,

$$\begin{aligned}
\mathbb{E}\left(\sum_{t=1}^{T-1} \mathbb{E}[d_t^2 \mid \mathcal{F}_{t-1}]\right)^2 &\leqslant (T-1)\sum_{t=1}^{T-1}(\mathbb{E}[d_t^2 \mid \mathcal{F}_{t-1}])^2 \leqslant (T-1)\sigma^4\|v\|^4\sum_{t=1}^{T-1}\mathbb{E}[\|z_t\|^4] \\
&\leqslant (T-1)^2 8(1 + 3d^2\sigma^4)\sigma^4\|v\|^4.
\end{aligned}$$

Hence,

$$\left(\mathbb{E}\left(\sum_{t=1}^{T-1}\mathbb{E}[d_t^2 \mid \mathcal{F}_{t-1}]\right)^2\right)^{1/4} \lesssim \sigma(1 + \sigma\sqrt{d})\|v\|\sqrt{T}.$$

We next bound,

$$\mathbb{E}[d_t^4] \leqslant \mathbb{E}[\|D_\theta h_\theta(z_t)v\|^4\|w_t\|^4] \leqslant \|v\|^4\mathbb{E}[\|z_t\|^4\|w_t\|^4] \leqslant \|v\|^4\sqrt{\mathbb{E}[\|z_t\|^8]}\sqrt{\mathbb{E}[\|w_t\|^8]}.$$

Since $\|w_t\|$ is a σ-sub-Gaussian random variable, we know that $\mathbb{E}[\|w_t\|^8] \lesssim \sigma^8 d^4$. On the other hand,

$$\mathbb{E}[\|z_t\|^8] = \mathbb{E}[\|\mu_{t-1} + w_{t-1}\|^8] \leqslant 128(1 + \mathbb{E}[\|w_{t-1}\|^8]) \lesssim 1 + \sigma^8 d^4.$$

Hence we have

$$\sum_{t=1}^{T-1}\mathbb{E}[d_t^4] \lesssim \|v\|^4 T\sigma^4 d^2(1 + \sigma^4 d^2).$$

Therefore,

$$\left(\mathbb{E}\left(\sum_{t=1}^{T-1}d_t\right)^4\right)^{1/4} \lesssim \|v\|\sigma(1 + \sigma\sqrt{d})T^{1/2}(1 + T^{-1/4}\sqrt{d}) \leqslant \|v\|\sigma(1 + \sigma\sqrt{d})T^{1/2},$$

where the last inequality holds since we assume $T \gtrsim d^2$. Now we set $v = \mathcal{I}(\theta)^{-1/2}\bar{v}$ for a unit norm $\bar{v}$, we have that

$$\sup_{\theta_0,\theta_1\in\Theta'} B_1(\theta_0,\theta_1) \lesssim \frac{\sigma(1 + \sigma\sqrt{d})T^{1/2}}{\sqrt{\inf_{\theta\in\Theta'}\lambda_{\min}(\mathcal{I}(\theta))}} \lesssim \sigma d(1 + \sigma\sqrt{d})\max\{1, 1/(\sigma A_{\star,\min})\}.$$

We now move to B_2. First we define the vector-valued function for a fixed $q \in \mathbb{R}^d$:

$$g(\theta; z, q) = D_\theta h_\theta(z)^\mathsf{T} q = M^\mathsf{T}(z)\,\mathsf{diag}(\cos(M(z)\theta))q.$$

The Jacobian of $g(\theta; z, q)$ is given by:

$$D_\theta g(\theta; z, q) = -M^\mathsf{T}(z)\,\mathsf{diag}(h_\theta(z) \odot q)M(z),$$

66

where $\odot$ denotes the Hadamard (entry-wise) product. Now we have

$$
\begin{aligned}
-\sigma^2 \nabla_\theta^2 \log p_\theta(z' \mid z) &= (D_\theta h_\theta(z))^{\mathsf{T}} (D_\theta h_\theta(z)) + (D_\theta(D_\theta h_\theta(z))^{\mathsf{T}})(h_\theta(z) - z') \\
&= (D_\theta h_\theta(z))^{\mathsf{T}} (D_\theta h_\theta(z)) + D_\theta g(\theta; z, h_\theta(z) - z') \\
&= M^{\mathsf{T}}(z)[\mathsf{diag}(\cos^2(M(z)\theta)) - \mathsf{diag}(h_\theta(z) \odot (h_\theta(z) - z'))]M(z).
\end{aligned}
$$

Hence, given a test vector $v \in \mathbb{R}^{d^2}$, given $z_{1:T} \sim p_\theta$, for all $t \in [T-1]$ we have:

$$
-\sigma^2 v^{\mathsf{T}} \nabla_\theta^2 \log p_\theta(z_{t+1} \mid z_t) v = v^{\mathsf{T}} M^{\mathsf{T}}(z_t) \mathsf{diag}(\cos^2(M(z_t)\theta)) M(z_t) v + v^{\mathsf{T}} M^{\mathsf{T}}(z_t) \mathsf{diag}(h_\theta(z_t) \odot w_t) M(z_t) v.
$$

We next compute:

$$
\mathbb{E}[(v^{\mathsf{T}} M^{\mathsf{T}}(z_t) \mathsf{diag}(\cos^2(M(z_t)\theta)) M(z_t) v)^2] \leqslant \mathbb{E}[\|M(z_t)v\|^4] \leqslant \|v\|^4 \mathbb{E}[\|z_t\|^4] \lesssim \|v\|^4 (1 + \sigma^4 d^2).
$$

Next we compute

$$
\begin{aligned}
\mathbb{E}[(v^{\mathsf{T}} M^{\mathsf{T}}(z_t) \mathsf{diag}(h_\theta(z_t) \odot w_t) M(z_t) v)^2] &= \sigma^2 \mathbb{E}[\|M(z_t)v \odot \sqrt{h_\theta(z_t)}\|_4^4] \\
&\leqslant \sigma^2 \mathbb{E}[\|M(z_t)v\|_4^4] \leqslant \sigma^2 \mathbb{E}[\|M(z_t)v\|^4] \\
&\leqslant \sigma^2 \|v\|^4 \mathbb{E}[\|z_t\|^4] \lesssim \sigma^2 \|v\|^4 (1 + \sigma^4 d^2).
\end{aligned}
$$

Hence,

$$
(\mathbb{E}[(v^{\mathsf{T}} \nabla_\theta^2 \log p_\theta(z_{1:T})v)^2])^{1/2} \leqslant \sum_{t=1}^{T-1} (\mathbb{E}[v^{\mathsf{T}} \nabla_\theta^2 \log p_\theta(z_{t+1} \mid z_t)v])^{1/2} \lesssim \frac{\|v\|^2}{\sigma^2} T(1+\sigma)(1+\sigma^2 d).
$$

Now we set $v = \mathcal{I}(\theta)^{-1/2} \bar{v}$ for a unit norm $\bar{v}$, we have that

$$
\begin{aligned}
\sup_{\theta_0, \theta_1 \in \Theta'} B_2(\theta_0, \theta_1) &\lesssim \frac{T(1+\sigma)(1+\sigma^2 d)/\sigma^2}{\inf_{\theta \in \Theta'} \lambda_{\min}(\mathcal{I}(\theta))} \\
&\lesssim d^2 \max\left\{1, \frac{1}{\sigma^2 A_{\star,\min}^2}\right\} \frac{(1+\sigma)(1+\sigma^2 d)}{\sigma^2} \\
&\lesssim d^3 \max\left\{\sigma, 1 + \frac{1}{\sigma^2 d}\right\} \max\left\{1, \frac{1}{\sigma^2 A_{\star,\min}^2}\right\}.
\end{aligned}
$$

Finally, we conclude that

$$
\sup_{\theta_0, \theta_1 \in \Theta'} \max\{B_1^2(\theta_0, \theta_1), B_2(\theta_0, \theta_1)\} \lesssim d^3 \max\left\{1, \sigma^4, \frac{1}{\sigma^2 d}\right\} \max\left\{1, \frac{1}{\sigma^2 A_{\star,\min}^2}\right\}. \tag{4.57}
$$

Step 3: Parameter error bound. Utilizing (4.54) and (4.57), we have that to verify the condition (3.22) for $\theta_0 = \theta_\star$ and $\theta_1 = \hat{\theta}_{m,T}^\varepsilon$ on $\mathcal{E}_1$, we need m to satisfy:

$$
m \gtrsim d^{10} T \cdot \max\left\{\frac{1}{d}, \frac{1}{\sigma^2}, \frac{1}{\sigma^4}\right\} \max\left\{1, \sigma^8, \frac{1}{\sigma^4 d^2}\right\} \max\left\{1, \frac{1}{\sigma^4 A_{\star,\min}^4}\right\} \cdot \log\left(\frac{c_0 RmT(d+\sigma^2)}{\sigma^2 \delta}\right).
$$

Defining

$$\Phi_2 := \max\left\{\frac{1}{d}, \frac{1}{\sigma^2}, \frac{1}{\sigma^4}\right\} \max\left\{1, \sigma^8, \frac{1}{\sigma^4 d^2}\right\} \max\left\{1, \frac{1}{\sigma^4 A_{\star,\min}^4}\right\},$$

by Proposition A.1 it suffices for m to satisfy:

$$m \gtrsim \Phi_2 d^{10} T \log\left(\frac{c_2 \Phi_2 d R T (d + \sigma^2)}{\sigma^2 \delta}\right). \tag{4.58}$$

From (4.56), we have that on Θ', the following lower bound $\inf_{\theta \in \Theta'} \lambda_{\min}(\mathcal{I}(\theta)) \gtrsim \frac{T}{d^2} \cdot \min\{1, \sigma^2 A_{\star,\min}^2\}$ holds. Since (a) (3.22) holds for $\theta_0 = \theta_\star$ and $\theta_1 = \hat{\theta}_{m,T}^\varepsilon$ implies (3.22) also holds for $\theta_0 = \theta_\star$ and any $\theta_1 \in \mathrm{conv}\{\hat{\theta}_{m,T}^\varepsilon, \theta_\star\}$ and (b) $\hat{\theta}_{m,T}^\varepsilon \in \Theta'$ implies $\mathrm{conv}\{\hat{\theta}_{m,T}^\varepsilon, \theta_\star\} \subset \Theta'$, by (3.23) from Proposition 3.9, we have on $\mathcal{E}_1$, for every $\theta \in \mathrm{conv}\{\hat{\theta}_{m,T}^\varepsilon, \theta_\star\}$,

$$\|\theta - \theta_\star\|^2 \lesssim \frac{d^2 \max\{1, 1/(\sigma^2 A_{\star,\min}^2)\}}{T} \mathrm{d}_H^2(\theta, \theta_\star),$$
$$\lesssim \max\{1, 1/(\sigma^2 A_{\star,\min}^2)\} \mathrm{d}_H^2(\theta, \theta_\star), \tag{4.59}$$

where the last inequality holds since we assume that $T \gtrsim d^2$.

Step 4: Verify FI radius. We will utilize Proposition A.4 to verify the FI radius condition (3.24). Since $\nabla_\theta \log p_\theta(z_{t+1} \mid z_t) = -\frac{1}{\sigma^2}(D_\theta h_\theta(z_t))^\mathsf{T}(h_\theta(z_t) - z_{t+1})$ for $t \geq 1$, we have for any $\theta \in \Theta$,

$$\mathcal{I}_{t+1}(\theta \mid z_{1:t}) = \frac{1}{\sigma^2}(D_\theta h_\theta(z_t))^\mathsf{T}(D_\theta h_\theta(z_t)), \quad t \in \{1, \dots, T-1\},$$
$$= \frac{1}{\sigma^2} M(z_t)^\mathsf{T} \mathsf{diag}(\cos^2(M(z_t)\theta)) M(z_t).$$

We also have the base case $\mathcal{I}_1(\theta) = 0$. Hence for $t \geq 1$, $\theta_1, \theta_2 \in \Theta$, and any unit-norm $v \in \mathbb{R}^{d^2}$,

$$|v^\mathsf{T}(\mathcal{I}_{t+1}(\theta_1 \mid z_{1:t}) - \mathcal{I}_{t+1}(\theta_2 \mid z_{1:t}))v| = \frac{1}{\sigma^2}|v^\mathsf{T} M(z_t)^\mathsf{T} \mathsf{diag}(\cos^2(M(z_t)\theta_1) - \cos^2(M(z_t)\theta_2)) M(z_t) v|$$
$$\leq \frac{2}{\sigma^2}\|M(z_t)v\|^2 \|M(z_t)(\theta_1 - \theta_2)\|$$
$$\leq \frac{2}{\sigma^2}\|M(z_t)\|_{\mathrm{op}}^3 \|\theta_1 - \theta_2\|$$
$$\leq \frac{2}{\sigma^2}\|z_t\|^3 \|\theta_1 - \theta_2\|.$$

Therefore by the variational form of the operator norm, we have for all $t \geq 1$ and $\theta_1, \theta_2 \in \Theta$:

$$\|\mathcal{I}_{t+1}(\theta_1 \mid z_{1:t}) - \mathcal{I}_{t+1}(\theta_2 \mid z_{1:t})\|_{\mathrm{op}} \leq \mathrm{Lip}_{t+1}(z_{1:t})\|\theta_1 - \theta_2\|, \quad \mathrm{Lip}_{t+1}(z_{1:t}) = \frac{2}{\sigma^2}\|z_t\|^3.$$

Note that for any $\theta \in \Theta$, $\mathbb{E}_{p_\theta}[\mathrm{Lip}_{t+1}(z_{1:t})] \lesssim \frac{1}{\sigma^2}(1 + \sigma^3 d^{3/2}) \asymp \max\{1/\sigma^2, \sigma d^{3/2}\}$, and consequently $\mathrm{Lip} \lesssim \max\{1/\sigma^2, \sigma d^{3/2}\}$. On the other hand, for a unit-norm $v \in \mathbb{R}^{d^2}$ and $\theta_1, \theta_2 \in \Theta$,

$$\mathbb{E}_{p_{\theta_1}}[(v^\mathsf{T} \mathcal{I}_{t+1}(\theta_2 \mid z_{1:t})v)^2] = \frac{1}{\sigma^4}\mathbb{E}_{p_{\theta_1}}[(v^\mathsf{T} M(z_t)^\mathsf{T} \mathsf{diag}(\cos^2(M(z_t)\theta_2)) M(z_t) v)^2]$$

$$\leqslant \frac{1}{\sigma^4}\mathbb{E}_{p_{\theta_1}}[\|M(z_t)v\|^4] \leqslant \frac{1}{\sigma^4}\mathbb{E}_{p_{\theta_1}}[\|z_t\|^4] \lesssim \frac{1}{\sigma^4}(1+\sigma^4 d^2) \asymp \max\{1/\sigma^4, d^2\}.$$

Hence, we have that $B_{\mathcal{I}} \lesssim \max\{1/\sigma^2, d\}$. By Proposition A.4 and (4.59) we have on $\mathcal{E}_1$ for any $\theta \in \text{conv}\{\hat{\theta}_{m,T}^\varepsilon, \theta_\star\}$,

$$\|\mathcal{I}(\theta) - \mathcal{I}(\theta_\star)\|_{\text{op}} \lesssim T\Big[\max\{1/\sigma^2, \sigma d^{3/2}\}\|\theta - \theta_\star\| + \max\{1/\sigma^2, d\}\mathrm{d}_H(\theta, \theta_\star)\Big]$$

$$\lesssim T\Big[\max\{1/\sigma^2, \sigma d^{3/2}\}\max\{1, 1/(\sigma A_{\star,\min})\} + \max\{1/\sigma^2, d\}\Big]\mathrm{d}_H(\theta, \theta_\star)$$

$$\lesssim T\max\{1/\sigma^2, \sigma d^{3/2}, d\}\max\{1, 1/(\sigma A_{\star,\min})\}\mathrm{d}_H(\theta, \theta_\star). \tag{4.60}$$

Let us define

$$\Phi_3 := \max\left\{\frac{1}{\sigma^4 d^3}, \sigma^2, \frac{1}{d}\right\}\max\left\{1, \frac{1}{\sigma^6 A_{\star,\min}^6}\right\}\max\left\{\frac{1}{d}, \frac{1}{\sigma^2}, \frac{1}{\sigma^4}\right\}.$$

Combining (4.54), (4.56), and (4.60), we have that on $\mathcal{E}_1$,

$$\sup_{\theta \in \text{conv}\{\hat{\theta}_{m,T}^\varepsilon, \theta_\star\}} \frac{\|\mathcal{I}(\theta) - \mathcal{I}(\theta_\star)\|_{\text{op}}}{\lambda_{\min}(\mathcal{I}(\theta_\star))}$$

$$\lesssim \sup_{\theta \in \text{conv}\{\hat{\theta}_{m,T}^\varepsilon, \theta_\star\}} d^2 \max\{1/\sigma^2, \sigma d^{3/2}, d\}\max\{1, 1/(\sigma^3 A_{\star,\min}^3)\}\mathrm{d}_H(\theta, \theta_\star)$$

$$= \sup_{\theta \in \text{conv}\{\hat{\theta}_{m,T}^\varepsilon, \theta_\star\}} d^{7/2} \max\{1/(\sigma^2 d^{3/2}), \sigma, 1/d^{1/2}\}\max\{1, 1/(\sigma^3 A_{\star,\min}^3)\}\mathrm{d}_H(\theta, \theta_\star)$$

$$\lesssim \sqrt{\frac{\Phi_3 d^{11} T}{m}\log\left(\frac{c_0 RmT(d+\sigma^2)}{\sigma^2 \delta}\right)}.$$

Hence by Proposition A.1, we have that (3.24) holds on $\mathcal{E}_1$ as long as m satisfies:

$$m \gtrsim \Phi_3 d^{11} T \log\left(\frac{c_3 \Phi_3 dRT(d+\sigma^2)}{\sigma^2 \delta}\right). \tag{4.61}$$

Step 5: Final result. If m satisfies conditions (4.55), (4.58), and (4.61), then on $\mathcal{E}_1$ we have from Proposition 3.9 and (4.52):

$$\|\hat{\theta}_{m,T}^\varepsilon - \theta_\star\|_{\mathcal{I}(\theta_\star)}^2 \lesssim \frac{d^2}{mT}\log\left(\frac{c_0 RmT(d+\sigma^2)}{\sigma^2 \delta}\right).$$

4.4 Sequence Modeling with Linear Attention

Since their introduction, transformer models and architectures have found popularity in modern sequence modeling tasks, finding use in fields such as language modeling, computer vision, and reinforcement learning [87–89]. Despite their widespread application however, full theoretical analysis of multi-layer transformer models is currently out of reach. As a result, stylized and simplified attention modules that isolate core mechanisms are commonly used in the literature as analytically tractable proxies for analyzing the full models. Single-layer linear self-attention models have been

used to explore the dynamics of in-context learning [90, 91] and emergent inductive biases in transformers [92, 93]. Recent works show that attention can operate as a max-margin token selection mechanism, even establishing an equivalence with hard-margin SVM [94, 95]. Furthermore, [96] established the global convergence of gradient descent for this framework, and a finite sample bounds were established by [97] with parameter estimation upper bounds; we take particular inspiration from this line of work, especially [97], and analyze a simple linear transformer [20] with a single-layer cross-attention and linear activation.

Let us consider a vocabulary of K tokens denoting the states $z \in \mathsf{Z}$, each assigned a d-dimensional embedding by an embedding matrix $E = [e_1 \cdots e_K]^\mathsf{T} \in \mathbb{R}^{K \times d}$, such that $\mathsf{Z} = \{e_i \mid i \in [K]\}$. We further emphasize that these embeddings $\{e_k\}_{k=1}^K$ are *not* the standard basis vectors, which we denote instead $\mathbf{1}(k) \in \mathbb{R}^K$ in this section. We assume that the first *two* tokens z_0, z_1 are drawn from a given initial distribution ρ_1 over $\mathsf{Z} \times \mathsf{Z}$ to create an initial state for cross-attention. To sample a new token $z_{t+1} \in \mathsf{Z}$ in a sequence $z_{0:t}$ where $z_{0:t} \in \mathbb{R}^{(t+1) \times d}$, we sample *auto-regressively* by taking the last token z_t to be the query token in a cross-attention layer given as:

$$p_\theta(z_{t+1} \mid z_{0:t}) = \mathbb{S}\big(\Phi(C(z_{0:t-1}V)^\mathsf{T}\mathbb{A}(z_{0:t-1}\mathcal{K}Q^\mathsf{T}z_t))\big), \quad t \in [T-1],$$

where $\mathcal{K}, Q, V \in \mathbb{R}^{d \times d}$ denote the key, query, and value matrices, $C \in \mathbb{R}^{(K-1) \times d}$ denotes the classifier head, $\mathbb{A}$ denotes an activation function, $\mathbb{S}$ denotes the softmax function, and $\Phi : \mathbb{R}^{K-1} \mapsto \mathbb{R}^K$ denotes the function that embeds $\Phi(x) := (x, 0)$.[13] We shall denote $\theta := \mathrm{vec}(\mathcal{K}Q^\mathsf{T})$ to be the parametrization such that $\Theta = \{\theta \in \mathbb{R}^{d^2} \mid \|\theta\| \leqslant R\}$, and we constrain $V = I_d$ and C to be fixed matrices. Finally, we use a linear activation [20], where we normalize by $1/t$ for key trajectory length t.[14] This all simplifies to the following:

$$p_\theta(z_{t+1} \mid z_{0:t}) = \mathbb{S}\left(\Phi\left(\frac{1}{t}Cz_{0:t-1}^\mathsf{T}z_{0:t-1}\mathrm{mat}(\theta)z_t\right)\right), \quad t \in [T-1], \tag{4.62}$$

where we recall that mat is the matricization function such that $\mathrm{mat}(\theta) \in \mathbb{R}^{d \times d}$.

Theorem 4.18. *Fix $\delta \in (0,1)$, and suppose that the embedding matrix E and classifier head C are both full column rank, with normalized embeddings such that $\max_{k \in [K]}\|e_k\| = 1$ and classifier head such that $\|C\|_{\mathrm{op}} \geqslant 1$. Let $\Theta = \{\theta \in \mathbb{R}^{d^2} \mid \|\theta\| \leqslant R\}$ for $R \geqslant 1$ and $d > 1$, and let $\hat{\theta}_{m,T}^\varepsilon$ denote the max FI discretized MLE estimator at resolution $\varepsilon = \frac{\delta}{2\sqrt{2m}}$. If the number of trajectories m and the trajectory length T satisfies*

$$m \gtrsim d^2 T_0^2 \log\left(\frac{c_0 d^2 T_0^2 R^2 \|C\|_{\mathrm{op}}^2 T}{\delta^2}\right), \quad T \gtrsim T_0, \quad T_0 := \frac{K^4 \kappa(C)^2}{\sigma_{\min}(E)^4} \exp(6R\|C\|_{\mathrm{op}}), \tag{4.63}$$

where $\kappa(C)$ denotes the condition number of C, then with probability at least $1 - \delta$ for $\delta \in (0,1)$ and for a universal positive constant c_0, we have

$$\|\hat{\theta}_{m,T}^\varepsilon - \theta_\star\|_{\bar{\mathcal{I}}(\theta_\star)}^2 \lesssim \frac{d^2 \log(c_0 R^2 \|C\|_{\mathrm{op}}^2 mT/\delta^2)}{mT}. \tag{4.64}$$

---

[13] The function Φ is introduced to allow for parameter recovery; otherwise parameterizing a distribution over $[K]$ using K logits is not identifiable, as softmax is invariant under affine transforms (i.e., $\mathbb{S}(x + c\mathbf{1}) = \mathbb{S}(x)$ for any $c \in \mathbb{R}$).

[14] This normalization ensures that the sum of key-query attention scores does not scale with trajectory length, which would increase the magnitude inside the outer softmax over time and cause exponential decay in the minimum probability of a token. This scaling is implicit in softmax activation settings, which would divide by the sum of the exponentiated weights. We note that this scaling is also used in practice in [98].

To the best of our knowledge, Theorem 4.18 is the first result for learning the parameters of an auto-regression linear transformer model in the multiple trajectory setting which achieves a nearly instance-optimal rate (cf. (4.64)), which also includes a rate of convergence that decreases with all the data mT instead of just the number of trajectories m. The most related result to Theorem 4.18 comes from [97, Corollary 4.3], which we will compare with in detail in a moment. Before discussion this related result, we make a few remarks on Theorem 4.18. First, we know that the assumption that both E and C are full column rank implies the constraint $d + 1 \leqslant K$, which states that the embedding size d is *less* than the vocabulary size K minus one. This is a realistic assumption in practice, as typical vocabulary sizes for modern LLMs are often in the 100k range, whereas typical embedding sizes are typically no more than 10k.[15] We next focus on the trajectory requirement on m in (4.63). We first note that the constraint on trajectory length $T \geqslant T_0$ in (4.63) can be elided at the expense of a more complex expression for the required number of trajectories, but the final rate would remain the same. Next, the requirement on m, ignoring the contributions of C, E, is $m \gtrsim \tilde{\Omega}(d^2 K^8 \exp(R))$. While the dependence on d is correct, we anticipate that the dependence on K, R is not sharp, and can be improved with further analysis. Similarly, in our analysis we show a bound of the form $\lambda_{\min}(\bar{\mathcal{I}}(\theta_\star)) \gtrsim K^{-4} \exp(-R)$ (also ignoring contributions of C, E), which implies from (4.64) a parameter recovery bound of $\|\hat{\theta}_{m,T}^\varepsilon - \theta_\star\|^2 \lesssim \tilde{O}(d^2 K^4 \exp(R)/(mT))$; as with the requirement on m, we anticipate this parameter recovery bound is also not optimal in its dependence on K, R (but is optimal in d, m, T).

Comparison with [97, Corollary 4.3].   As mentioned previously, the most comparable result to Theorem 4.18 is [97, Corollary 4.3]. Here, the authors also study a multi-trajectory data model, but one key difference is that in [97], the trajectory $z_{0:T}$ is not auto-regressively generated. Instead, there is a distribution $\mathcal{D}_X$ over *prompts* $z_{0:T-1}$, followed by a last token z_T generated from a self-attention model conditioned on the prompt $z_{0:T-1}$, resembling a standard supervised learning setup. Consequently, their final parameter recovery rate only decays with the number of trajectory m, in comparison to our rate (4.64) which decreases with the total data budget mT. Another difference between our two settings is a structural one. We choose to analyze a setting with linear activation instead of softmax activation $\mathbb{A}$, and we constrain the outputs of the classifier head C to Δ^{K-1} (the probability simplex in $\mathbb{R}^K$) by means of an outer softmax activation (i.e., we treat the outputs of the classifier head as logits, as is typically done in practice), as detailed in (4.62). On the other hand in [97] they consider a softmax activation $\mathbb{A}$, but omit the softmax activation after the classifier head. Hence, they require additional assumptions on the classifier head and the embeddings matrix to ensure that the output of the classifier head is a valid probability distribution. This may seem like a minor difference, but their setup requires that vocabulary embeddings E are linearly independent [97, Assumption 2.3], which requires that $d \geqslant K$ (i.e., the embedding dimension exceeds the vocabulary size). As we discussed previously, in practice we typically have the opposite trend (i.e., embedding dimension is much smaller than vocabulary size), which our model allows for.

With these remarks in place, we can now directly compare our bounds to [97], keeping in mind the differences in problem setup and assumptions described previously. To keep the comparison simple, we will suppress dependency on C, R, E, and only focus on m, T, K in the bounds. The main parameter recovery result in [97] states that with high probability:

$$m \gtrsim \tilde{\Omega}\left(\frac{K^2}{\alpha^2}\right) \implies \|\hat{\theta}_{m,T} - \theta_\star\|^2 \lesssim \tilde{O}\left(\frac{K^2}{\alpha^4 m}\right), \tag{4.65}$$

---

[15]For example, Meta's Llama3 8B model has a vocabulary size of 128k with an embedding dimension of 4096 [99].

where $\alpha > 0$ is the strong convexity constant of the population loss over a ball around $\theta_\star$. This quantity α is left unspecified in their argument; they only argue that $\alpha > 0$, but do not provide an explicit lower bound for it. Since for negative log likelihood, both Fisher information and Hessian of the population loss coincide, in the notation of our work, $\alpha = \inf_{\theta \in B(\theta_\star, r_0)} \lambda_{\min}(\mathbb{E}_{z_{0:T-1} \sim \mathcal{D}_X}[\mathcal{I}_T(\theta \mid z_{0:T-1})])$ (note that the conditional Fisher Information notation is defined in (A.1)) where r_0 is a localization parameter which we consider as a constant. With this in mind, Theorem 4.18 implies that with high probability:

$$m \gtrsim \tilde{\Omega}\left(\frac{d^2}{\bar{\alpha}^2}\right), \; T \gtrsim \frac{1}{\bar{\alpha}} \implies \|\hat{\theta}_{m,T}^{\varepsilon} - \theta_\star\|^2 \lesssim \tilde{O}\left(\frac{d^2}{\bar{\alpha}mT}\right), \tag{4.66}$$

where $\bar{\alpha} := \lambda_{\min}(\bar{\mathcal{I}}(\theta_\star))$. As mentioned previously, we show a lower bound on $\bar{\alpha} \gtrsim K^{-4}$, which again is most likely not optimal. Comparing (4.65) with (4.66), we see that the dependence on the parameter dimension (K for the former as they consider a subspace of $d \times d$ matrices with dimension $\leqslant K^2$, d for our case) is equivalent up to log factors. On the other hand, our bound yields an improvement on the dependence of the $\bar{\alpha}$ (vs. α) parameter in the final rate. Finally and most importantly, as our setting studies auto-regressive generation, our rate is able to capture the dependence on all the data points mT, rather than just the number of trajectories m.

4.4.1 Proof of Theorem 4.18

Step 1: Covering number bound. As earlier, we first estimate the covering number of $\mathcal{P}$ in the FI norm through the ℓ_2 covering number. Let $J := \begin{bmatrix} I_{K-1} \\ 0 \end{bmatrix} \in \mathbb{R}^{K \times (K-1)}$. If we denote $M_{0:t} := \frac{1}{t} z_t^\mathsf{T} \otimes (JC z_{0:t-1}^\mathsf{T} z_{0:t-1}) \in \mathbb{R}^{K \times d^2}$ and $\bar{M}_{0:t} := \frac{1}{t} z_t^\mathsf{T} \otimes (C z_{0:t-1}^\mathsf{T} z_{0:t-1}) \in \mathbb{R}^{(K-1) \times d}$, we can see that the following holds by vectorization:

$$\Phi\left(\frac{1}{t} C z_{0:t-1}^\mathsf{T} z_{0:t-1} \mathrm{mat}(\theta) z_t\right) = \Phi(\bar{M}_{0:t}\theta) = J\bar{M}_{0:t}\theta = M_{0:t}\theta$$

As such we can say $p_\theta(z_{t+1} \mid z_{0:t}) = \mathbb{S}(M_{0:t}\theta)_k$ for $z_{t+1} = e_k$. Next, since we have

$$\mathrm{d}_{\mathcal{I}_{\max}}(p_{\theta_0}, p_{\theta_1}) = \|\theta_0 - \theta_1\|_{\mathcal{I}_{\max}} \leqslant \sqrt{\lambda_{\max}(\mathcal{I}_{\max})}\|\theta_0 - \theta_1\|,$$

this motivates finding an expression for the Fisher information matrix and its maximum eigenvalue. If we let $(z_t) \in [K]$ denote the index of the token associated with entry z_t and $[M_{0:t}]_i$ denote the i-th row of $M_{0:t}$, we calculate the Hessian of the log likelihood:

$$\nabla_\theta \log p_\theta(z_{t+1} \mid z_{0:t}) = [M_{0:t}]_{(z_{t+1})} - M_{0:t}^\mathsf{T} \mathbb{S}(M_{0:t}\theta),$$
$$\nabla_\theta^2 \log p_\theta(z_{t+1} \mid z_{0:t}) = M_{0:t}^\mathsf{T} \left(\mathbb{S}(M_{0:t}\theta)\mathbb{S}(M_{0:t}\theta)^\mathsf{T} - \mathsf{diag}(\mathbb{S}(M_{0:t}\theta))\right) M_{0:t}.$$

If we denote the conditional expectation $\mathbb{E}_t^\theta[\cdot] \triangleq \mathbb{E}_{p_\theta}[\cdot \mid z_{0:t}]$,

$$\mathbb{E}_t^\theta[\nabla_\theta \log p_\theta(z_{t+1} \mid z_{0:t})] = 0,$$
$$\mathbb{E}_t^\theta[\nabla_\theta^2 \log p_\theta(z_{t+1} \mid z_{0:t})] = M_{0:t}^\mathsf{T} \left(\mathbb{S}(M_{0:t}\theta)\mathbb{S}(M_{0:t}\theta)^\mathsf{T} - \mathsf{diag}(\mathbb{S}(M_{0:t}\theta))\right) M_{0:t}.$$

Hence, the FI matrix can be represented as:

$$\mathcal{I}_{t+1}(\theta \mid z_{0:t}) = -\mathbb{E}_t^\theta[\nabla_\theta^2 \log p_\theta(z_{t+1} \mid z_{0:t})] = M_{0:t}^\mathsf{T} \left(\mathsf{diag}(\mathbb{S}(M_{0:t}\theta)) - \mathbb{S}(M_{0:t}\theta)\mathbb{S}(M_{0:t}\theta)^\mathsf{T}\right) M_{0:t}, \tag{4.67}$$

$$\mathcal{I}(\theta) = \sum_{t=1}^{T-1} \mathbb{E}_{z_{0:t} \sim p_\theta} \left[M_{0:t}^\mathsf{T} \left(\mathrm{diag}(\mathbb{S}(M_{0:t}\theta)) - \mathbb{S}(M_{0:t}\theta)\mathbb{S}(M_{0:t}\theta)^\mathsf{T} \right) M_{0:t} \right]. \tag{4.68}$$

We may see that for all $t \in [T-1]$, $\|M_{0:t}\|_{\mathrm{op}} \leqslant \sup_{k \in [K]} \|e_k\|^3 \|C\|_{\mathrm{op}} = \|C\|_{\mathrm{op}}$. Expanding out the multiplication and upper bounding gives our result.

$$\begin{aligned}
\mathcal{I}(\theta) &= \sum_{t=1}^{T-1} \mathbb{E}_{z_{0:t} \sim p_\theta} \left[M_{0:t}^\mathsf{T} \left(\mathrm{diag}(\mathbb{S}(M_{0:t}\theta)) - \mathbb{S}(M_{0:t}\theta)\mathbb{S}(M_{0:t}\theta)^\mathsf{T} \right) M_{0:t} \right] \\
&= \sum_{t=1}^{T-1} \mathbb{E}_{z_{0:t} \sim p_\theta} \left[M_{0:t}^\mathsf{T} \, \mathrm{diag}(\mathbb{S}(M_{0:t}\theta)) M_{0:t} - M_{0:t}^\mathsf{T} \mathbb{S}(M_{0:t}\theta)\mathbb{S}(M_{0:t}\theta)^\mathsf{T} M_{0:t} \right] \\
&\preccurlyeq \sum_{t=1}^{T-1} \mathbb{E}_{z_{0:t} \sim p_\theta} \left[M_{0:t}^\mathsf{T} \, \mathrm{diag}(\mathbb{S}(M_{0:t}\theta)) M_{0:t} \right] \\
&\preccurlyeq T \|C\|_{\mathrm{op}}^2 I_{d^2}.
\end{aligned}$$

This gives that for all $\theta \in \Theta$, $\lambda_{\max}(\mathcal{I}(\theta)) \leqslant T\|C\|_{\mathrm{op}}^2$: notably, this expression is agnostic of the exact parametrization. This allows us to set $\mathcal{I}_{\max} = T\|C\|_{\mathrm{op}}^2 I_{d^2}$, from which we have $\lambda_{\max}(\mathcal{I}_{\max}) = T\|C\|_{\mathrm{op}}^2$. If we fix some $\theta \in \Theta$ and let $\hat{\theta}$ denote its closest element in an ε-covering of Θ in ℓ_2, substituting $\mathcal{I}_{\max}$ into the previous bound on $\mathrm{d}_{\mathcal{I}_{\max}}(p_\theta, p_{\hat{\theta}})$ gives $\mathrm{d}_{\mathcal{I}_{\max}}(p_\theta, p_{\hat{\theta}}) \leqslant \sqrt{T}\|C\|_{\mathrm{op}}\varepsilon$. This implies the relation for $\varepsilon \in (0,1)$:

$$\mathcal{N}_{\mathcal{I}_{\max}}(\mathcal{P}, \varepsilon) \leqslant \mathcal{N}_{\|\cdot\|}\left(\Theta, \frac{\varepsilon}{\sqrt{T}\|C\|_{\mathrm{op}}} \right) \leqslant \left(3R \frac{\sqrt{T}\|C\|_{\mathrm{op}}}{\varepsilon} \right)^{d^2}.$$

Applying Theorem 3.6 with $\varepsilon = \frac{\delta}{2\sqrt{2m}}$ and $\eta = \frac{1}{2R\|C\|_{\mathrm{op}}\sqrt{mT}}$, and upper bounding $R\|C\|_{\mathrm{op}}\sqrt{mT}/\delta \leqslant (R\|C\|_{\mathrm{op}}\sqrt{mT}/\delta)^{d^2}$ in the logarithm, for some universal constant $c_0 > 0$ we obtain with probability at least $1 - \delta$,

$$\begin{aligned}
\sup_{\theta \in \mathrm{conv}\{\hat{\theta}_{m,T}^\varepsilon, \theta_\star\}} \mathrm{d}_H^2(\theta, \theta_\star) &\leqslant \inf_{\eta > 0} \left\{ \frac{6}{m} \log\left(\frac{2\mathcal{N}_{\mathcal{I}_{\max}}(\mathcal{P}, \varepsilon)}{\delta} \left\lceil \frac{1}{2\eta} \right\rceil \right) + \frac{3\eta^2}{4} \mathrm{diam}^2(\Theta) + 3\varepsilon^2 \right\} \\
&\lesssim \frac{1}{m} \log\left(\frac{c_0 \mathcal{N}_{\mathcal{I}_{\max}}(\mathcal{P}, \varepsilon) \left\lceil R\|C\|_{\mathrm{op}}\sqrt{mT} \right\rceil}{\delta} \right) + \frac{\mathrm{diam}^2(\Theta)}{R^2\|C\|_{\mathrm{op}}^2 mT} + \varepsilon^2 \\
&\lesssim \frac{d^2}{m} \log\left(\frac{c_0 R^2 \|C\|_{\mathrm{op}}^2 mT}{\delta^2} \right) + \frac{1 + \delta^2}{m}.
\end{aligned}$$

For satisfactory c_0, the final $\frac{1+\delta^2}{m}$ term can additionally be collapsed into the first term given that $d > 1$. Let us denote $\mathbf{1}(z) \in \{0,1\}^K$ to be the one-hot standard basis vector such that for $i \in [K]$, $[\mathbf{1}(z)]_i = 1$ if $z = e_i$ and $[\mathbf{1}(z)]_i = 0$ otherwise. When invoking Theorem 3.6, we can use the log-concave conclusion (3.16), since $\log p_\theta(z_{0:T})$ is given by:

$$\begin{aligned}
\log p_\theta(z_{0:T}) &= \log \rho_1(z_0, z_1) + \sum_{t=1}^{T-1} \log p_\theta(z_{t+1} \mid z_{0:t}) \\
&= \log \rho_1(z_0, z_1) + \sum_{t=1}^{T-1} \log(\langle \mathbb{S}(M_{0:t}\theta), \mathbf{1}(z_{t+1}) \rangle)
\end{aligned}$$

$$= \log \rho_1(z_0, z_1) + \sum_{t=1}^{T-1} \langle M_{0:t}\theta, \mathbf{1}(z_{t+1}) \rangle - \mathrm{LSE}(M_{0:t}\theta),$$

which is the sum of affine terms in θ minus the sum of terms which are given by taking the log-sum-exp (LSE) of a linear function of θ, which is convex [cf. 100, Section 3.1.5]. Hence, $\log p_\theta(z_{0:T})$ is a concave function by basic composition rules.

Step 2: Estimating B_1 and B_2. Crucial in bounding B_1 and B_2 is bounding the smallest eigenvalue of $\mathcal{I}(\theta_\star)$, so let us note the expansion of the Fisher information from eq. (4.68):

$$\mathcal{I}(\theta) = \sum_{t=1}^{T-1} \mathbb{E}_{z_{0:t} \sim p_\theta} \left[\bar{M}_{0:t}^\mathsf{T} J^\mathsf{T} \left(\mathsf{diag}(\mathbb{S}(M_{0:t}\theta)) - \mathbb{S}(M_{0:t}\theta)\mathbb{S}(M_{0:t}\theta)^\mathsf{T} \right) J \bar{M}_{0:t} \right].$$

We note that if one simply looks at $\mathsf{diag}(\mathbb{S}(M_{0:t}\theta)) - \mathbb{S}(M_{0:t}\theta)\mathbb{S}(M_{0:t}\theta)^\mathsf{T}$, the minimum eigenvalue would be zero since there is a constrained direction due to $\mathbb{S}(M_{0:t}\theta)$ lying on the $K-1$ dimensional simplex. Hence, we look at the reduced matrix by ignoring the last redundant coordinate, which restores a non-zero minimum eigenvalue. As such, we begin by examining the inner expression $J^\mathsf{T}(\mathsf{diag}(\mathbb{S}(M_{0:t}\theta)) - \mathbb{S}(M_{0:t}\theta)\mathbb{S}(M_{0:t}\theta)^\mathsf{T})J$. Proposition E.1 states that if we can show there exists $\mu > 0$ such that $\forall k \in [K]$, $\mathbb{S}(M_{0:t}\theta)_k \geqslant \mu$, then $\lambda_{\min}\left(J^\mathsf{T}(\mathsf{diag}(\mathbb{S}(M_{0:t}\theta)) - \mathbb{S}(M_{0:t}\theta)\mathbb{S}(M_{0:t}\theta)^\mathsf{T})J\right) \geqslant \frac{\mu}{4(K-1)}$. Let us now find such a μ by lower bounding $\log p_\theta(e_k \mid z_{0:t})$ and exponentiating it to find a bound for $p_\theta(e_k \mid z_{0:t})$ for arbitrary $k \in [K]$:

$$\log p_\theta(e_k \mid z_{0:t}) = [M_{0:t}]_k^\mathsf{T} \theta - \log \sum_{i=1}^K \exp\left([M_{0:t}]_i^\mathsf{T} \theta\right)$$

$$\geqslant \min_{i,j \in [K]} [M_{0:t}]_i^\mathsf{T} \theta - \log\left(K \exp\left([M_{0:t}]_j^\mathsf{T} \theta\right)\right)$$

$$\geqslant \min_{i,j \in [K]} \left([M_{0:t}]_i - [M_{0:t}]_j\right)^\mathsf{T} \theta - \log K.$$

We now exponentiate to recover a bound for the conditional likelihood:

$$p_\theta(e_k \mid z_{0:t}) \geqslant \min_{i,j \in [K]} \exp\left(\left([M_{0:t}]_i - [M_{0:t}]_j\right)^\mathsf{T} \theta - \log K\right)$$

$$= \min_{i,j \in [K]} \frac{1}{K} \exp\left(-\left([M_{0:t}]_i - [M_{0:t}]_j\right)^\mathsf{T} \theta\right).$$

Through Cauchy-Schwartz we may bound the quantity inside the exponent, giving our final bound.

$$\max_{i,j \in [K]} \left([M_{0:t}]_i - [M_{0:t}]_j\right)^\mathsf{T} \theta \leqslant \max_{i,j \in [K]} \|[M_{0:t}]_i - [M_{0:t}]_j\| \, \|\theta\|$$

$$\leqslant \max_{i \in [K]} 2\|[M_{0:t}]_i\| \, \|\theta\|$$

$$\leqslant 2R\|C\|_{\mathrm{op}},$$

$$\implies p_\theta(e_k \mid z_{0:t}) \geqslant \frac{1}{K} \exp\left(-2R\|C\|_{\mathrm{op}}\right).$$

Therefore we can set $\mu = \frac{1}{K} \exp\left(-2R\|C\|_{\mathrm{op}}\right)$ in order to satisfy the condition. This gives that $\lambda_{\min}\left(J^\mathsf{T}(\mathsf{diag}(\mathbb{S}(M_{0:t}\theta)) - \mathbb{S}(M_{0:t}\theta)\mathbb{S}(M_{0:t}\theta)^\mathsf{T})J\right) \geqslant \frac{\exp(-2\|C\|_{\mathrm{op}}R)}{4K(K-1)}$. Once we have this, we can extract the inner matrix from $\mathcal{I}(\theta_\star)$ by lower bounding the Rayleigh quotient, giving:

$$\mathbb{E}_{z_{0:t} \sim p_\theta}[\mathcal{I}_{t+1}(\theta \mid z_{0:t})] = \mathbb{E}_{z_{0:t} \sim p_\theta}\left[\bar{M}_{0:t}^\mathsf{T} J^\mathsf{T}\left(\mathsf{diag}(\mathbb{S}(M_{0:t}\theta)) - \mathbb{S}(M_{0:t}\theta)\mathbb{S}(M_{0:t}\theta)^\mathsf{T}\right) J\bar{M}_{0:t}\right]$$

$$\geq \frac{\exp\left(-2\|C\|_{\mathrm{op}}R\right)}{4K(K-1)} \mathbb{E}_{z_{0:t}\sim p_\theta}\left[\bar{M}_{0:t}^\mathsf{T}\bar{M}_{0:t}\right]$$

$$= \frac{\exp\left(-2\|C\|_{\mathrm{op}}R\right)}{4K(K-1)} \mathbb{E}_{z_{0:t}\sim p_\theta}\left[(z_t z_t^\mathsf{T}) \otimes \left(\frac{1}{t^2} z_{0:t-1}^\mathsf{T} z_{0:t-1} C^\mathsf{T} C z_{0:t-1}^\mathsf{T} z_{0:t-1}\right)\right]$$

$$= \frac{\exp\left(-2\|C\|_{\mathrm{op}}R\right)}{4K(K-1)} \mathbb{E}_{z_t\sim p_\theta}\left[z_t z_t^\mathsf{T}\right] \otimes \mathbb{E}_{z_{0:t-1}\sim p_\theta}\left[\frac{1}{t^2} z_{0:t-1}^\mathsf{T} z_{0:t-1} C^\mathsf{T} C z_{0:t-1}^\mathsf{T} z_{0:t-1}\right].$$

The smallest eigenvalue of this quantity can be lower bounded by the minimum eigenvalue of both sides of the Kronecker product. The first quantity can be handled by lower bounding the probability of seeing a particular token,

$$\lambda_{\min}(\mathbb{E}_{z_t\sim p_\theta}[z_t z_t^\mathsf{T}]) = \lambda_{\min}\left(\sum_{k\in[K]} (e_k e_k^\mathsf{T})\, p_\theta(e_k \mid z_{0:t-1})\right)$$

$$\geq \frac{1}{K}\sigma_{\min}^2(E)\exp(-2\|C\|_{\mathrm{op}}R).$$

For the second quantity, we note that the minimum eigenvalue function is concave so we may lower bound the expression by moving it into the expectation:

$$\lambda_{\min}\left(\mathbb{E}_{z_{0:t-1}\sim p_\theta}\left[\frac{1}{t^2} z_{0:t-1}^\mathsf{T} z_{0:t-1} C^\mathsf{T} C z_{0:t-1}^\mathsf{T} z_{0:t-1}\right]\right)$$

$$\geq \sigma_{\min}^2(C) \cdot \lambda_{\min}^2(\mathbb{E}_{z_{0:t-1}\sim p_\theta}[t^{-1} z_{0:t-1}^\mathsf{T} z_{0:t-1}])$$

$$\geq \sigma_{\min}^2(C) \cdot \frac{1}{t}\sum_{s=0}^{t-1} \lambda_{\min}^2(\mathbb{E}_{z_s\sim p_\theta|z_{0:s-1}}[z_s z_s^\mathsf{T}])$$

$$\geq \frac{1}{K}\sigma_{\min}^2(C)\sigma_{\min}^2(E)\exp(-2\|C\|_{\mathrm{op}}R).$$

Finally, we may put this all together for an expression for the minimum eigenvalues of the conditional and unconditional Fisher information matrices:

$$\lambda_{\min}\left(\mathbb{E}_{z_{0:t}\sim p_\theta}[\mathcal{I}_{t+1}(\theta \mid z_{0:t})]\right) \geq \frac{\sigma_{\min}^2(C)\sigma_{\min}^4(E)}{4K^3(K-1)} \exp(-6\|C\|_{\mathrm{op}}R), \tag{4.69}$$

$$\lambda_{\min}\left(\mathbb{E}_{z_{0:T}\sim p_\theta}[\mathcal{I}(\theta)]\right) \geq (T-1)\frac{\sigma_{\min}^2(C)\sigma_{\min}^4(E)}{4K^3(K-1)} \exp(-6\|C\|_{\mathrm{op}}R). \tag{4.70}$$

We can now begin working on finding bounds for the constants B_1 and B_2. Let us take a test vector $v \in \mathbb{R}^d$ with magnitude $\|v\| = \|\mathcal{I}(\theta_\star)^{1/2}\|_{\mathrm{op}}$ and analyze ψ_{t+1} defined as follows:

$$\psi_{t+1} := v^\mathsf{T}\left(\nabla_\theta \log p_\theta(z_{t+1} \mid z_{0:t})\right)$$

$$= v^\mathsf{T}\left([M_{0:t}]_{(z_{t+1})} - M_{0:t}^\mathsf{T}\mathbb{S}(M_{0:t}\theta)\right).$$

We can immediately see that $\mathbb{E}_{p_\theta}[\psi_{t+1} \mid z_{0:t}] = 0$. We may observe that $\mathbf{1}(z_{t+1}) - \mathbb{S}(M_{0:t}\theta)$ is a bounded random variable vector:

$$\|\mathbf{1}(z_{t+1}) - \mathbb{S}(M_{0:t}\theta)\|^2 = (1 - \mathbb{S}(M_{0:t}\theta)_{(z_{t+1})})^2 + \sum_{i:e_i\neq z_{t+1}} \mathbb{S}(M_{0:t}\theta)_i^2$$

$$\leqslant (1 - \mathbb{S}(M_{0:t}\theta)_{(z_{t+1})})^2 + \left(\sum_{i:e_i \neq z_{t+1}} \mathbb{S}(M_{0:t}\theta)_i\right)^2$$

$$= 2(1 - \mathbb{S}(M_{0:t}\theta)_{(z_{t+1})})^2 \leqslant 2.$$

From this we can observe the following:

$$\psi_{t+1} = v^\mathsf{T} \left([M_{0:t}]_{(z_{t+1})} - M_{0:t}^\mathsf{T}\mathbb{S}(M_{0:t}\theta)\right)$$
$$= v^\mathsf{T} M_{0:t}^\mathsf{T} \left(\mathbf{1}(z_{t+1}) - \mathbb{S}(M_{0:t}\theta)\right)$$
$$\leqslant \|v\|\|M_{0:t}\|_{\mathrm{op}}\|\mathbf{1}(z_{t+1}) - \mathbb{S}(M_{0:t}\theta)\|$$
$$\leqslant \sqrt{2}\|v\|\|C\|_{\mathrm{op}}.$$

This all gives that ψ_{t+1} is a zero-mean bounded random variable given by $\sigma^2 = 2\|v\|^2\|C\|_{\mathrm{op}}^2$, and we can see that $\langle v, \nabla_\theta \log p_\theta(z_{0:T})\rangle = \sum_{t=2}^T \psi_t$ is a martingale sum. This allows us to apply Azuma-Hoeffding (Theorem A.7), giving us that

$$\mathbb{P}\left[\langle v, \nabla_\theta \log p_\theta(z_{0:t})\rangle^4 \geqslant t\right] = \mathbb{P}\left[\sum_{s=2}^t |\psi_s| \geqslant u^{1/4}\right] \leqslant 2\exp\left(-\frac{\sqrt{u}}{4(t-1)\|v\|^2\|C\|_{\mathrm{op}}^2}\right),$$

and as such $\mathbb{E}[\langle v, \nabla_\theta \log p_\theta(z_{0:t})\rangle^4]^{1/4} \leqslant 4\sqrt{2T}\|v\|\|C\|_{\mathrm{op}}$. We note that applying Rosenthal's inequality for MDS (cf. Theorem A.6) retrieves a similar result, but requires a longer proof in order to express a bound for both terms. Taking $v = \mathcal{I}(\theta_\star)^{1/2}\bar{v}$ for unit vector $\bar{v}$ results in $B_1 \leqslant 4\sqrt{\frac{2T}{\lambda_{\min}(\mathcal{I}(\theta_\star))}}\|C\|_{\mathrm{op}}$.

Since the Hessian of the conditional log-likelihood has no dependence on the new token, bounding B_2 reduces to $\sup_{\theta \in \Theta}\|\mathcal{I}(\theta)^{-1/2}M_{0:t}^\mathsf{T}(\mathrm{diag}(\mathbb{S}(M_{0:t}\theta)) - \mathbb{S}(M_{0:t}\theta)\mathbb{S}(M_{0:t}\theta)^\mathsf{T})M_{0:t}\mathcal{I}(\theta)^{-1/2}\|_{\mathrm{op}}$ which may be bounded by $\frac{2}{\lambda_{\min}(\mathcal{I}(\theta))}\|C\|_{\mathrm{op}}^2$.

Step 3: Parameter error bound. From here, we can unlock the first set of bounds by verifying (3.22):

$$\sup_{\theta \in \mathrm{conv}\{\hat{\theta}_{m,T}^\varepsilon, \theta_\star\}} \mathrm{d}_H^2(\theta, \theta_\star) \lesssim \frac{d^2}{m}\log\left(\frac{c_0 R^2\|C\|_{\mathrm{op}}^2 mT}{\delta^2}\right) + \frac{1+\delta^2}{m}$$

$$\lesssim \frac{\lambda_{\min}(\mathcal{I}(\theta_\star))^2}{T^2\|C\|_{\mathrm{op}}^4}.$$

In order to satisfy this condition, we can take $m \geqslant \left(d^2\log\left(\frac{TR^2\|C\|_{\mathrm{op}}^2 m}{\delta^2}\right) + 1 + \delta^2\right) \cdot \frac{T^2\|C\|_{\mathrm{op}}^4}{\lambda_{\min}(\mathcal{I}(\theta_\star))^2}$. If we apply Proposition A.1 and expand out the minimum eigenvalue, for $T > 1$ this gives us:

$$m \gtrsim \frac{K^8\|C\|_{\mathrm{op}}^4\exp(12R\|C\|_{\mathrm{op}})}{\sigma_{\min}(C)^4\sigma_{\min}(E)^8}\left(1 + \delta^2 + d^2R\|C\|_{\mathrm{op}} + d^2\log\left(\frac{c_0 TK^8 d^2 R^2\|C\|_{\mathrm{op}}^6}{\delta^2\sigma_{\min}(C)^4\sigma_{\min}(E)^8}\right)\right). \tag{4.71}$$

Step 4: Verify FI radius. We now need to show (3.24) in order to unlock the second bound. To this end, we show a Lipschitz condition on the conditional Fisher information from eq. (4.67):

$$\|\mathcal{I}_{t+1}(\theta_1 \mid z_{0:t}) - \mathcal{I}_{t+1}(\theta_2 \mid z_{0:t})\|_{\mathrm{op}}$$

$$= \|M_{0:t}^{\mathsf{T}}(\mathsf{diag}(\mathbb{S}(M_{0:t}\theta_1)) - \mathbb{S}(M_{0:t}\theta_1)^{\otimes 2} - \mathsf{diag}(\mathbb{S}(M_{0:t}\theta_2)) + \mathbb{S}(M_{0:t}\theta_2)^{\otimes 2})M_{0:t}\|_{\mathrm{op}}$$

$$\leqslant \|M_{0:t}\|_{\mathrm{op}}^2 \left(\|\mathbb{S}(M_{0:t}\theta_1) - \mathbb{S}(M_{0:t}\theta_2)\|_\infty + \|\mathbb{S}(M_{0:t}\theta_1)^{\otimes 2} - \mathbb{S}(M_{0:t}\theta_2)^{\otimes 2}\|_{\mathrm{op}} \right)$$

$$\leqslant \|C\|_{\mathrm{op}}^2 \left(\|M_{0:t}\|_{\mathrm{op}}\|\theta_1 - \theta_2\| + (\|\mathbb{S}(M_{0:t}\theta_1)\| + \|\mathbb{S}(M_{0:t}\theta_2)\|) \|\mathbb{S}(M_{0:t}\theta_1) - \mathbb{S}(M_{0:t}\theta_2)\| \right)$$

$$\leqslant \|C\|_{\mathrm{op}}^2 \left(\|C\|_{\mathrm{op}}\|\theta_1 - \theta_2\| + 2\|\mathbb{S}(M_{0:t}\theta_1) - \mathbb{S}(M_{0:t}\theta_2)\| \right)$$

$$\leqslant 3\|C\|_{\mathrm{op}}^3\|\theta_1 - \theta_2\|.$$

This gives us that $\mathrm{Lip} \leqslant 3\|C\|_{\mathrm{op}}^3$. Let us now find an expression for the second moment bound:

$$B_{\mathcal{I}} = \sup_{\theta_1,\theta_2 \in \Theta_s} \sup_{v \in \mathbb{S}^{p-1}} \max_{t \in [T-1]} \|v^{\mathsf{T}}\mathcal{I}_{t+1}(\theta_1 \mid z_{0:t})v\|_{\mathcal{L}^2(p_{\theta_2})}$$

$$\leqslant \sup_{\theta_1,\theta_2 \in \Theta_s} \sup_{v \in \mathbb{S}^{p-1}} \max_{t \in [T-1]} \left(\mathbb{E}_{z_{0:t} \sim p_{\theta_2}} \mathbb{E}_{z_{t+1} \sim p_{\theta_1}} \left[\left(v^{\mathsf{T}} M_{0:t}^{\mathsf{T}} \left(\mathsf{diag}\left(\mathbb{S}(M_{0:t}\theta_1)\right) - \mathbb{S}(M_{0:t}\theta_1)^{\otimes 2} \right) M_{0:t}v \right)^2 \right] \right)^{1/2}$$

$$\leqslant \|C\|_{\mathrm{op}}^2.$$

With both a globally bounded Lipschitz constant and a finite $B_{\mathcal{I}}$, we can apply Proposition A.4 to bound the difference in Fisher informations:

$$\|\mathcal{I}(\theta_1) - \mathcal{I}(\theta_2)\|_{\mathrm{op}} \leqslant 3T\|C\|_{\mathrm{op}}^3\|\theta_1 - \theta_2\| + 2\sqrt{2}T\|C\|_{\mathrm{op}}^2 \mathrm{d}_H(p_{\theta_1}, p_{\theta_2}).$$

Note that the lower bound on $\lambda_{\min}(\mathcal{I}(\theta))$ in (4.70) is agnostic to the value of $\theta \in \Theta$. Therefore,

$$\lambda_{\min}(\mathcal{I}(\theta_\star, \hat{\theta})) \geqslant (T-1) \frac{\sigma_{\min}^2(C)\sigma_{\min}^4(E)}{4K^3(K-1)} \exp(-6\|C\|_{\mathrm{op}}R).$$

If we apply this to (3.23), we can upper bound the parameter distance:

$$\frac{1}{8}\|\theta_0 - \theta_1\|_{\mathcal{I}(\theta_0,\theta_1)}^2 \leqslant \mathrm{d}_H^2(\theta_0,\theta_1) \lesssim \frac{d^2}{m} \log\left(\frac{c_0 R^2 \|C\|_{\mathrm{op}}^2 mT}{\delta^2} \right) + \frac{1}{m} + \frac{\delta^2}{m},$$

$$\implies \|\theta_0 - \theta_1\| \leqslant \frac{2\sqrt{2}}{\sqrt{\lambda_{\min}(\mathcal{I}(\theta_0,\theta_1))}} \mathrm{d}_H(p_{\theta_0}, p_{\theta_1}).$$

If we put these together, we have the following:

$$\sup_{\theta \in \mathrm{conv}\{\theta_\star, \hat{\theta}_{m,T}^\varepsilon\}} \|\mathcal{I}(\theta_\star)^{-1/2}\mathcal{I}(\theta)\mathcal{I}(\theta_\star)^{-1/2} - I_{d^2}\|_{\mathrm{op}}$$

$$\leqslant \frac{1}{\lambda_{\min}(\mathcal{I}(\theta_\star))} \sup_{\theta \in \mathrm{conv}\{\theta_\star, \hat{\theta}_{m,T}^\varepsilon\}} \|\mathcal{I}(\theta) - \mathcal{I}(\theta_\star)\|_{\mathrm{op}}$$

$$\leqslant \frac{2\sqrt{2}T\|C\|_{\mathrm{op}}^2}{\lambda_{\min}(\mathcal{I}(\theta_\star))} \left(\frac{3\|C\|_{\mathrm{op}}}{\sqrt{\lambda_{\min}(\mathcal{I}(\theta_\star))}} + 1 \right) \sup_{\theta \in \mathrm{conv}\{\theta_\star, \hat{\theta}_{m,T}^\varepsilon\}} \mathrm{d}_H(p_\theta, p_{\theta_\star}).$$

If we take $T \gtrsim \frac{K^4\|C\|_{\mathrm{op}}^2}{\sigma_{\min}(C)^2\sigma_{\min}(E)^4} \exp(6R\|C\|_{\mathrm{op}})$ such that $\|C\|_{\mathrm{op}}\lambda_{\min}(\mathcal{I}(\theta_\star))^{-1/2} \lesssim 1$, this is bounded above by $1/2$ for:

$$m \gtrsim \frac{T^2\|C\|_{\mathrm{op}}^4}{\lambda_{\min}(\mathcal{I}(\theta_\star))^2} \left(d^2 \log\left(\frac{c_0 R^2 \|C\|_{\mathrm{op}}^2 mT}{\delta^2} \right) + 1 + \delta^2 \right).$$

Step 5: Final result. Finally, we can plug in our expression for the minimum eigenvalue in eq. (4.70) to take the following bound

$$m \gtrsim \frac{K^8 \|C\|_{\mathrm{op}}^4 \exp(12R\|C\|_{\mathrm{op}})}{\sigma_{\min}(C)^4 \sigma_{\min}(E)^8} \left(d^2 \log \left(\frac{c_0 R^2 \|C\|_{\mathrm{op}}^2 mT}{\delta^2} \right) + 1 + \delta^2 \right),$$

and applying Proposition A.1 extracts m to satisfy the Fisher radius:

$$m \gtrsim \frac{K^8 \|C\|_{\mathrm{op}}^4 \exp(12R\|C\|_{\mathrm{op}})}{\sigma_{\min}(C)^4 \sigma_{\min}(E)^8} \left(1 + \delta^2 + d^2 R\|C\|_{\mathrm{op}} + d^2 \log \left(\frac{c_0 T K^8 d^2 R^2 \|C\|_{\mathrm{op}}^6}{\delta^2 \sigma_{\min}(C)^4 \sigma_{\min}(E)^8} \right) \right).$$

Notably, this additionally satisfies (4.71). In this case, we unlock the following bound for $\delta \in (0,1)$:

$$\|\hat{\theta}_{m,T}^\varepsilon - \theta_\star\|_{\bar{\mathcal{I}}(\theta_\star)} \leqslant \frac{32}{3T} \mathrm{d}_H^2(\hat{\theta}_{m,T}^\varepsilon, \theta_\star)$$

$$\leqslant \sup_{\theta \in \mathrm{conv}\{\hat{\theta}_{m,T}^\varepsilon, \theta_\star\}} \frac{32}{3T} \mathrm{d}_H^2(\theta, \theta_\star)$$

$$\lesssim \frac{d^2}{mT} \log \left(\frac{c_0 R^2 \|C\|_{\mathrm{op}}^2 mT}{\delta^2} \right) + \frac{1 + \delta^2}{mT}$$

$$\lesssim \frac{d^2}{mT} \log \left(\frac{c_0 R^2 \|C\|_{\mathrm{op}}^2 mT}{\delta^2} \right).$$

5 Conclusion

We introduced the Hellinger localization framework for deriving nearly instance-optimal parameter recovery rates for multi-trajectory learning setups. We applied our framework to a diverse set of case studies, including a mixture of Markov chains example, a dependent linear regression problem with general noise distributions, a non-monotonic sinusoidal GLM example, and a linear attention sequence modeling setup. In each case, we showed that our Hellinger localization framework was able to provide nearly instance-optimal rates that significantly improve upon the prior art.

Our work further opens up several avenues for future investigation. We list out a few ideas, starting with technical improvements, and ending with more broader, high-level directions.

(a) *Extensions of self-normalization:* As discussed in Section 4.2.1, one particular drawback of our current framework is that it places un-necessary requirements on the regularity of the trajectory process $z_{1:T}$. Concretely, in the context of dependent linear regression, the process $z_{1:T}$ can not grow more than $\mathrm{poly}(T)$, which rules out e.g., recovering linear dynamical systems with spectral radius > 1. We believe this restriction is purely a technical limitation of our argument, which currently does not have a method to self-normalize as is done in analysis specialized for least-squares linear regression. The work of [80] which generalizes offset complexity to exp-concave losses is a natural starting point for such an inquiry.

(b) *Improved minimum trajectory requirements for non-log-concave families:* As we discussed in Section 3.4, another limitation of our current analysis is that whenever the family of distribution $\mathcal{P}$ is not log-concave, our requirements on the number of trajectory m grows

from $m \gtrsim \mathrm{polylog}(T)$ in the log-concave setting, to $m \gtrsim T \cdot \mathrm{polylog}(T)$. We believe this scaling should generally be improvable. One possible pathway is to utilize the local geodesic convexity of the squared Hellinger distance in the Fisher-Rao metric [101], and conduct our second-order Taylor analysis (cf. Proposition 3.9) over geodesics.

(c) *Non-realizable settings:* Our work is carried out in the realizable setting, i.e., where the data generating distribution $p_\star \in \mathcal{P}$. A natural and useful extension would be to allow for $p_\star \notin \mathcal{P}$, and study convergence to the best distribution in $\mathcal{P}$, i.e., $\theta_\star^{\mathcal{P}} := \arg\min_{p \in \mathcal{P}} \mathrm{KL}(p_\star \parallel p)$. One key technical challenge for the non-realizable setting is extending Theorem 3.6 to measure squared Hellinger distance $\mathrm{d}_H^2(\hat{p}_{m,T}^\varepsilon, p_\star^{\mathcal{P}})$ without relying on e.g., max divergence coverings (cf. Theorem 3.1), but instead allowing for some less stringent tail behavior for the log-likelihoods which is still practical to verify. This could also be useful in allowing Theorem 3.6 to apply directly to the MLE estimator and not its discretized counterpart (cf. Remark 3.7).

(d) *Applications to non-sequentially dependent data:* While our work focuses on sequentially-ordered stochastic processes, our main tools in Section 3 (i.e., Theorem 3.6 and Proposition 3.9) are actually agnostic to this sequential structure. It is only when we analyze the score function and observed information matrix moments (i.e., (3.20) and (3.21) from Proposition 3.9) that we impose a temporal dependence in the data. Hence, an interesting future direction is to apply our main tools to other problem settings with different correlation structures, such as for Ising models (cf. related work from Section 2) and other graph/network structures [102–104].

(e) *Applications for filtering and control problems:* Finally, through the case studies in Section 4, we have looked at parameter recovery in various types of dynamical systems. A natural next step is to consider the downstream control task where the recovered model parameters would be applied, by extending our results to enable task-specific optimal exploration for a broader family of parametric models and loss functions (cf. discussion in Section 4.2.1). Another direction is to apply our framework for filtering problems in state estimation, which can be cast as a latent maximum likelihood estimation problems. Here, an important sub-direction would be to study the application of our techniques to analyzing not just the exact MLE estimate, but also practical algorithms such as expectation-maximization and variational inference, which are necessary in situations where directly computing the MLE is computationally intractable.

Acknowledgments

References

[1] Stephen Tu, Roy Frostig, and Mahdi Soltanolkotabi. Learning from many trajectories. *Journal of Machine Learning Research*, 25(216):1–109, 2024.

[2] Ingvar Ziemann, Stephen Tu, George J. Pappas, and Nikolai Matni. Sharp rates in dependent learning theory: Avoiding sample size deflation for the square loss. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 62779–62802. PMLR, 21–27 Jul 2024.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen

Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc., 2020.

[5] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.

[6] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR, 06–09 Nov 2023.

[7] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017.

[8] Richard C. Bradley. Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions. *Probability Surveys*, 2(none):107–144, 2005.

[9] Bin Yu. Rates of Convergence for Empirical Processes of Stationary Mixing Sequences. *The Annals of Probability*, 22(1):94–116, 1994.

[10] Rajeeva L. Karandikar and M. Vidyasagar. Rates of uniform convergence of empirical means with mixing processes. *Statistics & Probability Letters*, 58(3):297–307, 2002. ISSN 0167-7152.

[11] Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary ϕ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11(26):789–814, 2010.

[12] Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, 2017.

[13] Qian Qin and James P Hobert. On the limitations of single-step drift and minorization in Markov chain convergence analysis. *The Annals of Applied Probability*, 31(4):1633–1659, 2021.

[14] Qian Qin, James P. Hobert, and Kshitij Khare. Estimating the spectral gap of a trace-class Markov operator. *Electronic Journal of Statistics*, 13(1):1790 – 1822, 2019.

[15] Daniel Hsu, Sham M. Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012. ISSN 0022-0000. JCSS Special Issue: Cloud Computing 2011.

[16] Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence off-policy evaluation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 3000–3006. AAAI Press, 2015. ISBN 0262511290.

[17] Yanxi Chen and H. Vincent Poor. Learning mixtures of linear dynamical systems. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 3507–3557. PMLR, 17–23 Jul 2022.

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[19] Ingvar Ziemann and Stephen Tu. Learning with little mixing. In *Advances in Neural Information Processing Systems*, volume 35, pages 4626–4637. Curran Associates, Inc., 2022.

[20] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5156–5165. PMLR, 13–18 Jul 2020.

[21] Ron Meir. Nonparametric time series prediction through adaptive model selection. *Machine learning*, 39(1): 5–34, 2000.

[22] Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-i.i.d. processes. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

[23] John C. Duchi, Alekh Agarwal, Mikael Johansson, and Michael I. Jordan. Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578, 2012.

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

[25] Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5610–5618. PMLR, 09–15 Jun 2019.

[26] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, 2018. ISSN 0005-1098.

[27] Suhas Kowshik, Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. Near-optimal offline and streaming algorithms for learning non-linear dynamical systems. In *Advances in Neural Information Processing Systems*, volume 34, pages 8518–8531. Curran Associates, Inc., 2021.

[28] Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: Upper and minimax bounds. *arXiv preprint arXiv:1305.4825*, 2013.

[29] Laurent Duchesne, Eric Feron, James Paduano, and Marty Brenner. Subspace identification with multiple data sets. In *AIAA, Guidance, Navigation and Control Conference*, 1996.

[30] Ivan Markovsky and Rik Pintelon. Identification of linear time-invariant systems from multiple experiments. *IEEE Transactions on Signal Processing*, 63(13):3549–3554, 2015.

[31] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2020.

[32] Yang Zheng and Na Li. Non-asymptotic identification of linear dynamical systems using multiple trajectories. *IEEE Control Systems Letters*, 5(5):1693–1698, 2021.

[33] Yu Xing, Benjamin Gravell, Xingkang He, Karl Henrik Johansson, and Tyler H. Summers. Identification of linear systems with multiplicative noise from multiple trajectory data. *Automatica*, 144:110486, 2022. ISSN 0005-1098.

[34] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online least squares estimation with self-normalized processes: An application to bandit problems. *arXiv preprint arXiv:1102.2670*, 2011.

[35] Shahar Mendelson. Learning without concentration. *J. ACM*, 62(3), June 2015. ISSN 0004-5411.

[36] Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge University Press, 2000.

[37] Tong Zhang. From ε-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006.

[38] Dylan Foster, Tuhin Sarkar, and Alexander Rakhlin. Learning nonlinear dynamical systems from a single trajectory. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120, pages 851–861. PMLR, 2020.

[39] Yury Polyanskiy and Yihong Wu. *Information theory: From coding to learning*. Cambridge University Press, 2025.

[40] Elad Hazan, Karan Singh, and Cyril Zhang. Learning linear dynamical systems via spectral filtering. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[41] Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Spectral filtering for general linear dynamical systems. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[42] Adam Block, Yuval Dagan, Noah Golowich, and Alexander Rakhlin. Smoothed online learning is as easy as statistical learning. 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 1716–1786. PMLR, 02–05 Jul 2022.

[43] Evan Dogariu, Anand Brahmbhatt, and Elad Hazan. Universal learning of nonlinear dynamics. *arXiv preprint arXiv:2508.11990*, 2025.

[44] Sourav Chatterjee. Estimation in spin glasses: A first step. *The Annals of Statistics*, 35(5):1931–1946, 2007.

[45] Bhaswar B. Bhattacharya and Sumit Mukherjee. Inference in Ising models. *Bernoulli*, 24(1):493–525, 2018.

[46] Guy Bresler. Efficiently learning ising models on arbitrary graphs. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, page 771–782, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450335362.

[47] Marc Vuffray, Sidhant Misra, Andrey Lokhov, and Michael Chertkov. Interaction screening: Efficient and sample-optimal learning of ising models. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[48] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, 2000.

[49] Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3, 2002.

[50] Vladimir Spokoiny. Parametric estimation. Finite sample theory. *The Annals of Statistics*, 40(6):2877–2909, 2012.

[51] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[52] Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset Rademacher complexity. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1260–1285, Paris, France, 03–06 Jul 2015. PMLR.

[53] Dylan J. Foster, Adam Block, and Dipendra Misra. Is behavior cloning all you need? understanding horizon in imitation learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 120602–120666. Curran Associates, Inc., 2024.

[54] Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. In *Advances in Neural Information Processing Systems*, volume 33, pages 20095–20107. Curran Associates, Inc., 2020.

[55] Jiawei Ge, Shange Tang, Jianqing Fan, and Chi Jin. On the provable advantage of unsupervised pretraining. In *The Twelfth International Conference on Learning Representations*, 2024.

[56] Ingvar Ziemann. A short information-theoretic analysis of linear auto-regressive learning. *arXiv preprint arXiv:2409.06437*, 2024.

[57] Constantinos Daskalakis, Nishanth Dikkala, and Nick Gravin. Testing symmetric Markov chains from a single trajectory. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 385–409. PMLR, 06–09 Jul 2018.

[58] Rishi Gupta, Ravi Kumar, and Sergei Vassilvitskii. On mixtures of Markov chains. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[59] Chinmaya Kausik, Kevin Tan, and Ambuj Tewari. Learning mixtures of Markov chains and MDPs. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15970–16017. PMLR, 23–29 Jul 2023.

[60] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[61] Fabian Spaeh and Charalampos Tsourakakis. Learning mixtures of Markov chains with quality guarantees. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 662–672, New York, NY, USA, 2023. Association for Computing Machinery.

[62] Geoffrey McLachlan and David Peel. *Finite Mixture Models*. Wiley, September 2000.

[63] Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20, January 2015.

[64] Jérôme Dedecker and Clémentine Prieur. New dependence coefficients. examples and applications to statistics. *Probability Theory and Related Fields*, 132(2):203–236, December 2004.

[65] Véronique Maume-Deschamps. Exponential inequalities and functional estimations for weak dependent data: Applications to dynamical systems. *Stochastics and Dynamics*, 06(04):535–560, December 2006.

[66] Florence Merlevède, Magda Peligrad, and Emmanuel Rio. *Bernstein inequality and moderate deviations under strong mixing conditions*, page 273–292. Institute of Mathematical Statistics, 2009.

[67] Wintenberger Olivier. Deviation inequalities for sums of weakly dependent time series. *Electronic Communications in Probability*, 15(none), January 2010.

[68] Hanyuan Hang and Ingo Steinwart. A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *The Annals of Statistics*, 45(2), April 2017.

[69] Zihao Yuan and Holger Dette. Exponential inequalities for some mixing processes and dynamic systems, 2025.

[70] Eckhard Liebscher. Towards a unified approach for proving geometric ergodicity and mixing properties of nonlinear autoregressive processes. *Journal of Time Series Analysis*, 26(5):669–689, August 2005.

[71] De Huang and Xiangyuan Li. Bernstein-type inequalities for Markov chains and Markov processes: A simple and robust proof, 2024.

[72] Yanxi Chen and H. Vincent Poor. Learning mixtures of linear dynamical systems. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 3507–3557. PMLR, 17–23 Jul 2022.

[73] Horia Mania, Michael I. Jordan, and Benjamin Recht. Active learning for nonlinear system identification with guarantees. *J. Mach. Learn. Res.*, 23(1), January 2022. ISSN 1532-4435.

[74] Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

[75] Andrew Wagenmaker, Guanya Shi, and Kevin G Jamieson. Optimal exploration for model-based rl in nonlinear systems. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 15406–15455. Curran Associates, Inc., 2023.

[76] Negin Musavi, Ziyao Guo, Geir Dullerud, and Yingying Li. Identification of analytic nonlinear dynamical systems with non-asymptotic guarantees. In *Advances in Neural Information Processing Systems*, volume 37, pages 85500–85522. Curran Associates, Inc., 2024.

[77] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, 2018. ISSN 0005-1098.

[78] Abhishek Roy, Krishnakumar Balasubramanian, and Murat A Erdogdu. On empirical risk minimization with dependent and heavy-tailed data. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8913–8926. Curran Associates, Inc., 2021.

[79] Vinay Kanakeri and Aritra Mitra. Outlier-robust linear system identification under heavy-tailed noise. In Necmiye Ozay, Laura Balzano, Dimitra Panagou, and Alessandro Abate, editors, *Proceedings of the 7th Annual Learning for Dynamics & Control Conference*, volume 283 of *Proceedings of Machine Learning Research*, pages 540–551. PMLR, 04–06 Jun 2025.

[80] Suhas Vijaykumar. Localization, convexity, and star aggregation. In *Advances in Neural Information Processing Systems*, volume 34, pages 4570–4581. Curran Associates, Inc., 2021.

[81] Andrew J Wagenmaker, Max Simchowitz, and Kevin Jamieson. Task-optimal exploration in linear dynamical systems. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10641–10652. PMLR, 18–24 Jul 2021.

[82] Bruce D. Lee, Ingvar Ziemann, George J. Pappas, and Nikolai Matni. Active learning for control-oriented identification of nonlinear systems. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pages 3011–3018, 2024. doi: 10.1109/CDC56724.2024.10885804.

[83] Yahya Sattar and Samet Oymak. Non-asymptotic and accurate learning of nonlinear dynamical systems. *Journal of Machine Learning Research*, 23(140):1–49, 2022.

[84] 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.

[85] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7462–7473. Curran Associates, Inc., 2020.

[86] Svante Janson. *Gaussian Hilbert spaces*. Number 129. Cambridge University Press, 1997.

[87] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[88] Wenzhe Li, Hao Luo, Zichuan Lin, Chongjie Zhang, Zongqing Lu, and Deheng Ye. A survey on transformers in reinforcement learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. Survey Certification.

[89] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[90] Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

[91] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35151–35174. PMLR, 23–29 Jul 2023.

[92] Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5793–5831. PMLR, 17–23 Jul 2022.

[93] Yuandong Tian, Yiping Wang, Beidi Chen, and Simon Shaolei Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[94] Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[95] Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023.

[96] Yingcong Li, Yixiao Huang, Muhammed Emrullah Ildiz, Ankit Singh Rawat, and Samet Oymak. Mechanics of next token prediction with self-attention. In *AISTATS*, pages 685–693, 2024.

[97] Muhammed Emrullah Ildiz, Yixiao Huang, Yingcong Li, Ankit Singh Rawat, and Samet Oymak. From self-attention to Markov models: Unveiling the dynamics of generative transformers. In *Forty-first International Conference on Machine Learning*, 2024.

[98] Shen Zhuoran, Zhang Mingyuan, Zhao Haiyu, Yi Shuai, and Li Hongsheng. Efficient attention: Attention with linear complexities. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3530–3538, 2021.

[99] Llama Team. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[100] Stephen P. Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[101] Boris Khesin, Jonatan Lenells, Gerard Misiolek, and Stephen C Preston. Geometry of diffeomorphism groups, complete integrability and optimal transport. *arXiv preprint arXiv:1105.0643*, 2011.

[102] Nicolas Usunier, Massih R. Amini, and Patrick Gallinari. Generalization error bounds for classifiers trained with interdependent data. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005.

[103] Liva Ralaivola, Marie Szafranski, and Guillaume Stempfel. Chromatic PAC-Bayes bounds for non-iid data: Applications to ranking and stationary β-mixing processes. *Journal of Machine Learning Research*, 11(65): 1927–1956, 2010.

[104] Cosma Rohilla Shalizi and Alessandro Rinaldo. Consistency under sampling of exponential random graph models. *The Annals of Statistics*, 41(2):508 − 535, 2013.

[105] Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2826–2836. PMLR, 18–24 Jul 2021.

[106] Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

[107] D. L. Burkholder. Distribution function inequalities for martingales. *The Annals of Probability*, 1(1), February 1973.

[108] Pawel Hitczenko. Best constants in martingale version of Rosenthal's inequality. *The Annals of Probability*, 18 (4), October 1990.

[109] Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[110] Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer Science & Business Media, 2013.

[111] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional Gaussians with the same mean. *arXiv preprint arXiv:1810.08693*, 2018.

A   Additional Results for Hellinger Localization

Proposition A.1 (cf. [105, Lemma F.2]). *Suppose that $\nu, a, b \geqslant 0$. Then, we have:*

$$m \geqslant (1+\nu)^{\nu} a \log^{\nu}((1+\nu)^{\nu} ab) \implies m \geqslant a \log^{\nu}(bm).$$

Proposition A.2. *Let p, q be two measures. Fix $\delta \in (0,1)$ and $m \in \mathbb{N}_+$. We have that:*

$$\mathrm{d}_H(p,q) \leqslant \frac{\delta}{\sqrt{2m}} \implies \mathrm{d}_H(p^{\otimes m}, q^{\otimes m}) \leqslant \delta.$$

Proof. We have that $\mathrm{d}_H^2(p^{\otimes m}, q^{\otimes m}) = 2(1 - \rho(p,q)^m)$ where $\rho(p,q) := \int \sqrt{pq}\mathrm{d}\mu$ denotes the Bhattacharyya coefficient between p and q (note that $\mathrm{d}_H^2(p,q) = 2(1 - \rho(p,q))$). Let $\rho = \rho(p,q)$. Note that if $\rho \geqslant (1 - \delta^2/2)^{1/m}$ then we have $\sqrt{2(1-\rho^m)} \leqslant \delta$. On the other hand since $(1 - \delta^2/2)^{1/m} \leqslant \exp(-\delta^2/(2m))$, it suffices to take $\rho \geqslant \exp(-\delta^2/(2m))$. The latter condition is equal to $\mathrm{d}_H^2(p,q)/2 \leqslant 1 - \exp(-\delta^2/(2m))$. Using the inequality $\exp(-x) \leqslant 1 - x + x^2/2$ valid for $x \in [0,1]$, we have that:

$$\exp(-\delta^2/(2m)) \leqslant 1 - \frac{\delta^2}{2m}\left(1 - \frac{\delta^2}{4m}\right) \leqslant 1 - \frac{\delta^2}{4m} \implies 1 - \exp(-\delta^2/(2m)) \geqslant \frac{\delta^2}{4m}.$$

Hence, we have shown that:

$$\mathrm{d}_H(p,q) \leqslant \frac{\delta}{\sqrt{2m}} \implies \mathrm{d}_H(p^{\otimes m}, q^{\otimes m}) \leqslant \delta.$$

$\square$

Proposition A.3. *Let μ, ν be two probability measures on the same measure space X, and let $f : \mathsf{X} \mapsto \mathbb{R}$ be a real-valued function. We have:*

$$|\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]| \leqslant \sqrt{2}(\|f\|_{L^2(\mu)} + \|f\|_{L^2(\nu)})\mathrm{d}_H(\mu,\nu).$$

Proof. Let λ be a common measure and let p_μ, p_ν denote the resulting Radon-Nikodym derivatives. We have:

$$
\begin{aligned}
\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f] &= \int f(x)(p_\mu(x) - p_\nu(x))\mathrm{d}\lambda \\
&= \int f(x)(\sqrt{p_\mu(x)} + \sqrt{p_\nu(x)})(\sqrt{p_\mu(x)} - \sqrt{p_\nu(x)})\mathrm{d}\lambda \\
&\leqslant \sqrt{\int f(x)^2(\sqrt{p_\mu(x)} + \sqrt{p_\nu(x)})^2\mathrm{d}\lambda}\sqrt{\int (\sqrt{p_\mu(x)} - \sqrt{p_\nu(x)})^2\mathrm{d}\lambda} \\
&\leqslant \sqrt{2\int f(x)^2 p_\mu(x)\mathrm{d}\lambda + 2\int f(x)^2 p_\nu(x)\mathrm{d}\lambda} \cdot \mathrm{d}_H(\mu,\nu) \\
&= \sqrt{2}(\|f\|_{L^2(\mu)} + \|f\|_{L^2(\nu)})\mathrm{d}_H(\mu,\nu).
\end{aligned}
$$

Above, the first inequality is Cauchy-Schwarz, and the second is $(a+b)^2 \leqslant 2(a^2 + b^2)$. The claim now follows by reversing the role of μ, ν. $\square$

Proposition A.4. *For $t \in [T]$, define the conditional Fisher information matrices as:*

$$\mathcal{I}_t(\theta \mid z_{1:t-1}) := -\mathbb{E}_{p_\theta}\left[\nabla^2 \log p_\theta(z_t \mid z_{1:t-1}) \mid z_{1:t-1}\right]. \tag{A.1}$$

(We interpret $z_{1:0}$ to condition on no information.) Let $\Theta' \subseteq \Theta$, and suppose that the following conditions hold:

(a) *For all $t \in [T]$, we have for a.e. $z_{1:t-1} \in \mathsf{Z}^{t-1}$ and $\theta_1, \theta_2 \in \Theta'$,*

$$\|\mathcal{I}_t(\theta_1 \mid z_{1:t-1}) - \mathcal{I}_t(\theta_2 \mid z_{1:t-1})\|_{\mathrm{op}} \leqslant \mathrm{Lip}_t(z_{1:t-1}) \|\theta_1 - \theta_2\|.$$

(b) *The following bound on the Lipschitz conditions holds:*

$$\mathrm{Lip} := \sup_{\theta \in \Theta_s} \max_{t \in [T]} \mathbb{E}_{p_\theta}[\mathrm{Lip}_t(z_{1:t-1})] < \infty.$$

(c) *The following second moment bound holds:*

$$B_{\mathcal{I}} := \sup_{\theta_1, \theta_2 \in \Theta_s} \sup_{v \in \mathbb{S}^{p-1}} \max_{t \in [T]} \|v^\mathsf{T} \mathcal{I}_t(\theta_1 \mid z_{1:t-1}) v\|_{\mathcal{L}^2(p_{\theta_2})} < \infty.$$

Then we have:

$$\|\mathcal{I}(\theta_1) - \mathcal{I}(\theta_2)\|_{\mathrm{op}} \leqslant T\left[\mathrm{Lip} \cdot \|\theta_1 - \theta_2\| + 2\sqrt{2}B_{\mathcal{I}} \cdot \mathrm{d}_H(p_{\theta_1}, p_{\theta_2})\right].$$

Proof. We first use the tower property to decompose $\mathcal{I}(\theta)$ as:

$$\mathcal{I}(\theta) = \sum_{t=1}^{T} \mathbb{E}_{p_\theta}[\mathcal{I}_t(\theta \mid z_{1:t-1})].$$

Hence we have for a fixed unit-norm $v \in \mathbb{R}^p$,

$$\begin{aligned}
|v^\mathsf{T}(\mathcal{I}(\theta_1) - \mathcal{I}(\theta_2))v| &= \left|\sum_{t=1}^{T} \mathbb{E}_{p_{\theta_1}}[v^\mathsf{T}\mathcal{I}_t(\theta_1 \mid z_{1:t-1})v] - \mathbb{E}_{p_{\theta_2}}[v^\mathsf{T}\mathcal{I}_t(\theta_2 \mid z_{1:t-1})v]\right| \\
&= \left|\sum_{t=1}^{T} \mathbb{E}_{p_{\theta_1}}[v^\mathsf{T}(\mathcal{I}_t(\theta_1 \mid z_{1:t-1}) - \mathcal{I}_t(\theta_2 \mid z_{1:t-1}))v] + (\mathbb{E}_{p_{\theta_1}} - \mathbb{E}_{p_{\theta_2}})[v^\mathsf{T}I_t(\theta_2 \mid z_{1:t-1})v]\right| \\
&\leqslant \sum_{t=1}^{T} \mathbb{E}_{p_{\theta_1}}[\mathrm{Lip}_t(z_{1:t-1})]\|\theta_1 - \theta_2\| + \sum_{t=1}^{T} |(\mathbb{E}_{p_{\theta_1}} - \mathbb{E}_{p_{\theta_2}})[v^\mathsf{T}I_t(\theta_2 \mid z_{1:t-1})v]| \\
&\leqslant T\mathrm{Lip} \cdot \|\theta_1 - \theta_2\| + 2\sqrt{2}TB_{\mathcal{I}} \cdot \mathrm{d}_H(p_{\theta_1}, p_{\theta_2}),
\end{aligned}$$

where the last inequality holds from Proposition A.3. The claim now follows by the variational form of the operator norm for symmetric matrices. $\square$

Proposition A.5 (cf. [106, Lemma A.4]). *Let $(X_t)_{t \in \mathbb{N}_+}$ be a sequence of real-valued random variables adapted to a filtration $(\mathcal{F}_t)_{t \in \mathbb{N}_+}$. With probability at least $1 - \delta$ for all $\tau \in \mathbb{N}_+$:*

$$\sum_{t=1}^{\tau} -\log(\mathbb{E}[e^{-X_t} \mid \mathcal{F}_{t-1}]) \leqslant \sum_{t=1}^{\tau} X_t + \log(1/\delta).$$

By negating X_t, the following inequality also holds with probability at least $1 - \delta$ for all $\tau \in \mathbb{N}_+$:

$$\sum_{t=1}^{\tau} X_t \leqslant \sum_{t=1}^{\tau} \log(\mathbb{E}[e^{X_t} \mid \mathcal{F}_{t-1}]) + \log(1/\delta).$$

Theorem A.6 (Rosenthal's inequality for MDS, [107, 108]). *Let $(d_n)_{n \geq 1}$ be a martingale difference sequence (MDS) adapted to a filtration $(\mathcal{F}_n)_{n \geq 1}$. For any $2 \leq p < \infty$,*

$$\left(\mathbb{E} \left| \sum_{k=1}^n d_k \right|^p \right)^{1/p} \leq C_p \left\{ \left(\mathbb{E} \left(\sum_{k=1}^n \mathbb{E}[d_k^2 \mid \mathcal{F}_{k-1}] \right)^{p/2} \right)^{1/p} + \left(\sum_{k=1}^n \mathbb{E}|d_k|^p \right)^{1/p} \right\},$$

where the constant C_p only depends on p.

Theorem A.7 (Azuma-Hoeffding, [109]). *Let $(\{(D_k, \mathcal{F}_k)\}_{k=1}^\infty)$ be a martingale difference sequence for which there are constants $\{(a_k, b_k)\}_{k=1}^n$ such that $D_k \in [a_k, b_k]$ almost surely for all $k \in [n]$. Then, for all $t \geq 0$,*

$$\mathbb{P}\left[\left| \sum_{k=1}^n D_k \right| \geq t \right] \leq 2 \exp\left(-\frac{2t^2}{\sum_{k=1}^n (b_k - a_k)^2} \right).$$

We next restate and prove Proposition 3.3.

Proposition 3.3. *For any $\theta_0, \theta_1 \in \Theta$ such that $\mathrm{conv}(\theta_0, \theta_1) \subseteq \Theta$, we have:*

$$d_H(p_{\theta_0}, p_{\theta_1}) \leq \frac{1}{2} d_{\mathrm{FI}}(p_{\theta_0}, p_{\theta_1}) \leq \frac{1}{2} d_{\mathcal{I}_{\max}}(p_{\theta_0}, p_{\theta_1}).$$

Proof. For $\theta \in \Theta$ and $z \in \mathsf{Z}^T$, define $h(\theta; z) := \sqrt{p_\theta(z)}$. We take the first derivative of $\theta \mapsto h(\theta; z)$:

$$\nabla_\theta h(\theta; z) = \frac{1}{2} \sqrt{p_\theta(z)} \nabla_\theta \log p_\theta(z).$$

The integral form of Taylor's theorem yields for $\mu^{\otimes T}$ a.e. $z \in \mathsf{Z}^T$:

$$h(\theta_1; z) = h(\theta_0; z) + \int_0^1 \langle \nabla_\theta h(\theta(s); z), \theta_1 - \theta_0 \rangle ds, \quad \theta(s) := (1-s)\theta_0 + s\theta_1.$$

Hence, we have, overloading $\mu = \mu^{\otimes T}$,

$$
\begin{aligned}
d_H^2(p_{\theta_0}, p_{\theta_1}) &= \int (h(\theta_1; z) - h(\theta_0; z))^2 d\mu \\
&= \int \left(\int_0^1 \langle \nabla_\theta h(\theta(s); z), \theta_1 - \theta_0 \rangle ds \right)^2 d\mu \\
&\leq \int \int_0^1 (\langle \nabla_\theta h(\theta(s); z), \theta_1 - \theta_0 \rangle)^2 ds d\mu && \text{[Jensen's inequality]} \\
&= \int_0^1 \int (\langle \nabla_\theta h(\theta(s); z), \theta_1 - \theta_0 \rangle)^2 d\mu ds && \text{[Fubini's lemma]} \\
&= \frac{1}{4} \int_0^1 \int (\langle \nabla_\theta \log p_{\theta(s)}(z), \theta_1 - \theta_0 \rangle)^2 p_{\theta(s)}(z) d\mu ds \\
&= \frac{1}{4} d_{\mathrm{FI}}^2(p_{\theta_0}, p_{\theta_1}).
\end{aligned}
$$

$\square$

Proposition A.8 (Hellinger identifiability under general conditions). *In addition to the assumptions on $\mathcal{P} = \{p_\theta \mid \theta \in \Theta\}$ stated in Section 3.1, assume that Θ is compact, $\mathcal{I}(\theta_\star)$ has full rank, and that the map $\theta \mapsto d_H^2(\theta, \theta_\star)$ has Lipschitz Hessians. Then, $\mathcal{P}$ is (γ_1, γ_2)-identifiable (cf. Definition 3.11) for some positive γ_1, γ_2.*

Proof. For every $\delta > 0$, we define the set $\Theta_\delta := \Theta \cap \{\theta \in \mathbb{R}^p \mid \|\theta - \theta_\star\| \geqslant \delta\}$. As Θ_δ is the intersection of two closed sets, Θ_δ is closed. Furthermore, since $\Theta_\delta \subset \Theta$, it is also bounded, and hence compact. Define $r(\theta) := \mathrm{d}_H^2(\theta, \theta_\star)$, and $r_\delta := \inf\{r(\theta) \mid \theta \in \Theta_\delta\}$. Since r_δ is the infimum of a continuous function over a compact set, the infimum is achieved by some $\theta_\delta \in \Theta_\delta$. Furthermore, since $\theta_\delta \neq \theta_\star$ by definition, we must have that $r_\delta = r(\theta_\delta) > 0$ (since $r(\theta) = 0$ iff $p_\theta = p_{\theta_\star}$, and we assumed that $p_{\theta_\star}$ is uniquely represented in $\mathcal{P}$, i.e., $p_\theta = p_{\theta_\star}$ iff $\theta = \theta_\star$). Hence, this yields the conclusion:

$$\forall \theta \in \Theta, \ r(\theta) \leqslant r_\delta/2 \Longrightarrow \|\theta - \theta_\star\| \leqslant \delta.$$

By the star-convexity assumption of Θ around $\theta_\star$, the function $r(\theta)$ is C^2 on the entire ray between θ and $\theta_\star$, and hence by the integral form of Taylor's theorem we obtain with $\Delta := \theta - \theta_\star$:

$$r(\theta) = r(\theta_\star) + \langle \nabla r(\theta_\star), \Delta \rangle + \int_0^1 (1-t)\Delta^\mathsf{T} \nabla^2 r((1-t)\theta_\star + t\theta)\Delta \mathrm{d}t$$

$$\overset{(a)}{=} \frac{1}{2}\Delta^\mathsf{T}\nabla^2 r(\theta_\star)\Delta + \int_0^1 (1-t)\Delta^\mathsf{T}[\nabla^2 r((1-t)\theta_\star + t\theta) - \nabla^2 r(\theta_\star)]\Delta \mathrm{d}t$$

$$\overset{(b)}{\geqslant} \frac{1}{2}\Delta^\mathsf{T}\nabla^2 r(\theta_\star)\Delta - L\int_0^1 t(1-t)\|\Delta\|^3 \mathrm{d}t = \frac{1}{2}\Delta^\mathsf{T}\nabla^2 r(\theta_\star)\Delta - \frac{L}{6}\|\Delta\|^3.$$

where (a) holds since $r(\theta_\star) = 0$ and $\nabla r(\theta_\star) = 0$ and (b) holds by the assumption that $r(\theta)$ has Lipschitz Hessians. Suppose now that that $\nabla^2 r(\theta_\star)$ is non-degenerate; we will check this momentarily. Hence, we have:

$$\|\Delta\| \leqslant \delta_0 := \frac{3\lambda_{\min}(\nabla^2 r(\theta_\star))}{2} \Longrightarrow r(\theta) \geqslant \frac{\lambda_{\min}(\nabla^2 r(\theta_\star))}{4}\|\Delta\|^2.$$

We therefore have:

$$r(\theta) \leqslant r_{\delta_0/2} \Longrightarrow \|\Delta\| \leqslant \delta_0 \Longrightarrow \|\Delta\|^2 \leqslant \frac{4}{\lambda_{\min}(\nabla^2 r(\theta_\star))}r(\theta),$$

which shows the desired Hellinger identifiability. To finish the proof, we confirm that $\nabla^2 r(\theta_\star)$ is non-degenerate. A standard computation (as done in the proof of Proposition 3.9) shows that $\nabla^2 r(\theta_\star) = \frac{1}{2}\mathcal{I}(\theta_\star)$, from which the proof concludes. $\square$

Proposition A.9. *Suppose that Y is a zero-mean random variable satisfying for some $\theta \in (0,1)$ and $\alpha > 0$,*

$$\log \mathbb{E}[\exp(\lambda Y)] \leqslant h(\lambda) := \alpha \log\left(e^\lambda \theta + 1 - \theta\right) - \alpha\theta\lambda, \quad \lambda \in \mathbb{R}.$$

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

$$|Y| \leqslant 2\sqrt{2e\theta(1-\theta)\alpha\log(2/\delta)} + 2\log(2/\delta).$$

Proof. We first differentiate $h(\lambda)$ twice:

$$h'(\lambda) = \frac{\alpha\theta e^\lambda}{\theta e^\lambda + (1-\theta)} - \alpha\theta, \quad h''(\lambda) = \frac{\alpha\theta(1-\theta)e^\lambda}{(\theta e^\lambda + (1-\theta))^2}.$$

We next observe that:

$$|\lambda| \leqslant 1 \implies \frac{e^\lambda}{(\theta e^\lambda + (1-\theta))^2} \leqslant \min\left\{\frac{1}{\theta^2 e^\lambda}, \frac{e^\lambda}{(1-\theta)^2}\right\} \leqslant e \min\left\{\frac{1}{\theta^2}, \frac{1}{(1-\theta)^2}\right\} \leqslant 4e.$$

Hence by a second order Taylor expansion of $h(\lambda)$:

$$|\lambda| \leqslant 1 \implies h(\lambda) \leqslant h(0) + h'(0)\lambda + \frac{1}{2}\sup_{|c|\leqslant 1} h''(c)\lambda^2 \leqslant 2e\alpha\theta(1-\theta)\lambda^2.$$

Therefore Y is $(2\sqrt{e\alpha\theta(1-\theta)}, 1)$-sub-Exponential [see e.g., 109, Chapter 2], and hence from using a sub-Exponential tail bound [109, Proposition 2.9] we have with probability at least $1 - \delta$,

$$|Y| \leqslant \sqrt{8e\alpha\theta(1-\theta)\log(2/\delta)} + 2\log(2/\delta).$$

$\square$

B Additional Derivations for Two-State Markov Chain Example

Log-likelihood and FI matrix computations. We start with the expression for the log-likelihood. The conditional log-likelihood is:

$$\log p_\theta(z' \mid z) = \log\theta \cdot \mathbb{1}\{z' = z\} + \log(1-\theta) \cdot \mathbb{1}\{z' \neq z\}.$$

Hence, we have:

$$\log p_\theta(z_{1:T}) = \log\theta \cdot N_{\text{stay}}(z_{1:T}) + \log(1-\theta) \cdot N_{\text{switch}}(z_{1:T}) + \log\rho_1(z_1),$$

where $N_{\text{stay}}(z_{1:T}) := \sum_{t=1}^{T-1} \mathbb{1}\{z_{t+1} = z_t\}$ and $N_{\text{switch}}(z_{1:T}) := T - 1 - N_{\text{stay}}(z_{1:T})$. We immediately see that $\mathcal{P}$ is log-concave (cf. Definition 3.5). We next compute the first and second derivatives of the conditional log-likelihood:

$$\partial_\theta \log p_\theta(z' \mid z) = \frac{1}{\theta}\mathbb{1}\{z' = z\} - \frac{1}{1-\theta}\mathbb{1}\{z' \neq z\},$$

$$\partial_\theta^2 \log p_\theta(z' \mid z) = -\left[\frac{1}{\theta^2}\mathbb{1}\{z' = z\} + \frac{1}{(1-\theta)^2}\mathbb{1}\{z' \neq z\}\right].$$

Taking conditional expectations,

$$\mathbb{E}_{p_\theta}[\partial_\theta \log p_\theta(z_{t+1} \mid z_t) \mid z_t] = 0, \quad \mathbb{E}_{p_\theta}[\partial_\theta^2 \log p_\theta(z_{t+1} \mid z_t)] = -\frac{1}{\theta(1-\theta)}.$$

Hence, the FI matrix is:

$$\mathcal{I}(\theta) = -\mathbb{E}_{p_\theta}[\partial_\theta^2 \log p_\theta(z_{1:T})] = -\sum_{t=1}^{T-1}\mathbb{E}_{p_\theta}[\partial_\theta^2 \log p_\theta(z_{t+1} \mid z_t)] = \frac{T-1}{\theta(1-\theta)}.$$

Hence, for every $\theta \in \Theta$, we have:

$$\mathcal{I}(\theta) = \frac{T-1}{\theta(1-\theta)} \leqslant \frac{T-1}{\mu(1-\mu)} =: \mathcal{I}_{\max}.$$

We also have the bound $\mathrm{diam}(\Theta) \leqslant \frac{T-1}{\mu(1-\mu)}$.

Moment computations. We now turn to the moment computations of $\mathbb{E}_{p_\theta}[(\partial_\theta \log p_\theta(z_{1:T}))^4]$ and $\mathbb{E}_{p_\theta}[(\partial_\theta^2 \log p_\theta(z_{1:T}))^2]$. Using the expressions for the conditional log-likelihoods, we have that:

$$\partial_\theta \log p_\theta(z_{1:T}) = \frac{1}{\theta} N_{\text{stay}}(z_{1:T}) - \frac{1}{1-\theta}(T - 1 - N_{\text{stay}}(z_{1:T})),$$

$$\partial_\theta^2 \log p_\theta(z_{1:T}) = -\left[\frac{1}{\theta^2} N_{\text{stay}}(z_{1:T}) + \frac{1}{(1-\theta)^2}(T - 1 - N_{\text{stay}}(z_{1:T}))\right].$$

We also observe that $N_{\text{stay}}(z_{1:T}) \sim \text{Bin}(T - 1, \theta)$ when $z_{1:T} \sim p_\theta$. Hence, utilizing standard moment expressions for binomial distributions, it is straightforward to compute:

$$\mathbb{E}_{p_\theta}[(\partial_\theta \log p_\theta(z_{1:T}))^4] = (T - 1)\left(\frac{1}{\theta^3} + \frac{1}{(1-\theta)^3}\right) + 3(T - 1)(T - 2)\frac{1}{\theta^2(1-\theta)^2},$$

$$\mathbb{E}_{p_\theta}[(\partial_\theta^2 \log p_\theta(z_{1:T}))^2] = (T - 1)\left(\frac{1}{\theta^3} + \frac{1}{(1-\theta)^3}\right) + (T - 1)(T - 2)\frac{1}{\theta^2(1-\theta)^2}.$$

Direct analysis. Let $\psi^{(i)} := N_{\text{stay}}(z_{1:T}^{(i)})$ for $i \in [m]$ and $T' := T-1$. We know that $\psi^{(i)} \sim \text{Bin}(T', \theta_\star)$. Hence, its MGF is given as $\mathbb{E}[\exp(\lambda\psi^{(i)})] = (\theta_\star e^\lambda + (1 - \theta_\star))^{T'}$ for $\lambda \in \mathbb{R}$. Since $\psi^{(i)}$ are iid across $i \in [m]$, we have $\mathbb{E}[\exp(\lambda \sum_{i=1}^m \psi^{(i)})] = (\theta_\star e^\lambda + (1 - \theta_\star))^{mT'}$. Hence, defining $Y := \sum_{i=1}^m \psi^{(i)} - mT'\theta_\star$, we have that $\mathbb{E}[Y] = 0$ and

$$\log \mathbb{E}[\exp(\lambda Y)] = mT' \log(\theta_\star \varepsilon^\lambda + (1 - \theta_\star)) - mT'\theta_\star\lambda, \quad \lambda \in \mathbb{R}.$$

From Proposition A.9, with probability at least $1 - \delta$ (over $\mathcal{D}_{m,T}$), we have

$$|Y| \lesssim \sqrt{mT\sigma_\star^2 \log(2/\delta)} + \log(2/\delta).$$

Dividing both sides by mT', we have with probability at least $1 - \delta$,

$$|\hat{\theta}_{m,T} - \theta_\star| \lesssim \sqrt{\frac{\sigma_\star^2 \log(1/\delta)}{mT}} + \frac{\log(1/\delta)}{mT}.$$

Call this event $\mathcal{E}_2$. From this, we finally obtain:

$$mT \gtrsim \sigma_\star^{-2} \log(1/\delta) \implies |\hat{\theta}_{m,T} - \theta_\star|^2 \lesssim \frac{\sigma_\star^2 \log(1/\delta)}{mT} \text{ on } \mathcal{E}_2,$$

which is the claimed rate in (3.37).

C   Additional Results for Mixture of Two-State Markov Chains

Proposition C.1. *Given real matrix $A \in \mathbb{R}^{kd \times kd}$ with*

$$A = \begin{pmatrix} A_{11} & \cdots & A_{1k} \\ \vdots & \ddots & \vdots \\ A_{k1} & \cdots & A_{kk} \end{pmatrix}, \quad \forall\ 1 \leqslant i, j \leqslant k,\ A_{ij} \in \mathbb{R}^{d \times d},$$

and a real symmetric matrix $A' \in \mathbb{R}^{kd \times kd}$ with

$$A' = \begin{pmatrix} A'_{11} & \cdots & A'_{1k} \\ \vdots & \ddots & \vdots \\ A'^{\mathsf{T}}_{1k} & \cdots & A_{kk} \end{pmatrix}, \quad \forall\, 1 \leqslant i, j \leqslant k,\; A'_{ij} \in \mathbb{R}^{d \times d},\; A'_{ii} \text{ symmetric,}$$

such that for any $1 \leqslant i, j \leqslant d$, $\|A_{ij}\|_{\mathrm{op}} \leqslant \|A'_{ij}\|_{\mathrm{op}}$, we have

$$A \preccurlyeq \mathsf{blk\text{-}diag}\left\{ \left(\sum_{j=1}^{k} \|A'_{1j}\|_{\mathrm{op}} \right) I_d,\; \ldots,\; \left(\sum_{j=1}^{k} \|A'_{kj}\|_{\mathrm{op}} \right) I_d \right\}.$$

Proof. Given a test vector $v \in \mathbb{R}^{kd}$, we decompose v into corresponding blocks:

$$v = \begin{pmatrix} v_1^{\mathsf{T}} & \ldots & v_k^{\mathsf{T}} \end{pmatrix}^{\mathsf{T}}, \quad \forall\, 1 \leqslant i \leqslant k,\; v_i \in \mathbb{R}^d.$$

Then we have:

$$
\begin{aligned}
v^{\mathsf{T}} A v &= \sum_{i=1}^{k} v_i^{\mathsf{T}} A_{ii} v_i + 2 \sum_{1 \leqslant i < j \leqslant k} v_i^{\mathsf{T}} A_{ij} v_j \\
&\overset{(a)}{\leqslant} \sum_{i=1}^{k} \|A'_{ii}\|_{\mathrm{op}} \|v_i\|^2 + \sum_{1 \leqslant i < j \leqslant k} \|A'_{ij}\|_{\mathrm{op}} \|v_i\| \|v_j\| \\
&\overset{(b)}{\leqslant} \sum_{i=1}^{k} \|A'_{ii}\|_{\mathrm{op}} \|v_i\|^2 + \sum_{1 \leqslant i < j \leqslant k} \|A'_{ij}\|_{\mathrm{op}} \left(\|v_i\|^2 + \|v_j\|^2 \right) \\
&\overset{(c)}{\leqslant} \sum_{i=1}^{k} \|A'_{ii}\|_{\mathrm{op}} \|v_i\|^2 + \sum_{1 \leqslant i \neq j \leqslant k} \|A'_{ij}\|_{\mathrm{op}} \|v_i\|^2 \\
&= \sum_{i=1}^{k} \left(\sum_{j=1}^{k} \|A'_{ij}\|_{\mathrm{op}} \right) \|v_i\|^2 \\
&= v^{\mathsf{T}} \mathsf{blk\text{-}diag}\left\{ \left(\sum_{j=1}^{k} \|A'_{1j}\|_{\mathrm{op}} \right) I_d,\; \ldots,\; \left(\sum_{j=1}^{k} \|A'_{kj}\|_{\mathrm{op}} \right) I_d \right\} v,
\end{aligned}
$$

where for (a) we applied the Cauchy-Schwarz inequality, for (b) we applied the AM-GM inequality, and for (c) we used the fact that for any square matrix M, $\|M\|_{\mathrm{op}} = \|M^{\mathsf{T}}\|_{\mathrm{op}}$. $\qquad\square$

Corollary C.2. *Given real matrix $A \in \mathbb{R}^{d \times d}$ and real symmetric matrix $A' \in \mathbb{R}^{d \times d}$ such that for any $1 \leqslant i, j \leqslant d$, $|A_{ij}| \leqslant A'_{ij}$, we have*

$$A \preccurlyeq \mathsf{diag}\left\{ \sum_{j=1}^{d} A'_{1j},\; \ldots,\; \sum_{j=1}^{d} A'_{dj} \right\}.$$

Proof. This is a special case of Proposition C.1 with $k = 1$. $\qquad\square$

Proposition C.3. *Given positive definite matrices $A_1,\, A_2 \in \mathbb{R}^{d \times d}$, a vector-valued function $M_1 \colon \mathbb{X} \to \mathbb{R}^d$, and a matrix valued function $M_2 \colon \mathbb{X} \to \mathbb{R}^{d \times d}$. Assume that $A_1 \succcurlyeq A_2$. Let $\psi \colon \mathbb{R}^{\mathbb{X}} \to \mathbb{R}_{\geqslant 0}$ be a non-negative linear functional. Then we have*

$$\sup_{\|v\|=1} \psi\left(v^{\mathsf{T}} A_1^{-1/2} M_1(x) \right) \leqslant \sup_{\|v\|=1} \psi\left(v^{\mathsf{T}} A_2^{-1/2} M_1(x) \right),$$

$$\sup_{\|v\|=1} \psi\left(v^\mathsf{T} A_1^{-1/2} M_2(x) A_1^{-1/2} v\right) \leqslant \sup_{\|v\|=1} \psi\left(v^\mathsf{T} A_2^{-1/2} M_2(x) A_2^{-1/2} v\right).$$

In particular, we have for any $p, q > 0$, suppose

$$\sup_{\|v\|=1} \left\|v^\mathsf{T} A_2^{-1/2} M_1(x)\right\|_{\mathcal{L}^p(\mu)} < \infty, \quad \sup_{\|v\|=1} \left\|v^\mathsf{T} A_2^{-1/2} M_2(x) A_2^{-1/2} v\right\|_{\mathcal{L}^q(\mu)} < \infty,$$

then,

$$\sup_{\|v\|=1} \left\|v^\mathsf{T} A_1^{-1/2} M_1(x)\right\|_{\mathcal{L}^p(\mu)} \leqslant \sup_{\|v\|=1} \left\|v^\mathsf{T} A_2^{-1/2} M_1(x)\right\|_{\mathcal{L}^p(\mu)},$$

$$\sup_{\|v\|=1} \left\|v^\mathsf{T} A_1^{-1/2} M_2(x) A_1^{-1/2} v\right\|_{\mathcal{L}^q(\mu)} \leqslant \sup_{\|v\|=1} \left\|v^\mathsf{T} A_2^{-1/2} M_2(x) A_2^{-1/2} v\right\|_{\mathcal{L}^q(\mu)}.$$

Proof. Denote $u := A_1^{-1/2} v$, by symmetricity, we have

$$\begin{aligned}
\sup_{\|v\|=1} \psi\left(v^\mathsf{T} A_1^{-1/2} M_1(x)\right) &= \sup_{\|v\|=1} \psi\left(\left(A_1^{-1/2} v\right)^\mathsf{T} M_1(x)\right) \\
&= \sup_{\|A_1^{1/2} u\|=1} \psi\left(u^\mathsf{T} M_1(x)\right) \\
&= \sup_{\|u\|_{A_1}=1} \psi\left(u^\mathsf{T} M_1(x)\right) \\
&= \sup_{u} \psi\left(\left(\frac{u}{\|u\|_{A_1}}\right)^\mathsf{T} M_1(x)\right) \\
&= \sup_{u} \frac{\psi\left(u^\mathsf{T} M_1(x)\right)}{\sqrt{u^\mathsf{T} A_1 u}} \\
&\leqslant \sup_{u} \frac{\psi\left(u^\mathsf{T} M_1(x)\right)}{\sqrt{u^\mathsf{T} A_2 u}} = \sup_{\|v\|=1} \psi\left(v^\mathsf{T} A_2^{-1/2} M_1(x)\right),
\end{aligned}$$

and similarly

$$\begin{aligned}
\sup_{\|v\|=1} \psi\left(v^\mathsf{T} A_1^{-1/2} M_2(x) A_1^{-1/2} v\right) &= \sup_{\|v\|=1} \psi\left(\left(A_1^{-1/2} v\right)^\mathsf{T} M_2(x) \left(A_1^{-1/2} v\right)\right) \\
&= \sup_{\|A_1^{1/2} u\|=1} \psi\left(u^\mathsf{T} M_2(x) u\right) \\
&= \sup_{\|u\|_{A_1}=1} \psi\left(u^\mathsf{T} M_2(x) u\right) \\
&= \sup_{u} \psi\left(\left(\frac{u}{\|u\|_{A_1}}\right)^\mathsf{T} M_2(x) \left(\frac{u}{\|u\|_{A_1}}\right)\right) \\
&= \sup_{u} \frac{\psi\left(u^\mathsf{T} M_2(x) u\right)}{u^\mathsf{T} A_1 u} \\
&\leqslant \sup_{u} \frac{\psi\left(u^\mathsf{T} M_2(x) u\right)}{u^\mathsf{T} A_2 u} = \sup_{\|v\|=1} \psi\left(v^\mathsf{T} A_2^{-1/2} M_2(x) A_1^{-1/2} v\right).
\end{aligned}$$

$\square$

Proposition C.4. *Let Δ be a $d \times d$ matrix, and M, N be two $d \times d$ positive definite matrices satisfying $M \succcurlyeq N$. We have:*

$$\|M^{-1/2}\Delta M^{-1/2}\|_{\mathrm{op}} \le \|N^{-1/2}\Delta N^{-1/2}\|_{\mathrm{op}}.$$

Proof. Observe that:

$$
\begin{aligned}
\|M^{-1/2}\Delta M^{-1/2}\|_{\mathrm{op}}^2 &= \lambda_{\max}(M^{-1/2}\Delta^{\mathsf{T}}M^{-1}\Delta M^{-1/2}) \\
&\overset{(a)}{\le} \lambda_{\max}(M^{-1/2}\Delta^{\mathsf{T}}N^{-1}\Delta M^{-1/2}) \\
&= \lambda_{\max}(M^{-1/2}\Delta^{\mathsf{T}}N^{-1/2}\cdot N^{-1/2}\Delta M^{-1/2}) \\
&\overset{(b)}{=} \lambda_{\max}(N^{-1/2}\Delta M^{-1/2}\cdot M^{-1/2}\Delta^{\mathsf{T}}N^{-1/2}) \\
&= \lambda_{\max}(N^{-1/2}\Delta M^{-1}\Delta^{\mathsf{T}}N^{-1/2}) \\
&\overset{(c)}{\le} \lambda_{\max}(N^{-1/2}\Delta N^{-1}\Delta^{\mathsf{T}}N^{-1/2}) \\
&= \|N^{-1/2}\Delta N^{-1/2}\|_{\mathrm{op}}^2,
\end{aligned}
$$

where (a) follows since $M \succcurlyeq N > 0$ implies $N^{-1} \succcurlyeq M^{-1}$ [see e.g., 110, Prop. V.1.6] and so conjugating both sides of the latter inequality by $\Delta M^{-1/2}$ yields $M^{-1/2}\Delta^{\mathsf{T}}M^{-1}\Delta M^{-1/2} \preccurlyeq M^{-1/2}\Delta^{\mathsf{T}}N^{-1}\Delta M^{-1/2}$, (b) uses $\lambda(AB) = \lambda(BA)$ for two square matrices A, B, where $\lambda(\cdot)$ refers to spectrum of its argument and (c) uses the same argument as (a). $\qquad\square$

D Hellinger Identifiability for Sinusoidal GLMs

Here we cover the necessary Gaussian anti-concentration results and local identifibility needed in the proof of Theorem 4.17. We believe these set of results to be of independent interest.

Proposition D.1. *Let $a \in \mathbb{R}$, $t \in (0, 1)$, and $\sigma > 0$. We have for $g \sim \mathsf{N}(0, 1)$,*

$$\mathbb{P}_g(|\cos(\sigma g + a)| \le t) \le \left(1 + \frac{3\sqrt{\pi/2}}{\sigma}\right)\left(1 - \frac{2\cos^{-1}(t)}{\pi}\right).$$

Hence there exists a universal constant c_0 such that:

$$\forall t > 0, \ \mathbb{P}_g(|\cos(\sigma g + a)| \le t) \le c_0 \max\{1, 1/\sigma\}t.$$

Note that since $\cos(x) = \sin(x + \pi/2)$, the same result also holds for sin *in place of* cos.

Proof. Fix a $t \in (0, 1)$. For $k \in \mathbb{Z}$, define the interval:

$$I_k := \left[k\pi + \cos^{-1}(t), (k+1)\pi - \cos^{-1}(t)\right].$$

Since these intervals are disjoint and their union covers the t-sub-level set of $|\cos(x)|$, i.e.,

$$\bigsqcup_{k\in\mathbb{Z}} I_k = \{x \in \mathbb{R} \mid |\cos(x)| \le t\},$$

95

we have that, letting $X \sim \mathsf{N}(a, \sigma^2)$,

$$\mathbb{P}_g(|\cos(\sigma g + a)| \leqslant t) = \mathbb{P}\left(X \in \bigsqcup_{k \in \mathbb{Z}} I_k\right) = \sum_{k \in \mathbb{Z}} \mathbb{P}(X \in I_k) = \sum_{k \in \mathbb{Z}} \int_{I_k} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-a)^2}{2\sigma^2}} \, \mathrm{d}x.$$

Define $k_0 := \inf\{k \in \mathbb{Z} \mid k\pi + \cos^{-1}(t) \geqslant a\}$, and $\phi_k(x) := x - (k - k_0)\pi$ for $k \geqslant k_0$. Now for any $k \geqslant k_0$ and $x \in I_k$, we have that $\phi_k(x) \in I_{k_0}$, and furthermore:

$$\begin{aligned}
\frac{\exp(-(x-a)^2/(2\sigma^2))}{\exp(-(\phi_k(x)-a)^2/(2\sigma^2))} &= \exp\left(-\frac{1}{2\sigma^2}\left[(x-a)^2 - (\phi_k(x) - a)^2\right]\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left[((x-a) + (\phi_k(x) - a))(x - \phi_k(x))\right]\right) \\
&= \exp\left(-\frac{(k-k_0)\pi}{2\sigma^2}((x-a) + (\phi_k(x) - a))\right) \\
&\leqslant \exp\left(-\frac{(k-k_0)^2\pi^2}{2\sigma^2}\right).
\end{aligned}$$

The last inequality holds since: (a) we know that $\phi_k(x) - a \geqslant 0$ since $\phi_k(x) \in I_{k_0}$ and every $x \in I_{k_0}$ satisfies $x \geqslant a$ by definition, and (b) since $x \in I_k$ and $k \geqslant k_0$, we know that $x - a = (k-k_0)\pi + \phi_k(x) - a \geqslant (k - k_0)\pi$. Hence, for each $k \geqslant k_0$,

$$\begin{aligned}
\mathbb{P}(X \in I_k) &= \int_{I_k} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-a)^2}{2\sigma^2}} \, \mathrm{d}x \leqslant e^{-(k-k_0)^2\pi^2/(2\sigma^2)} \int_{I_k} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\phi_k(x)-a)^2}{2\sigma^2}} \, \mathrm{d}x \\
&= e^{-(k-k_0)^2\pi^2/(2\sigma^2)} \mathbb{P}(X \in I_{k_0}) \leqslant e^{-(k-k_0)^2\pi^2/(2\sigma^2)} \sup_{k \in \mathbb{Z}} \mathbb{P}(X \in I_k).
\end{aligned}$$

We now consider the case when $k \leqslant \bar{k}_0 := k_0 - 2$. We next define $\bar{\phi}_k(x) := x + (\bar{k}_0 - k)\pi$ for $k \leqslant \bar{k}_0$. Similar to before, we have that for any $k \leqslant \bar{k}_0$ and $x \in I_k$, $\bar{\phi}_k(x) \in I_{\bar{k}_0}$. Also similar to before, we can show that:

$$\frac{\exp(-(x-a)^2/(2\sigma^2))}{\exp(-(\bar{\phi}_k(x)-a)^2/(2\sigma^2))} \leqslant \exp\left(-\frac{(\bar{k}_0 - k)^2\pi^2}{2\sigma^2}\right),$$

and hence for $k \leqslant \bar{k}_0$,

$$\mathbb{P}(X \in I_k) \leqslant e^{-(\bar{k}_0 - k)^2\pi^2/(2\sigma^2)} \mathbb{P}(X \in I_{\bar{k}_0}) \leqslant e^{-(\bar{k}_0 - k)^2\pi^2/(2\sigma^2)} \sup_{k \in \mathbb{Z}} \mathbb{P}(X \in I_k).$$

Consequently,

$$\begin{aligned}
\sum_{k \in \mathbb{Z}} \mathbb{P}(X \in I_k) &= \sum_{k \leqslant \bar{k}_0} \mathbb{P}(X \in I_k) + \sum_{k \geqslant k_0} \mathbb{P}(X \in I_k) + \mathbb{P}(X \in I_{k_0 - 1}) \\
&\leqslant \sup_{k \in \mathbb{Z}} \mathbb{P}(X \in I_k) \left[1 + \sum_{k \leqslant \bar{k}_0} e^{-(\bar{k}_0 - k)^2\pi^2/(2\sigma^2)} + \sum_{k \geqslant k_0} e^{-(k-k_0)^2\pi^2/(2\sigma^2)}\right] \\
&= \sup_{k \in \mathbb{Z}} \mathbb{P}(X \in I_k) \left[1 + 2\sum_{k \in \mathbb{N}} e^{-k^2\pi^2/(2\sigma^2)}\right].
\end{aligned}$$

Next since $x \mapsto e^{-x^2\pi^2/(2\sigma^2)}$ is decreasing on $\mathbb{R}_{\geqslant 0}$ we have that

$$\sum_{k\in\mathbb{N}} e^{-k^2\pi^2/(2\sigma^2)} = 1 + \sum_{k\geqslant 1} e^{-k^2\pi^2/(2\sigma^2)} \leqslant 1 + \int_0^\infty e^{-x^2\pi^2/(2\sigma^2)}\mathrm{d}x = 1 + \frac{\sigma}{\sqrt{2\pi}}.$$

Since $|I_k| = \pi - 2\cos^{-1}(t)$, we also have

$$\sup_{k\in\mathbb{Z}} \mathbb{P}(X \in I_k) \leqslant \frac{\pi - 2\cos^{-1}(t)}{\sqrt{2\pi}\sigma}.$$

Therefore,

$$\sum_{k\in\mathbb{Z}} \mathbb{P}(X \in I_k) \leqslant \left(3 + \frac{2\sigma}{\sqrt{2\pi}}\right)\frac{\pi - 2\cos^{-1}(t)}{\sqrt{2\pi}\sigma} = \left(1 + \frac{3\sqrt{\pi/2}}{\sigma}\right)\left(1 - \frac{2\cos^{-1}(t)}{\pi}\right).$$

$\square$

The following result is similar to [38, Lemma 10], except for sin instead of ReLU activations.

Proposition D.2. *Let $u_1, u_2 \in \mathbb{R}^d$. There exists a universal positive constants γ_0, c_0 such that for all $\gamma \in [0, \gamma_0]$,*

$$\mathbb{E}_{z\sim\mathsf{N}(0,\sigma^2 I_d)}\big[(\sin(\langle u_1, z\rangle) - \sin(\langle u_2, z\rangle))^2\big] \leqslant \gamma^2 \implies \|u_1 - u_2\|^2 \leqslant \frac{c_0\gamma^2}{\sigma^2}.$$

Proof. We first use the identity

$$\sin(\langle u_1, z\rangle) - \sin(\langle u_2, z\rangle) = 2\cos\left(\frac{\langle u_1 + u_2, z\rangle}{2}\right)\sin\left(\frac{\langle u_1 - u_2, z\rangle}{2}\right),$$

so that

$$(\sin(\langle u_1, z\rangle) - \sin(\langle u_2, z\rangle))^2 = 4\cos^2\left(\frac{\langle u_1 + u_2, z\rangle}{2}\right)\sin^2\left(\frac{\langle u_1 - u_2, z\rangle}{2}\right).$$

We fix $\delta = 1/4$ and consider two cases.

Case $\sigma\|u_1 + u_2\| \leqslant 1/\sqrt{2\log(2/\delta)}$. We define two events:

$$\mathcal{E}_1 := \{|\langle u_1 + u_2, z\rangle| \leqslant \sigma\|u_1 + u_2\|\sqrt{2\log(2/\delta)}\},$$
$$\mathcal{E}_2 := \{|\sin(\langle u_1 - u_2, z\rangle/2)| \geqslant \delta/c_0 \cdot \min\{1, \sigma/2 \cdot \|u_1 - u_2\|\}\},$$

where c_0 is from Proposition D.1. By standard Gaussian concentration results, we know that $\mathbb{P}(\mathcal{E}_1^c) \leqslant \delta$. From Proposition D.1 we also know that $\mathbb{P}(\mathcal{E}_2^c) \leqslant \delta$. By a union bound, we have $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geqslant 1 - 2\delta = 1/2$. Putting these together,

$$\mathbb{E}_{z\sim\mathsf{N}(0,\sigma^2 I_d)}\big[(\sin(\langle u_1, z\rangle) - \sin(\langle u_2, z\rangle))^2\big]$$
$$= 4\mathbb{E}_{z\sim\mathsf{N}(0,\sigma^2 I_d)}\left[\cos^2\left(\frac{\langle u_1 + u_2, z\rangle}{2}\right)\sin^2\left(\frac{\langle u_1 - u_2, z\rangle}{2}\right)\right]$$
$$\geqslant 4\mathbb{E}_{z\sim\mathsf{N}(0,\sigma^2 I_d)}\left[\cos^2\left(\frac{\langle u_1 + u_2, z\rangle}{2}\right)\sin^2\left(\frac{\langle u_1 - u_2, z\rangle}{2}\right)\mathbb{1}\{\mathcal{E}_1 \cap \mathcal{E}_2\}\right]$$
$$\geqslant 4\mathbb{E}_{z\sim\mathsf{N}(0,\sigma^2 I_d)}\left[\cos^2(1/2)(\delta/c_0)^2\min\{1, \sigma^2/4 \cdot \|u_1 - u_2\|^2\}\mathbb{1}\{\mathcal{E}_1 \cap \mathcal{E}_2\}\right]$$
$$\geqslant 2\cos^2(1/2)(\delta/c_0)^2\min\{1, \sigma^2/4 \cdot \|u_1 - u_2\|^2\}.$$

97

Case $\sigma\|u_1 + u_2\| > 1/\sqrt{2\log(2/\delta)}$. In this case, we define two events:

$$\mathcal{E}_1 := \{|\cos(\langle u_1 + u_2, z\rangle/2)| \geqslant \delta/c_0 \cdot \min\{1, \sigma/2 \cdot \|u_1 + u_2\|\}\},$$
$$\mathcal{E}_2 := \{|\sin(\langle u_1 - u_2, z\rangle/2)| \geqslant \delta/c_0 \cdot \min\{1, \sigma/2 \cdot \|u_1 - u_2\|\}\},$$

where again c_0 is from Proposition D.1. Using Proposition D.1 and a union bound, we have $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_1) \geqslant 1/2$. Furthermore,

$$\mathbb{E}_{z\sim\mathsf{N}(0,\sigma^2 I_d)}[(\sin(\langle u_1, z\rangle) - \sin(\langle u_2, z\rangle))^2]$$
$$\geqslant 4\mathbb{E}_{z\sim\mathsf{N}(0,\sigma^2 I_d)}\left[\cos^2\left(\frac{\langle u_1 + u_2, z\rangle}{2}\right)\sin^2\left(\frac{\langle u_1 - u_2, z\rangle}{2}\right)\mathbb{1}\{\mathcal{E}_1 \cap \mathcal{E}_2\}\right]$$
$$\geqslant 4(\delta/c_0)^4 \min\{1, \sigma^2/4 \cdot \|u_1 + u_2\|^2\} \min\{1, \sigma^2/4 \cdot \|u_1 - u_2\|^2\}\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2)$$
$$\geqslant 2(\delta/c_0)^4 \min\{1, 1/(8\log(2/\delta))\} \min\{1, \sigma^2/4 \cdot \|u_1 - u_2\|^2\}.$$

Combining both cases. Combining both cases, we have that:

$$\mathbb{E}_{z\sim\mathsf{N}(0,\sigma^2 I_d)}[(\sin(\langle u_1, z\rangle) - \sin(\langle u_2, z\rangle))^2] \geqslant c_1 \min\{1, \sigma^2/4 \cdot \|u_1 - u_2\|^2\},$$

where $c_1 > 0$ is a universal constant (recall that $\delta = 1/4$ is fixed). Hence, for any $\gamma^2 \leqslant c_1/2$, we must have that

$$\mathbb{E}_{z\sim\mathsf{N}(0,\sigma^2 I_d)}[(\sin(\langle u_1, z\rangle) - \sin(\langle u_2, z\rangle))^2] \leqslant \gamma^2 \implies \min\{1, \sigma^2/4 \cdot \|u_1 - u_2\|^2\} = \sigma^2/4 \cdot \|u_1 - u_2\|^2,$$

otherwise we would have the contradiction $c_1/2 \geqslant \gamma^2 \geqslant c_1$. Hence, we conclude

$$\mathbb{E}_{z\sim\mathsf{N}(0,\sigma^2 I_d)}[(\sin(\langle u_1, z\rangle) - \sin(\langle u_2, z\rangle))^2] \leqslant \gamma^2 \implies \|u_1 - u_2\|^2 \leqslant \frac{4\gamma^2}{c_1 \sigma^2}.$$

$\square$

Fact D.3 (Hellinger distance for multivariate Gaussians [cf. 111]). *Let $\mathsf{N}(\mu_i, \Sigma_i)$ for $i \in \{1,2\}$ be two multivariate Gaussians in $\mathbb{R}^d$. The squared Hellinger distance has the following closed-form expression:*

$$\frac{1}{2}\mathrm{d}_H^2(\mathsf{N}(\mu_1, \Sigma_1), \mathsf{N}(\mu_2, \Sigma_2)) = 1 - \frac{\det(\Sigma_1)^{1/4}\det(\Sigma_2)^{1/4}}{\det((\Sigma_1 + \Sigma_2)/2)^{1/2}}\exp\left\{-\frac{1}{8}(\mu_1 - \mu_2)^\mathsf{T}\left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1}(\mu_1 - \mu_2)\right\}.$$

Hence, a special case when $\Sigma_1 = \Sigma_2 = \sigma^2 I_d$ is:

$$\frac{1}{2}\mathrm{d}_H^2(\mathsf{N}(\mu_1, \sigma^2 I_d), \mathsf{N}(\mu_2, \sigma^2 I_d)) = 1 - \exp\left(-\frac{1}{8\sigma^2}\|\mu_1 - \mu_2\|^2\right).$$

Proposition D.4. *Fix parameters $A_1, A_2 \in \mathbb{R}^{d\times d}$ and let $\theta_i = \mathrm{vec}(A_i) \in \mathbb{R}^{d^2}$ for $i \in \{1,2\}$. For any $\gamma \geqslant 0$, we have that $\mathrm{d}_H^2(p_{\theta_1}(z_{1:2}), p_{\theta_2}(z_{1:2})) \leqslant \gamma^2$ implies the following bound holds:*

$$\max_{j\in[d]}\mathbb{E}_{z_1\sim\mathsf{N}(0,\sigma^2 I_d)}[(\sin(\langle A_1[j], z_1\rangle) - \sin(\langle A_2[j], z_1\rangle))^2] \leqslant 4\max\{2\sigma^2, 1\}\gamma^2.$$

Here, $A_i[j] \in \mathbb{R}^d$ denotes the j-th row of A_i.

Proof. Since $z_1 \sim \mathsf{N}(0, \sigma^2 I_d)$ regardless of θ, we have using Fact D.3,

$$\frac{1}{2}\mathrm{d}_H^2(p_{\theta_1}(z_{1:2}), p_{\theta_2}(z_{1:2})) = \frac{1}{2}\mathbb{E}_{z_1}[\mathrm{d}_H^2(\mathsf{N}(\sin(A_1 z_1), \sigma^2 I_d), \mathsf{N}(\sin(A_2 z_1), \sigma^2 I_d))]$$

$$= 1 - \mathbb{E}_{z_1}\left[\exp\left(-\frac{1}{8\sigma^2}\|\sin(A_1 z_1) - \sin(A_2 z_1)\|^2\right)\right].$$

Hence,

$$\mathrm{d}_H^2(p_{\theta_1}(z_{1:2}), p_{\theta_2}(z_{1:2})) \leqslant \gamma^2 \iff 1 - \frac{\gamma^2}{2} \leqslant \mathbb{E}_{z_1}\left[\exp\left(-\frac{1}{8\sigma^2}\|\sin(A_1 z_1) - \sin(A_2 z_1)\|^2\right)\right].$$

Let $c = \max\{8, 4/\sigma^2\}$, $x \in [0, 2]$, and observe that by the inequality $\exp(-x) \leqslant 1 - x + x^2/2$ which is valid for all $x \geqslant 0$,

$$\exp\left(-\frac{x^2}{8\sigma^2}\right) \leqslant \exp\left(-\frac{x^2}{c\sigma^2}\right) \leqslant 1 - \frac{x^2}{c\sigma^2} + \frac{x^4}{2c^2\sigma^4} \leqslant 1 - \frac{x^2}{c\sigma^2} + \frac{2x^2}{c^2\sigma^4}$$

$$= 1 - \frac{x^2}{c\sigma^2}\left(1 - \frac{2}{c\sigma^2}\right) \leqslant 1 - \frac{x^2}{2c\sigma^2}.$$

Fixing any index $j_0 \in [d]$, we now observe that:

$$\mathbb{E}_{z_1}\left[\exp\left(-\frac{1}{8\sigma^2}\|\sin(A_1 z_1) - \sin(A_2 z_1)\|^2\right)\right]$$

$$\leqslant \mathbb{E}_{z_1}\left[\exp\left(-\frac{1}{8\sigma^2}(\sin(\langle A_1[j_0], z_1\rangle) - \sin(\langle A_2[j_0], z_1\rangle))^2\right)\right].$$

$$\leqslant 1 - \frac{1}{2c\sigma^2}\mathbb{E}_{z_1}[(\sin(\langle A_1[j_0], z_1\rangle) - \sin(\langle A_2[j_0], z_1\rangle))^2].$$

From this, we conclude that

$$\mathrm{d}_H^2(p_{\theta_1}(z_{1:2}), p_{\theta_2}(z_{1:2})) \leqslant \gamma^2 \implies \mathbb{E}_{z_1}[(\sin(\langle A_1[j_0], z_1\rangle) - \sin(\langle A_2[j_0], z_1\rangle))^2] \leqslant c\sigma^2\gamma^2.$$

Since $j_0 \in [d]$ is arbitrary, the claim follows. $\qquad\square$

Proposition D.5. *Fix parameters* $A_1, A_2 \in \mathbb{R}^{d \times d}$ *and let* $\theta_i = \mathrm{vec}(A_i) \in \mathbb{R}^{d^2}$ *for* $i \in \{1, 2\}$. *There exists universal positive constants* γ_0, c_0 *such that for all* $\gamma \in [0, \gamma_0/\max\{1, \sigma\}]$,

$$\mathrm{d}_H^2(p_{\theta_1}(z_{1:2}), p_{\theta_2}(z_{1:2})) \leqslant \gamma^2 \implies \max_{j \in [d]}\|A_1[j] - A_2[j]\|_F^2 \leqslant c_0 \max\{1, 1/\sigma^2\}\gamma^2.$$

Proof. Given the condition $\mathrm{d}_H^2(p_{\theta_1}(z_{1:2}), p_{\theta_2}(z_{1:2})) \leqslant \gamma^2$, from Proposition D.4 for every $j \in [d]$,

$$\mathbb{E}_{z_1 \sim \mathsf{N}(0, \sigma^2 I_d)}[(\sin(\langle A_1[j], z_1\rangle) - \sin(\langle A_2[j], z_1\rangle))^2] \lesssim \max\{1, \sigma^2\}\gamma^2.$$

Next, from Proposition D.2, this implies that for every $j \in [d]$,

$$\|A_1[j] - A_2[j]\|^2 \lesssim \max\{1, 1/\sigma^2\}\gamma^2.$$

$\qquad\square$

E Additional Results for Sequence Modeling

Proposition E.1. *Suppose that for $\Psi : \mathbb{R}^d \mapsto \mathbb{R}^{d-1}$, $d > 1$ defined as*

$$\Psi(v) := J^\top v \quad for \quad J := \begin{bmatrix} I_{d-1} \\ 0 \end{bmatrix}.$$

For $p \in \mathbb{R}^d_{>0}$ such that $\|p\|_1 = 1$ with $\mu := \min_{i \in [d]} p_i > 0$, it holds that

$$\lambda_{\min}\left(\operatorname{diag}(\Psi(p)) - \Psi(p)\Psi(p)^\top\right) \geqslant \frac{\mu}{4(d-1)}.$$

Proof. For convenience, we define $q = \Psi(p)$ such that $q \in \mathbb{R}^{d-1}$. Let us begin by briefly establishing an upperbound for the minimum eigenvalue. First, noting that $\langle \mathbb{1}_{d-1}, q \rangle = 1 - p_d$ for all ones vector $\mathbb{1}_{d-1} \in \mathbb{R}^{d-1}$ and setting $v = \frac{1}{\sqrt{d-1}} \mathbb{1}_{d-1}$ such that $\|v\| = 1$, we can see

$$\lambda_{\min}(\operatorname{diag}(q) - qq^\top) \leqslant v^\top(\operatorname{diag}(q) - qq^\top)v$$
$$= \frac{1}{d-1}\left[\langle \mathbb{1}_{d-1}, q \rangle - \langle \mathbb{1}_{d-1}, q \rangle^2\right]$$
$$= \frac{p_d(1-p_d)}{d-1}.$$

Therefore, for $d > 1$ we can conclude that $\lambda_{\min}(\operatorname{diag}(q) - qq^\top) < p_d$. Now, for $j := \arg\min_{i \in [d-1]} q_i$ let us set v to be the basis vector in $\mathbb{R}^{d-1}$ such that $v_j = 1$ and $v_i = 0$ for all $i \in [d-1] \smallsetminus \{j\}$.

$$\lambda_{\min}(\operatorname{diag}(q) - qq^\top) \leqslant v^\top(\operatorname{diag}(q) - qq^\top)v$$
$$= q_j - q_j^2 < q_j.$$

Putting these together, we can see that for $d > 1$, $\lambda_{\min}(\operatorname{diag}(q) - qq^\top) < \min\{q_j, p_d\} = \mu$. Now let us move on to the lowerbound. Let λ be the smallest eigenvalue of $\operatorname{diag}(q) - qq^\top$: this means that

$$0 = \det(\lambda I - (\operatorname{diag}(q) - qq^\top)) = \det(\lambda I - \operatorname{diag}(q) + qq^\top).$$

Since we have established that $\lambda \notin \{p_1, \ldots, p_{d-1}\}$, we can see that the matrix $\lambda I - \operatorname{diag}(q)$ must be invertible. Hence, by the matrix determinant lemma,

$$\det(\lambda I - \operatorname{diag}(q))(1 + q^\top(\lambda I - \operatorname{diag}(q))^{-1}q) = 0,$$

and since $\det(\lambda I - \operatorname{diag}(q)) \neq 0$, we can see that

$$1 + q^\top(\lambda I - \operatorname{diag}(q))^{-1}q = 0,$$
$$\sum_{i=1}^{d-1} \frac{p_i^2}{p_i - \lambda} = 1.$$

Let us define $f_a(\lambda) := a/(a - \lambda)$ for some $a > 0$ and $\lambda \neq a$, such that $f_a'(\lambda) = a/(a-\lambda)^2$. By the mean value theorem, for some $c \in [0, \lambda]$,

$$f_a(\lambda) = f_a(0) + \lambda f_a'(c) = 1 + \frac{a\lambda}{(a-c)^2} \leqslant 1 + \frac{a\lambda}{(a-\lambda)^2}.$$

Using this in the original equation and noting $\sum_{i=1}^{d-1} p_i \leqslant 1 - \mu$ we have

$$\sum_{i=1}^{d-1} \frac{p_i^2}{p_i - \lambda} = \sum_{i=1}^{d-1} p_i \cdot f_{p_i}(\lambda) \leqslant \sum_{i=1}^{d-1} p_i \left(1 + \frac{p_i \lambda}{(p_i - \lambda)^2}\right) \leqslant 1 - \mu + \sum_{i=1}^{d-1} \frac{p_i^2 \lambda}{(p_i - \lambda)^2}.$$

Next, it is quick to check that $x \mapsto x^2/(x - \lambda)^2$ is *decreasing* for $x > \lambda$, since $\frac{\mathrm{d}}{\mathrm{d}x} x^2/(x - \lambda)^2 = -2\lambda x/(x - \lambda)^3$. Since we have established that the minimum eigenvalue is upperbounded by μ (for $d > 1$), we have that $p_i \geqslant \mu > \lambda$ for all $i \in [d]$ so we can upperbound the expression by lowerbounding the elements of p for

$$\sum_{i=1}^{d-1} \frac{p_i^2 \lambda}{(p_i - \lambda)^2} \leqslant (d - 1) \frac{\mu^2 \lambda}{(\mu - \lambda)^2}.$$

Now we write $\lambda(c) = c\mu$ for $c \in (0, 1)$, and compute a c_0 such that for all $c \leqslant c_0$, $(d - 1) \frac{\mu^2 \lambda(c)}{(\mu - \lambda(c))^2} < \mu$:

$$\frac{c}{(1 - c)^2}\mu = \frac{\mu^2 c\mu}{(1 - c)^2 \mu^2} < \frac{\mu}{d - 1} \iff \frac{c}{(1 - c)^2} < \frac{1}{d - 1}.$$

For the RHS, it suffices to take $c < 1/(4(d - 1))$. Thus, we conclude that:

$$\lambda_{\min}(\mathsf{diag}(q) - qq^{\mathsf{T}}) \geqslant \frac{\mu}{4(d - 1)}.$$

$\square$