On counterfactual inference with unobserved confounding

Abhin Shah
MIT
abhin@mit.edu

Raaz Dwivedi
MIT & Harvard University
raaz@mit.edu

Devavrat Shah
MIT
devavrat@mit.edu

Gregory W. Wornell
MIT
gww@mit.edu

Abstract

Given an observational study with n independent but heterogeneous units and one p-dimensional sample per unit containing covariates, interventions, and outcomes, our goal is to learn the counterfactual distribution for each unit. We consider studies with unobserved confounding which introduces statistical biases between interventions and outcomes as well as exacerbates the heterogeneity across units. Modeling the underlying joint distribution as an exponential family and under suitable conditions, we reduce learning the n unit-level counterfactual distributions to learning n exponential family distributions with heterogeneous parameters and only one sample per distribution. We introduce a convex objective that pools all n samples to jointly learn all n parameters and provide a unit-wise mean squared error bound that scales linearly with the metric entropy of the parameter space. For example, when the parameters are s-sparse linear combination of k known vectors, the error is $O(s \log k/p)$. En route, we derive sufficient conditions for compactly supported distributions to satisfy the logarithmic Sobolev inequality.

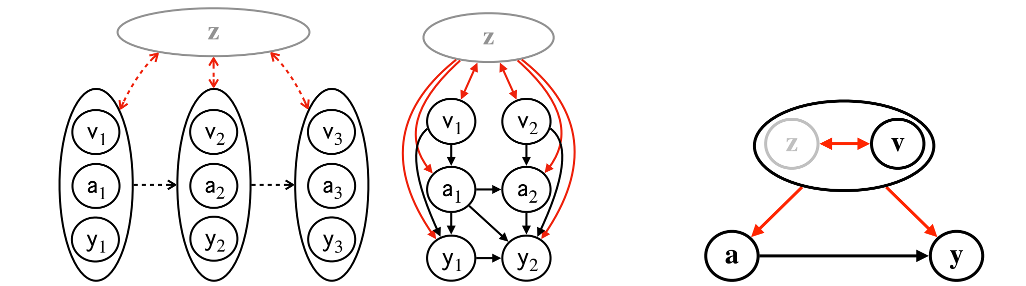
1 Introduction

We are interested in the problem of unit-level counterfactual inference owing to the increasing importance of personalized decision-making in many domains. As a motivating example, consider a recommender system interacting with a user over time. At each time, the user is exposed to a product based on observed demographic factors as well as certain unobserved factors, and the user's engagement level is recorded. The engagement level at any time can depend sequentially on the prior interaction in addition to the ongoing interaction (see Fig. 1(a)). The system can also sequentially adapt its recommendation. Given historical data of many heterogeneous users, the system wants to infer each user's average engagement level if it were exposed to a different sequence of products while the observed and the unobserved factors remain unchanged. This task is challenging since: (a) the *unobserved* factors could give rise to spurious associations, (b) the users could be *heterogeneous* in that they may have different responses to same sequence of products, and (c) each user only provides a *single* interaction trajectory.

In a general problem, we consider an observational setting where a unit undergoes interventions denoted by $\mathbf{a}$. We denote the outcomes of interest by $\mathbf{y}$, and allow the interventions $\mathbf{a}$ and the outcomes $\mathbf{y}$ to be confounded by observed covariates $\mathbf{v}$ as well as unobserved covariates $\mathbf{z}$. The graphical structure shown in Fig. 1(b) captures these interactions and is at the heart of our problem. We consider n heterogeneous and independent units indexed by $i \in [n] \triangleq \{1, \cdots, n\}$, and assume access to one observation per unit with $\boldsymbol{v}^{(i)}$, $\boldsymbol{a}^{(i)}$, and $\boldsymbol{y}^{(i)}$ denoting the realizations of $\mathbf{v}$, $\mathbf{a}$, and $\mathbf{y}$ for unit i respectively.

We operate within the Neyman-Rubin potential outcomes framework [21, 24] and denote the potential outcome of unit $i \in [n]$ under interventions $\boldsymbol{a}$ by $\boldsymbol{y}^{(i)}(\boldsymbol{a})$. Given the realizations $\big\{(\boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}, \boldsymbol{y}^{(i)})\big\}_{i=1}^{n}$, our goal is to answer counterfactual questions for these n units, e.g., what

NeurIPS 2022Workshop on Causal Machine Learning for Real-World Impact (CML4Impact 2022).



(a) A graphical model for sequential recommender system (b) A generic model for our setting

Figure 1: Two graphical models illustrating aspects of our work. The directed and the bi-directed arrows indicate causation and association respectively. The arrows involving the unobserved factors/covariates $\mathbf{z}$ are colored red. Panel **(a)** visualizes a sequential recommender system interacting with a user for 3 time points where v_t, a_t, and y_t denote the user's observed demographic factors, the product exposed to the user, and the user's engagement level respectively at time t, and $\mathbf{z}$ denotes unobserved factors. The left plot illustrates the dependency of the observed variables (v_t, a_t, y_t) at time t, on the observed variables at time $t-1$ via $\dashrightarrow$, and on the unobserved covariates $\mathbf{z}$ via $\leftarrow\dashrightarrow$; these dependencies for time 1 and 2 are expanded in the right plot. We do not assume any knowledge of such low-level causal links between elements of $\mathbf{z}$, $\mathbf{v}$, $\mathbf{a}$, and $\mathbf{y}$. Panel **(b)** exhibits a generic graphical model depicting the relationship between $\mathbf{z}$, $\mathbf{v}$, $\mathbf{a}$, and $\mathbf{y}$ for every unit. Our methodology works for any graphical model consistent with the graphical model in Panel **(b)**.

would the potential outcomes $\boldsymbol{y}^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})$ for interventions $\widetilde{\boldsymbol{a}}^{(i)} \neq \boldsymbol{a}^{(i)}$ be, while the observed and unobserved covariates remain unchanged? Under the stable unit treatment value assumption (SUTVA), i.e., the potential outcomes of unit i are not affected by the interventions at other units, learning unit-level counterfactual distributions is equivalent to learning unit-level conditional distributions

$$\left\{ f_{\mathbf{y}|\mathbf{a},\mathbf{z},\mathbf{v}}(\mathbf{y} = \cdot | \mathbf{a} = \cdot, \boldsymbol{z}^{(i)}, \boldsymbol{v}^{(i)}) \right\}_{i=1}^{n}. \tag{1}$$

Here, the i-th distribution represents the conditional distribution for the outcomes $\mathbf{y}$ as a function of the interventions $\mathbf{a}$, while keeping the observed covariates $\mathbf{v}$ and the unobserved covariates $\mathbf{z}$ fixed at the corresponding realizations for unit i, i.e., $\boldsymbol{v}^{(i)}$ and $\boldsymbol{z}^{(i)}$ respectively.

It is infeasible to answer such questions without any structural assumptions due to two key challenges: (a) unobserved confounding and (b) heterogeneity in unit-level conditional distributions. First, the unobserved covariates $\mathbf{z}$ introduces spurious statistical dependence between interventions and outcomes, known as *unobserved confounding*, which results in biased estimates. Second, the realizations $\left\{(\boldsymbol{z}^{(i)}, \boldsymbol{v}^{(i)})\right\}_{i=1}^{n}$ could be different for different units leading to *heterogeneity* in conditional distributions across units. The heterogeneity is crucial since we only observe one realization, namely the outcomes $\boldsymbol{y}^{(i)}(\boldsymbol{a}^{(i)})$ under the interventions $\boldsymbol{a}^{(i)}$, that is consistent with the unit-level conditional distribution $f_{\mathbf{y}|\mathbf{a},\mathbf{z},\mathbf{v}}(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{z}^{(i)}, \boldsymbol{v}^{(i)})$. As a result, one needs to learn n heterogeneous conditional distributions while having access to only one sample from each of them.

In this work, we model the joint distribution of the unobserved covariates, the observed covariates, the intervention, and the outcomes of interest as an exponential family distribution in accordance with the principle of maximum entropy.[1] With this modeling assumption, we show that both the aforementioned challenges can be tackled. In particular, we show that the n unit-level conditional distributions in (1) lead to n distributions from the same exponential family albeit with parameters that vary across units. The parameter corresponding to the i^{th} unit, say $\phi^{\star(i)}$, captures the effect of $\boldsymbol{z}^{(i)}$, and helps tackle the challenge of unobserved confounding. However, the challenge still remains to learn n heterogeneous exponential family distributions with one sample per distribution. This has been addressed in two specific scenarios in the literature: (a) if the unobserved confounding is identical across units, i.e., the parameters $\{\phi^{\star(i)}\}_{i=1}^{n}$ were all equal, then the challenge boils down to learning parameters of a single exponential family distribution from n samples which has been well-studied (see [26] for an overview); (b) if $\mathbf{v}$, $\mathbf{a}$, and $\mathbf{y}$ take binary values and have pairwise

[1]Exponential family distributions are the maximum entropy distributions given linear constraints on distributions such as bounded moments (see [8, 10]). The exponential family considered in this work arise when the first and second moments of the joint vector $(\mathbf{z}, \mathbf{v}, \mathbf{a}, \mathbf{y})$ are bounded.

interactions, and the dependencies between them are known, then the challenge boils down to learning certain parameters of an Ising model with one sample which has been studied in [12, 18]. In this work, we consider a generalized setting, where $\mathbf{v}$, $\mathbf{a}$, and $\mathbf{y}$ can be either discrete, continuous, or both, and assume no knowledge of the underlying dependencies.

Our contributions. As the main contribution, this work introduces a method to learn unit-level counterfactual distributions from observational studies in the presence of unobserved confounding with one sample per unit. Specifically, for every unit $i \in [n]$, we reduce learning its counterfactual distribution to learning $\phi^{\star(i)}$ from one sample $(\boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}, \boldsymbol{y}^{(i)})$ where $\phi^{\star(i)}$ is the parameter of the exponential family distribution corresponding to unit i. Our technical contributions are as follows.

1. We introduce a convex (and strictly proper) loss function (Def. 1) that pools the data $\{(\boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}, \boldsymbol{y}^{(i)})\}_{i=1}^{n}$ across all n samples to jointly learn all n parameters $\{\phi^{\star(i)}\}_{i=1}^{n}$.
2. For every unit i, we establish that the mean squared errors of our estimates of (a) $\phi^{\star(i)}$ (Thm. 1), and (b) the expected potential outcomes under alternate interventions (Thm. 2) scale linearly with the metric entropy of the parameter space, i.e., the set $\phi^{\star(i)}$ comes from. In particular, when $\phi^{\star(i)}$ is s-sparse linear combination of k known vectors (see Sec. 4), the error—just with one sample—decays as $O(s \log k/p)$, where p is the dimension of the tuple $(\mathbf{v}, \mathbf{a}, \mathbf{y})$.
3. As part of our analysis, we (a) derive sufficient conditions for a continuous random vector supported on a compact set to satisfy the logarithmic Sobolev inequality (Prop. 2) and (b) provide new concentration bounds for arbitrary functions of a continuous random vector that satisfies the logarithmic Sobolev inequality (Prop. 3). These results could be of independent interest.

Notation. For a deterministic sequence $u_1, \cdots, u_n$, we let $\boldsymbol{u} := (u_1, \cdots, u_n)$. For a random sequence $\mathsf{u}_1, \cdots, \mathsf{u}_n$, we let $\mathbf{u} := (\mathsf{u}_1, \cdots, \mathsf{u}_n)$. For a vector $\boldsymbol{u} \in \mathbb{R}^p$, we use u_t to denote its t^{th} coordinate and $u_{-t} \in \mathbb{R}^{p-1}$ to denote the vector after deleting the t^{th} coordinate. For a matrix $\mathbf{M} \in \mathbb{R}^{p \times p}$, we denote the element in t^{th} row and u^{th} column by $\mathbf{M}_{tu}$, the t^{th} row by $\mathbf{M}_t$, and the matrix maximum norm by $\|\mathbf{M}\|_{\max}$.

Related work. We provide an overview of related work that focus either on exponential family learning or on unit-level counterfactual inference with unobserved confounding. We refer the reader to [17, 15] for counterfactual inference with no unobserved confounding (as well as closely related concepts of ignorability in statistics and selection on the observables in economics). Likewise, we refer the reader to [22, 23] for counterfactual inference when the underlying causal mechanism (i.e., the directed acyclic graph) is known unlike our work.

Exponential family learning. There has been extensive work on learning parameters of a single exponential family distribution from multiple samples (see [26] for an overview). Of particular interest are the works that focus on learning sparse Markov random fields with (a) discrete variables [28] and (b) continuous variables [27] which inspire our loss function. Recently, a few works [18, 12] have focused on learning Ising model, i.e., sparse Markov random fields with binary variables, with one sample. However, these works focus on special cases where either the dependencies between the variables or a specific subset of the parameters are already known.

Unit-level counterfactual inference. For unit-level inference with unobserved confounding, prior work has largely focused on latent factor models, where the interventions and potential outcomes are assumed to be independent conditional on latent factors. These include popular frameworks of difference-in-differences [9, 5], synthetic controls [2, 1], synthetic difference-in-differences [6], and synthetic interventions [3]. A recent work [13] provides a latent factor model based approach for counterfactual inference in sequential experiments where the treatment mechanism is designed and known, and there is no confounding by definition. Notably these works allow only for finitely many interventions, and need multiple units to be simultaneously treated with the same interventions for a period of time (for their estimation strategies to work). Another key difference is that these works directly learn the outcomes, and not the distributions like we do.

2 Problem formulation

We consider a counterfactual inference task where units go through $p_a \geq 1$ interventions. For every unit, we observe $p_y \geq 1$ outcomes of interest. The interventions and the outcomes could be confounded by $p_v \geq 1$ observed covariates as well as $p_z \geq 1$ unobserved covariates. Additionally, the observed covariates and the unobserved covariates could be arbitrarily associated. We denote the random vector associated with the interventions, the outcomes, the observed covariates, and the unobserved covariates by $\mathbf{a} \triangleq (a_1, \cdots, a_{p_a}) \in \mathcal{A}^{p_a}$, $\mathbf{y} = (y_1, \cdots, y_{p_y}) \in \mathcal{Y}^{p_y}$, $\mathbf{v} \triangleq (v_1, \cdots, v_{p_v}) \in \mathcal{V}^{p_v}$, and $\mathbf{z} \triangleq (z_1, \cdots, z_{p_z}) \in \mathcal{Z}^{p_z}$ respectively where $\mathcal{A}, \mathcal{Y}, \mathcal{V}$, and $\mathcal{Z}$ denote the support of interventions, outcomes, observed covariates, and unobserved covariates respectively. We allow these sets to contain discrete, continuous, or mixed values. We summarize the causal relationship between the random vectors $\mathbf{z}$, $\mathbf{v}$, $\mathbf{a}$, and $\mathbf{y}$ in Fig. 1(b) where we denote the arbitrary association between $\mathbf{z}$ and $\mathbf{v}$ by a bi-directed arrow, and the causal association between (i) $(\mathbf{z}, \mathbf{v})$ and $\mathbf{a}$, (ii) $(\mathbf{z}, \mathbf{v})$ and $\mathbf{y}$, and (iii) $\mathbf{a}$ and $\mathbf{y}$ by directed arrows. Fig. 1(a) provides an example for sequential recommender systems covered by our work where (i) a_{t+1} depends on a_t in addition to v_{t+1} and $\mathbf{z}$, and (ii) y_{t+1} depends on a_t and y_t in addition to a_{t+1}, v_{t+1} and $\mathbf{z}$. We do not assume any knowledge of such low-level causal links between elements of $\mathbf{z}$, $\mathbf{v}$, $\mathbf{a}$, and $\mathbf{y}$. Our methodology works for any graphical model consistent with the graphical model in Fig. 1(b).

Data We are interested in answering counterfactual questions regarding n independent but heterogeneous units in a population. To do so, we assume access to one observation of the observed covariates, the interventions, and the outcomes per unit, and index it by $i \in [n]$, i.e., $\boldsymbol{v}^{(i)}$, $\boldsymbol{a}^{(i)}$, and $\boldsymbol{y}^{(i)}$ denote the realizations of $\mathbf{v}$, $\mathbf{a}$, and $\mathbf{y}$ for unit i respectively. For every realized tuple $(\boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}, \boldsymbol{y}^{(i)})$, there is a corresponding realization $\boldsymbol{z}^{(i)}$ of the unobserved covariates $\mathbf{z}$ that is unobserved.

Potential outcomes framework We adopt the potential outcomes framework and denote the potential outcomes of unit $i \in [n]$ under interventions $\boldsymbol{a} \in \mathcal{A}^{p_a}$ by $\boldsymbol{y}^{(i)}(\boldsymbol{a})$. We work under the stable unit treatment value assumption (SUTVA) where the potential outcomes of any unit i are unaffected by the interventions at other units. In fact, we assume independence across units implying the potential outcomes of any unit i are also unaffected by the covariates and the potential outcomes at other units. Then, the observations are related to the potential outcomes as $\boldsymbol{y}^{(i)} = \boldsymbol{y}^{(i)}(\boldsymbol{a}^{(i)})$ for all $i \in [n]$. To establish equivalence between unit-level counterfactual distribution and unit-level conditional distribution, consider unit $i \in [n]$ and fix the observed covariates and the unobserved covariates at $\boldsymbol{v}^{(i)}$ and $\boldsymbol{z}^{(i)}$ respectively. Then, let $\widetilde{\boldsymbol{y}}^{(i)}$ be a realization of $\mathbf{y}$ when $\mathbf{a} = \widetilde{\boldsymbol{a}}^{(i)}$. We are interested in the distribution of the potential outcomes of unit i for interventions $\widetilde{\boldsymbol{a}}^{(i)}$, i.e., the distribution of $\boldsymbol{y}^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})|\mathbf{z} = \boldsymbol{z}^{(i)}, \mathbf{v} = \boldsymbol{v}^{(i)}$. Under the causal framework considered here (see Fig. 1(b)), it is equivalent to the distribution of $\boldsymbol{y}^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})|\mathbf{a} = \widetilde{\boldsymbol{a}}^{(i)}, \mathbf{z} = \boldsymbol{z}^{(i)}, \mathbf{v} = \boldsymbol{v}^{(i)}$ since $(\mathbf{z}, \mathbf{v})$ satisfy ignorability [22, 17], i.e., the potential outcomes are independent of the interventions given $(\mathbf{z}, \mathbf{v})$. Further, under SUTVA, it is equivalent to the distribution of $\widetilde{\boldsymbol{y}}^{(i)}|\mathbf{a} = \widetilde{\boldsymbol{a}}^{(i)}, \mathbf{z} = \boldsymbol{z}^{(i)}, \mathbf{v} = \boldsymbol{v}^{(i)}$, i.e., $f_{\mathbf{y}|\mathbf{a},\mathbf{z},\mathbf{v}}(\mathbf{y} = \cdot|\mathbf{a} = \widetilde{\boldsymbol{a}}^{(i)}, \boldsymbol{z}^{(i)}, \boldsymbol{v}^{(i)})$. Therefore, our goal is to learn the n unit-level conditional distributions in (1). To that end, we model the joint distribution of the $\mathbf{z}$, $\mathbf{v}$, $\mathbf{a}$, and $\mathbf{y}$ belong to an exponential family.

Modeling data as exponential family Let $\widetilde{p} \triangleq p_z + p_v + p_a + p_y$. We parameterize the joint probability distribution $f_{\mathbf{w}}$ of the $\widetilde{p}$-dimensional random vector $\mathbf{w} \triangleq (\mathbf{z}, \mathbf{v}, \mathbf{a}, \mathbf{y})$ by natural statistics $\mathbf{w}$ and $\mathbf{w}\mathbf{w}^\top$, and by natural parameters $\phi \in \mathbb{R}^{\widetilde{p} \times 1}$ and $\Phi \in \mathbb{R}^{\widetilde{p} \times \widetilde{p}}$ as follows

$$f_{\mathbf{w}}(\boldsymbol{w}; \phi, \Phi) \propto \exp\Big(\phi^\top \boldsymbol{w} + \boldsymbol{w}^\top \Phi \boldsymbol{w}\Big), \quad \text{where} \quad \boldsymbol{w} \triangleq (\boldsymbol{z}, \boldsymbol{v}, \boldsymbol{a}, \boldsymbol{y}),$$

and $\boldsymbol{z} \triangleq (z_1, \cdots, z_{p_z})$, $\boldsymbol{v} \triangleq (v_1, \cdots, v_{p_v})$, $\boldsymbol{a} \triangleq (a_1, \cdots, a_{p_a})$, and $\boldsymbol{y} \triangleq (y_1, \cdots, y_{p_y})$ denote realizations of $\mathbf{z}$, $\mathbf{v}$, $\mathbf{a}$, and $\mathbf{y}$ respectively. Without loss of generality, we can assume Φ to be a symmetric matrix. Then, recognizing that the (unit-level) conditional distribution of $\mathbf{y}$ conditioned on $\mathbf{a} = \boldsymbol{a}$, $\mathbf{z} = \boldsymbol{z}$, and $\mathbf{v} = \boldsymbol{v}$ belongs to an exponential family with natural statistics $\mathbf{y}$ and $\mathbf{y}\mathbf{y}^\top$, we can write

$$f_{\mathbf{y}|\mathbf{a},\mathbf{z},\mathbf{v}}(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{z}, \boldsymbol{v}) \propto \exp\Big(\big[\phi^{(y)^\top} + 2\boldsymbol{z}^\top \Phi^{(z,y)} + 2\boldsymbol{v}^\top \Phi^{(v,y)} + 2\boldsymbol{a}^\top \Phi^{(a,y)}\big]\boldsymbol{y} + \boldsymbol{y}^\top \Phi^{(y,y)} \boldsymbol{y}\Big), \quad (2)$$

where $\phi^{(y)} \in \mathbb{R}^{p \times 1}$ is the component of ϕ corresponding to $\mathbf{y}$ and $\Phi^{(u,y)} \in \mathbb{R}^{p_u \times p_y}$ is the component of Φ corresponding to $\mathbf{u}$ and $\mathbf{y}$ for all $\mathbf{u} \in \{\mathbf{z}, \mathbf{v}, \mathbf{a}, \mathbf{y}\}$. We make two key observations: (a) The term $\Phi^{(z,y)\top}\boldsymbol{z}$ captures the effect of unobserved covariates $\boldsymbol{z}$ on $f_{\mathbf{y}|\mathbf{a},\mathbf{z},\mathbf{v}}(\mathbf{y} = \cdot|\mathbf{a} = \cdot, \boldsymbol{z}, \boldsymbol{v})$ in (2). (b) The task of learning $f_{\mathbf{y}|\mathbf{a},\mathbf{z},\mathbf{v}}(\mathbf{y} = \cdot|\mathbf{a} = \cdot, \boldsymbol{z}, \boldsymbol{v})$ in (2) as a function of $\mathbf{a}$ reduces to learning

$$\text{(i) } \phi^{(y)} + 2\Phi^{(z,y)\top}\boldsymbol{z} + 2\Phi^{(v,y)\top}\boldsymbol{v}, \qquad \text{(ii) } \Phi^{(a,y)}, \quad \text{and} \quad \text{(iii) } \Phi^{(y,y)}. \qquad (3)$$

Now, we argue that learning (i), (ii), and (iii) in (3) is subsumed in learning the parameters of the (unit-level) conditional distribution $f_{\mathsf{x}|\mathsf{z}}$ of the random vector $\mathsf{x} \triangleq (\mathsf{v}, \mathsf{a}, \mathsf{y})$ conditioned on $\mathsf{z} = \boldsymbol{z}$. To that end, we recognize that $f_{\mathsf{x}|\mathsf{z}}$ belongs to an exponential family with natural statistics x and $\mathsf{x}\mathsf{x}^\top$. For all $\mathsf{u} \in \{\mathsf{v}, \mathsf{a}, \mathsf{y}\}$, let $\phi^{(u)} \in \mathbb{R}^{p_u \times 1}$ be the component of ϕ corresponding to u, and $\Phi^{(z,u)} \in \mathbb{R}^{p_z \times p_u}$ be the component of Φ corresponding to z and u. We parameterize $f_{\mathsf{x}|\mathsf{z}}$ as

$$f_{\mathsf{x}|\mathsf{z}}\big(\boldsymbol{x}|\boldsymbol{z}; \theta(\boldsymbol{z}), \Theta\big) \propto \exp\Big([\theta(\boldsymbol{z})]^\top \boldsymbol{x} + \boldsymbol{x}^\top \Theta \boldsymbol{x}\Big) \text{ where } \theta(\boldsymbol{z}) \triangleq \begin{bmatrix} \phi^{(v)} + 2\Phi^{(z,v)\top}\boldsymbol{z} \\ \phi^{(a)} + 2\Phi^{(z,a)\top}\boldsymbol{z} \\ \phi^{(y)} + 2\Phi^{(z,y)\top}\boldsymbol{z} \end{bmatrix} \in \mathbb{R}^{p\times 1},\ \boldsymbol{x} \triangleq \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{a} \\ \boldsymbol{y} \end{bmatrix}, \tag{4}$$

$p \triangleq p_v + p_a + p_y$, $\Theta \in \mathbb{R}^{p\times p}$ denotes the component of Φ corresponding to x, and $\boldsymbol{v}$, $\boldsymbol{a}$, and $\boldsymbol{y}$ denote realizations of v, a, and y respectively. Now, if we learn $\theta(\boldsymbol{z})$ and Θ, we can obtain (i), (ii), and (iii) in (3) for any $\mathsf{v} = \boldsymbol{v}$ by using the appropriate components of $\theta(\boldsymbol{z})$ and Θ.

Inference tasks Let $f_{\mathsf{w}}(\cdot; \phi^*, \Phi^*)$ denote the true data generating distribution (Sec. 2) of w, and let $f_{\mathsf{x}|\mathsf{z}}\big(\cdot\,|\boldsymbol{z}; \theta^\star(\boldsymbol{z}), \Theta^\star\big)$ denote the true distribution of x conditioned on $\mathsf{z} = \boldsymbol{z}$. Then, for all $i \in [n]$, we note that the realization $\boldsymbol{x}^{(i)} \triangleq (\boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}, \boldsymbol{y}^{(i)})$ is consistent with the conditional distribution $f_{\mathsf{x}|\mathsf{z}}\big(\cdot\,|\boldsymbol{z}^{(i)}; \theta^\star(\boldsymbol{z}^{(i)}), \Theta^\star\big)$ where we do not observe $\boldsymbol{z}^{(i)}$. Our primary goal is to learn the n unit-level counterfactual distributions, which, as per our earlier discussion simplifies to estimating

$$\text{(i) Unit-level parameters } \theta^{\star(i)} \triangleq \theta^\star(\boldsymbol{z}^{(i)}) \text{ for } i \in [n], \text{ and (ii) Population-level parameter } \Theta^\star. \tag{5}$$

Our secondary goal is to estimate the expected potential outcomes for any given unit i (with $\mathsf{z} = \boldsymbol{z}^{(i)}, \mathsf{v} = \boldsymbol{v}^{(i)}$) and an alternate intervention $\widetilde{\boldsymbol{a}}^{(i)}$. In particular, our estimand of interest is

$$\mu^{(i)}(\widetilde{\boldsymbol{a}}^{(i)}) \triangleq \mathbb{E}[\boldsymbol{y}^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})|\mathsf{z} = \boldsymbol{z}^{(i)}, \mathsf{v} = \boldsymbol{v}^{(i)}]. \tag{6}$$

Assumptions We now state the assumptions we require to estimate the parameters in (5) and the expected potential outcomes in (6). For the ease of exposition, we let the sets $\mathcal{V}$, $\mathcal{A}$, and $\mathcal{Y}$ be continuous valued. However, our results apply equally to discrete and mixed cases. Without loss of generality, we let these sets be equal, and let $\mathcal{X} \triangleq \mathcal{V} = \mathcal{A} = \mathcal{Y}$ be bounded and symmetric around 0, i.e., $\mathcal{X} = \{-x_{\max}, x_{\max}\}$ where $x_{\max} < \infty$.

Assumption 1 (Bounded and sparse parameters)**.** *The true model parameters satisfy the following:*

(a) $\theta^{\star(i)}$ *and* $\Theta^\star$ *are bounded for all* $i \in [n]$,[2] *i.e.,* $\max\big\{\max_{i\in[n]} \big\|\theta^{\star(i)}\big\|_\infty, \|\Theta^\star\|_{\max}\big\} \leq \alpha$.

(b) $\Theta^\star$ *is row-wise sparse and has zeros on the diagonal, i.e.,* $\|\Theta_t^\star\|_0 \leq \beta$ *and* $\Theta_{tt}^\star = 0$ *for all* $t \in [p]$. *These imply that each* $\mathsf{x}_t \in \mathsf{x}$ *interacts with only a few other* $\mathsf{x}_u \in \mathsf{x}$ *in* (4).

Below, we define the set Λ_θ such that it contains all $p \times 1$ vectors θ satisfying Assum. 1(a) and the set Λ_Θ such that it contains all $p \times p$ symmetric matrices Θ satisfying Assum. 1(a) and (b).

$$\Lambda_\theta \triangleq \big\{\theta \in \mathbb{R}^{p\times 1} : \|\theta\|_\infty \leq \alpha\big\} \text{ and}$$
$$\Lambda_\Theta \triangleq \Big\{\Theta \in \mathbb{R}^{p\times p} : \Theta = \Theta^\top, \Theta_{tt} = 0 \text{ for all } t \in [p], \|\Theta\|_{\max} \leq \alpha, \max_{t\in[p]} \|\Theta_t\|_0 \leq \beta\Big\}. \tag{7}$$

3 Algorithm

We propose a computationally tractable loss function to estimate the unit-level and the population-level parameters in (5). Our algorithm jointly learns all the parameters of interest by pooling the observations across all n units and exploiting the exponential family structure of v, a, and y conditioned on $\mathsf{z} = \boldsymbol{z}$ in (4). In particular, our loss explicitly utilizes the fact that the population-level parameter $\Theta^\star$ is shared across units. We defer our estimate of the expected potential outcomes in (6) to App. A.1.

[2]This bound is necessary for model identifiability [25].

Loss function and parameter estimate. Consider any $t \in [p]$. Our loss function is inspired by the conditional distribution $f_{\mathsf{x}_t|\mathsf{x}_{-t},\mathsf{z}}$ of the random variable x_t conditioned on $\mathsf{x}_{-t} = \boldsymbol{x}_{-t}$ and $\mathsf{z} = \boldsymbol{z}$ which is given by

$$f_{\mathsf{x}_t|\mathsf{x}_{-t},\mathsf{z}}\big(x_t|\boldsymbol{x}_{-t}, \boldsymbol{z}; \theta_t(\boldsymbol{z}), \Theta_t\big) \propto \exp\left([\theta_t(\boldsymbol{z}) + 2\Theta_t^\top \boldsymbol{x}]x_t\right), \tag{8}$$

where $\theta_t(\boldsymbol{z})$ is the t^{th} element of $\theta(\boldsymbol{z})$, Θ_t is the t^{th} row of Θ, and $\boldsymbol{x}$ denotes a realization of x.

Definition 1 (**Loss function**)**.** *Given a sample $\boldsymbol{x}^{(i)}$ for every unit $i \in [n]$, our loss function maps $\underline{\Theta} \in \mathbb{R}^{p\times(n+p)}$ to $\mathcal{L}(\underline{\Theta}) \in \mathbb{R}$ defined as*

$$\mathcal{L}(\underline{\Theta}) = \frac{1}{n}\sum_{t\in[p]}\sum_{i\in[n]}\exp\left(-[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}]x_t^{(i)}\right) \text{where } \underline{\Theta} \triangleq \begin{bmatrix}\underline{\Theta}_1^\top \\ \vdots \\ \underline{\Theta}_p^\top\end{bmatrix}, \text{with } \underline{\Theta}_t \triangleq \{\theta_t^{(1)}, \cdots, \theta_t^{(n)}, \Theta_t\}. \tag{9}$$

The loss function defined above has many useful properties. We state one such property below with a proof in App. B.

Proposition 1 (**Proper loss function**)**.** *The loss function $\mathcal{L}(\cdot)$ is strictly proper, i.e., $\underline{\Theta}^\star = \arg\min_{\underline{\Theta}\in\Lambda_\theta^n\times\Lambda_\Theta} \mathbb{E}\big[\mathcal{L}(\underline{\Theta})\big]$.*

Our estimate of $\underline{\Theta}^\star$ (defined analogous to $\underline{\Theta}$) is obtained by minimizing the convex function $\mathcal{L}(\underline{\Theta})$ over all $\theta^{(i)} \in \Lambda_\theta$ for all $i \in [n]$ and $\Theta \in \Lambda_\Theta$, i.e.,

$$\widehat{\underline{\Theta}} \in \underset{\underline{\Theta}\in\Lambda_\theta^n\times\Lambda_\Theta}{\arg\min}\ \mathcal{L}(\underline{\Theta}). \tag{10}$$

We note (10) is a convex optimization problem, and a projected gradient descent algorithm returns an ϵ-optimal estimate with $\tau = O(p/\epsilon)$ iterations[3] where $\widehat{\underline{\Theta}}_\epsilon$ is said to be an ϵ-optimal estimate if $\mathcal{L}(\widehat{\underline{\Theta}}_\epsilon) \le \mathcal{L}(\widehat{\underline{\Theta}}) + \epsilon$ for any $\epsilon > 0$.

4 Main results

In this section, we provide our guarantee on estimating the unit-level and the population-level parameters in (5). We defer our guarantee on estimating the expected potential outcomes in (6) to App. A.2. Our guarantees depend on the metric entropy of the set Λ_θ which provides some notion of the richness of Λ_θ. To that end, we define the notions of ε-covering number and metric entropy.

Definition 2 (ε-covering number and metric entropy)**.** *Given a set $\mathcal{V} \subset \mathbb{R}^p$ and a scalar $\varepsilon > 0$, we use $\mathcal{C}(\mathcal{V}, \varepsilon)$ to denote the ε-covering number of $\mathcal{V}$ with respect to $\|\cdot\|_2$, i.e., $\mathcal{C}(\mathcal{V}, \varepsilon)$ denotes the minimum cardinality over all possible covers $\mathcal{U} \subset \mathcal{V}$ that satisfy*

$$\mathcal{V} \subset \cup_{u\in\mathcal{U}}\mathcal{B}(u; \varepsilon),$$

where $\mathcal{B}(u; \varepsilon) \triangleq \{v \in \mathbb{R}^p : \|u - v\|_2 \le \varepsilon\}$ denotes a ball of radius ε in $\mathbb{R}^p$ centered at u. Further, we use $\mathcal{M}(\mathcal{V}, \varepsilon)$ to denote the metric entropy of $\mathcal{V}$, i.e., $\mathcal{M}(\mathcal{V}, \varepsilon) \triangleq \log \mathcal{C}(\mathcal{V}, \varepsilon)$. Lastly, we use the shorthand notation $\mathcal{M}_\theta(\varepsilon) = \mathcal{M}(\Lambda_\theta, \varepsilon)$ to denote the metric entropy of Λ_θ.

The following result provides the estimation error for the estimate $\widehat{\underline{\Theta}}$ obtained in (10). We divide the proof across App. C and App. D.[4]

Theorem 1 (**Guarantee on parameter estimate**)**.** *Fix an $\varepsilon > 0$ and $\delta \in (0, 1)$, and define*

$$R(\varepsilon,\delta) \triangleq \max\{ce^{c'\beta}\sqrt{\log\log(2p/\delta)) + \mathcal{M}_\theta(\eta)}, \varepsilon\gamma\},\ \ \eta \triangleq \frac{ce^{-c'\beta}}{\gamma},\ \text{and } \gamma \triangleq \max_{\theta,\bar{\theta}\in\Lambda_\theta} \frac{\|\theta - \bar{\theta}\|_1}{\|\theta - \bar{\theta}\|_2}. \tag{11}$$

[3]This follows from [7, Theorem 10.6] by noting that $\mathcal{L}(\underline{\Theta})$ is $O(p)$ smooth function of $\underline{\Theta}$.

[4]To simplify presentation, we use c and c' to denote universal constants or constants that depend on the model-parameters α and $x_{\max}$, and can take a different value in each appearance.

Then, with probability at least $1-\delta$*, the estimates* $\widehat{\Theta}, \widehat{\theta}^{(1)}, \cdots, \widehat{\theta}^{(n)}$ *defined in* (10) *satisfy*

$$\max_{t\in[p]} \|\widehat{\Theta}_t - \Theta_t^\star\|_2 \leq \varepsilon \qquad \textit{whenever} \quad n \geq \frac{ce^{c'\beta}\log\frac{p}{\delta}}{\varepsilon^4} \quad \textit{and}$$

$$\max_{i\in[n]} \|\widehat{\theta}^{(i)} - \theta^{\star(i)}\|_2 \leq R(\varepsilon, \delta/n) \quad \textit{whenever} \quad n \geq \frac{ce^{c'\beta}(\log\frac{np}{\delta} + \mathcal{M}_\theta(\eta))}{\varepsilon^4}.$$

Example. Consider the case where $\theta^{\star(i)}$ is s-sparse linear combination of k known vectors for all $i \in [n]$, i.e., $\theta^{\star(i)} = \mathbf{B}\mathbf{a}^{(i)}$ for a known matrix $\mathbf{B} \in \mathbb{R}^{p\times k}$ and $\mathbf{a}^{(i)} \in \mathbb{R}^{k\times 1}$ with $\|\mathbf{a}^{(i)}\|_0 \leq s$ such that $\|\theta^{\star(i)}\|_\infty \leq \alpha$. Then, the parameter set Λ_θ can be re-parameterized as the set $\Lambda_{\mathbf{a}} = \{\mathbf{a} \in \mathbb{R}^{k\times 1} : \|\mathbf{a}\|_0 \leq s \text{ and } \|\mathbf{B}\mathbf{a}\|_\infty \leq \alpha\}$ whose metric entropy is given by $\mathcal{M}_{\mathbf{a}}(\eta) = \left(1 + \frac{c}{\eta}\right)^s \binom{k}{s}$ for some constant c [12, Corollary 4]. Using this bound, and substituting the worst case $\gamma = \sqrt{p}$, the guarantees in Thm. 1 simplify. We capture this in the following corollary by focusing on the mean squared error.

Corollary 1. *Fix an* $\varepsilon > 0$ *and* $\delta \in (0,1)$*. Suppose* $\theta^{\star(i)}$ *is* s*-sparse linear combination of* k *known vectors for all* $i \in [n]$*. Then, with probability at least* $1-\delta$*, the estimates* $\{\widehat{\theta}^{(i)}\}_{i=1}^n$ *defined in* (10) *satisfy*

$$\max_{i\in[n]} \mathrm{MSE}(\widehat{\theta}^{(i)}, \theta^{\star(i)}) \leq \frac{\max\{\varepsilon^2, ce^{c'\beta}s\log pk\}}{p} \quad \textit{whenever} \quad n \geq \frac{ce^{c'\beta}sp^2\log\frac{npk}{\delta}}{\varepsilon^4}.$$

As a result, we see that the unit-wise mean squared error for parameter estimation scale as $O(s\log k)/p$ when (i) the true parameters are s-sparse linear combination of k known vectors and (ii) n scales as $O(sp^2\log spk)$.

References

[1] A. Abadie, A. Diamond, and J. Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, 2010. (Cited on page 3.)

[2] A. Abadie and J. Gardeazabal. The economic costs of conflict: A case study of the basque country. *American Economic Review*, 2003. (Cited on page 3.)

[3] A. Agarwal, D. Shah, and D. Shen. Synthetic interventions. *arXiv preprint arXiv:2006.07691*, 2021. (Cited on page 3.)

[4] S. Aida and D. Stroock. Moment estimates derived from poincaré and logarithmic sobolev inequalities. *Mathematical Research Letters*, 1(1):75–86, 1994. (Cited on page 41.)

[5] J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009. (Cited on page 3.)

[6] D. Arkhangelsky, S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager. Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118, December 2021. (Cited on page 3.)

[7] M. Bande and N. He. Lower bounds & projected gradient descent. 2016. (Cited on page 6.)

[8] O. Barndorff-Nielsen. *Information and exponential families: in statistical theory*. John Wiley & Sons, 2014. (Cited on page 2.)

[9] M. Bertrand, E. Duflo, and S. Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1):249–275, 2004. (Cited on page 3.)

[10] L. D. Brown. Fundamentals of statistical exponential families: with applications in statistical decision theory. Ims, 1986. (Cited on page 2.)

[11] R. Busa-Fekete, D. Fotakis, B. Szörényi, and M. Zampetakis. Optimal learning of mallows block model. In *Conference on Learning Theory*, pages 529–532. PMLR, 2019. (Cited on pages 12 and 45.)

[12] Y. Dagan, C. Daskalakis, N. Dikkala, and A. V. Kandiros. Learning ising models from one or multiple samples. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 161–168, 2021. (Cited on pages 3, 7, 9, 34, 44, and 45.)

[13] R. Dwivedi, S. Murphy, and D. Shah. Counterfactual inference for sequential experimental design. *arXiv preprint arXiv:2202.06891*, 2022. (Cited on page 3.)

[14] W. Ghang, Z. Martin, and S. Waruhiu. The sharp log-sobolev inequality on a compact interval. 2014. (Cited on page 46.)

[15] M. Hernán and J. Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020. (Cited on page 3.)

[16] R. Holley and D. W. Stroock. Logarithmic sobolev inequalities and stochastic ising models. 1986. (Cited on page 46.)

[17] G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015. (Cited on pages 3 and 4.)

[18] V. Kandiros, Y. Dagan, N. Dikkala, S. Goel, and C. Daskalakis. Statistical estimation from dependent data. In *International Conference on Machine Learning*, pages 5269–5278. PMLR, 2021. (Cited on pages 3 and 9.)

[19] M. Ledoux. Logarithmic sobolev inequalities for unbounded spin systems revisited. In *Séminaire de Probabilités XXXV*, pages 167–194. Springer, 2001. (Cited on page 46.)

[20] K. Marton. Logarithmic sobolev inequalities in discrete product spaces: a proof by a transportation cost distance. *arXiv preprint arXiv:1507.02803*, 2015. (Cited on pages 33, 34, and 38.)

[21] J. Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10(1):1–51, 1923. (Cited on page 1.)

[22] J. Pearl. *Causality*. Cambridge university press, 2009. (Cited on pages 3 and 4.)

[23] J. Pearl, M. Glymour, and N. P. Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. (Cited on page 3.)

[24] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974. (Cited on page 1.)

[25] N. P. Santhanam and M. J. Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory*, 58(7):4117–4134, 2012. (Cited on page 5.)

[26] A. Shah, D. Shah, and G. Wornell. A computationally efficient method for learning exponential family distributions. *Advances in Neural Information Processing Systems*, 34:15841–15854, 2021. (Cited on pages 2 and 3.)

[27] A. Shah, D. Shah, and G. Wornell. On learning continuous pairwise markov random fields. In *International Conference on Artificial Intelligence and Statistics*, pages 1153–1161. PMLR, 2021. (Cited on page 3.)

[28] M. Vuffray, S. Misra, and A. Y. Lokhov. Efficient learning of discrete graphical models. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124017, 2022. (Cited on page 3.)

Acknowledgments

This work was supported, in part, by NSF under Grant No. DMS-2023528 as part of the Foundations of Data Science Institute (FODSI), the MIT-IBM Watson AI Lab under Agreement No. W1771646, MIT-IBM projects on Time Series and Causal Inference as well as project with DSO National Laboratory.

The authors thank Thomas Courtade, Yuzhou Gu, Anuran Makur, Wenlong Mou, Felix Otto, and Yury Polyanskiy for helpful pointers regarding Logarithmic Sobolev inequalities. The authors also thank Yuval Dagan and Anthimos Vardis Kandiros for helpful discussion about [12] and [18].

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

A Causal estimate and associated guarantees

In this section, we provide our estimate of the expected potential outcomes under alternate intervention and the corresponding estimation error.

A.1 Causal estimate

Assuming the estimate $\widehat{\underline{\Theta}}$ of $\underline{\Theta}^\star$ is given from Sec. 3, we define our estimate of the expected potential outcome $\mu^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})$ (see (6)) for any given unit $i \in [n]$ under an alternate intervention $\widetilde{\boldsymbol{a}}^{(i)} \in \mathcal{A}^{p_a}$. First, we identify $\widehat{\Phi}^{(u,y)} \in \mathbb{R}^{p_u \times p_y}$ to be the component of $\widehat{\Theta}$ corresponding to $\mathbf{u}$ and $\mathbf{y}$ for all $\mathbf{u} \in \{\mathbf{v}, \mathbf{a}, \mathbf{y}\}$ and $\widehat{\theta}^{(i,y)} \in \mathbb{R}^{p_y}$ to be the component of $\widehat{\theta}^{(i)}$ corresponding to $\mathbf{y}$. Then, our estimate of the conditional distribution of $\mathbf{y}$ as a function of the interventions $\mathbf{a}$, while keeping $\mathbf{v}$ (observed) and $\mathbf{z}$ (unobserved) fixed at the corresponding realizations for unit i, i.e., $\boldsymbol{v}^{(i)}$ and $\boldsymbol{z}^{(i)}$ respectively, is as follows:

$$\widehat{f}^{(i)}_{\mathbf{y}|\mathbf{a}}(\boldsymbol{y}|\boldsymbol{a}) \propto \exp\Big(\big[\widehat{\theta}^{(i,y)} + 2\boldsymbol{v}^{(i)\top}\widehat{\Phi}^{(v,y)} + 2\boldsymbol{a}^\top\widehat{\Phi}^{(a,y)}\big]\boldsymbol{y} + \boldsymbol{y}^\top\widehat{\Phi}^{(y,y)}\boldsymbol{y}\Big). \tag{12}$$

Finally, our estimate of $\mu^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})$ is given by

$$\widehat{\mu}^{(i)}(\widetilde{\boldsymbol{a}}^{(i)}) \triangleq \mathbb{E}_{\widehat{f}^{(i)}_{\mathbf{y}|\mathbf{a}}}[\mathbf{y}|\mathbf{a} = \widetilde{\boldsymbol{a}}^{(i)}]. \tag{13}$$

A.2 Guarantee on outcome estimate

The following result provides the estimation error for the estimate $\widehat{\mu}^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})$ (see (13)) of the expected potential outcomes for any unit $i \in [n]$ under an alternate intervention $\widetilde{\boldsymbol{a}}^{(i)} \in \mathcal{A}^{p_a}$. The result requires the operator norms of the following matrices to remain bounded for minor perturbation in the parameters: (i) the covariance matrix of $\mathbf{y}$ conditioned on $\mathbf{a}$, $\mathbf{z}$, and $\mathbf{v}$ and (ii) the cross-covariance matrix of $\mathbf{y}$ and $y_t\mathbf{y}$ conditioned on $\mathbf{a}$, $\mathbf{z}$, and $\mathbf{v}$ for all $t \in [p_y]$. We note that the expectation in the aforementioned covariance matrix and cross-covariance matrices is with respect to the conditional distribution of $\mathbf{y}$ conditioned on $\mathbf{a} = \boldsymbol{a}$, $\mathbf{z} = \boldsymbol{z}$, and $\mathbf{v} = \boldsymbol{v}$ which is fully parameterized by θ and Θ, i.e., replace $\theta(\boldsymbol{z})$ by θ in (4), and we require the operator norms of these matrices to remain bounded for perturbations in θ and Θ. Formally, for perturbations in θ and Θ such that θ, Θ belong to the set $\mathbb{B}$, we have

$$\sup_{\theta,\Theta\in\mathbb{B}} \max\Big\{\|\mathbb{C}\mathrm{ov}_{\theta,\Theta}(\mathbf{y},\mathbf{y}|\boldsymbol{a},\boldsymbol{z},\boldsymbol{v})\|_{\mathrm{op}}, \max_{t\in[p_y]}\|\mathbb{C}\mathrm{ov}_{\theta,\Theta}(\mathbf{y}, y_t\mathbf{y}|\boldsymbol{a},\boldsymbol{z},\boldsymbol{v})\|_{\mathrm{op}}\Big\} \leq M(\mathbb{B}), \tag{14}$$

where $M(\mathbb{B})$ is a constant that depends on the set $\mathbb{B}$. For simplicity, we assume $p_v = p_a = p_y$. See the proof below for the general case.

Theorem 2 (**Guarantee on outcome estimate**). *Fix an $\varepsilon > 0$ and $\delta \in (0,1)$. Then, with probability at least $1-\delta$, the estimates $\{\widehat{\mu}^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})\}_{i=1}^n$ defined in (13) for any $\{\widetilde{\boldsymbol{a}}^{(i)} \in \mathcal{A}^{p_a}\}_{i=1}^n$ satisfy*

$$\max_{i\in[n]} \frac{\|\mu^{(i)}(\widetilde{\boldsymbol{a}}^{(i)}) - \widehat{\mu}^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})\|_2}{M(\mathbb{B}_i)} \leq R(\varepsilon, \delta/n) + p\varepsilon \quad \text{whenever} \quad n \geq \frac{ce^{c'\beta}(\log\frac{np}{\delta} + \mathcal{M}_\theta(\eta))}{\varepsilon^4},$$

where $R(\varepsilon,\delta)$ and η were defined in (11), $M(\mathbb{B})$ was defined in (14), and

$$\mathbb{B}_i \triangleq \big\{\theta \in \Lambda_\theta : \|\theta - \theta^{\star(i)}\|_2 \leq R(\varepsilon, \delta/n)\big\} \times \big\{\Theta \in \Lambda_\Theta : \max_{t\in[p]}\|\Theta_t - \Theta_t^\star\|_2 \leq \varepsilon\big\}.$$

Proof. Fix any unit $i \in [n]$ and an alternate intervention $\widetilde{\boldsymbol{a}}^{(i)} \in \mathcal{A}^{p_a}$. Then, we have

$$\mu^{(i)}(\widetilde{\boldsymbol{a}}^{(i)}) \stackrel{(6)}{=} \mathbb{E}[\boldsymbol{y}^{(i)}(\widetilde{\boldsymbol{a}}^{(i)})|\mathbf{z} = \boldsymbol{z}^{(i)}, \mathbf{v} = \boldsymbol{v}^{(i)}] \stackrel{(a)}{=} \mathbb{E}[\mathbf{y}|\mathbf{a} = \widetilde{\boldsymbol{a}}^{(i)}, \mathbf{z} = \boldsymbol{z}^{(i)}, \mathbf{v} = \boldsymbol{v}^{(i)}],$$

where (a) follows because the unit-level counterfactual distribution is equivalent to unit-level conditional distribution under the causal framework considered as described in Sec. 2. To obtain a convenient expression for $\mathbb{E}[\mathbf{y}|\mathbf{a} = \widetilde{\boldsymbol{a}}^{(i)}, \mathbf{z} = \boldsymbol{z}^{(i)}, \mathbf{v} = \boldsymbol{v}^{(i)}]$, we identify $\Phi^{\star(u,y)} \in \mathbb{R}^{p_u \times p_y}$ to be the component of $\Theta^\star$ corresponding to $\mathbf{u}$ and $\mathbf{y}$ for all $\mathbf{u} \in \{\mathbf{v}, \mathbf{a}, \mathbf{y}\}$ and $\theta^{\star(i,y)} \in \mathbb{R}^{p_y}$ to be the component of $\theta^{\star(i)}$ corresponding to $\mathbf{y}$. Then, the conditional distribution of $\mathbf{y}$ as a function of the interventions $\mathbf{a}$, while keeping $\mathbf{v}$ and $\mathbf{z}$ fixed at the corresponding realizations for unit i, i.e., $\boldsymbol{v}^{(i)}$ and $\boldsymbol{z}^{(i)}$ respectively, can be written as

$$f^{(i)}_{\mathbf{y}|\mathbf{a}}(\boldsymbol{y}|\boldsymbol{a}) \propto \exp\Big(\big[\theta^{\star(i,y)} + 2\boldsymbol{v}^{(i)\top}\Phi^{\star(v,y)} + 2\boldsymbol{a}^\top\Phi^{\star(a,y)}\big]\boldsymbol{y} + \boldsymbol{y}^\top\Phi^{\star(y,y)}\boldsymbol{y}\Big). \tag{15}$$

Therefore, we have

$$\mathbb{E}[\mathsf{y}|\mathsf{a} = \widetilde{\boldsymbol{a}}^{(i)}, \mathsf{z} = \boldsymbol{z}^{(i)}, \mathsf{v} = \boldsymbol{v}^{(i)}] = \mathbb{E}_{f^{(i)}_{\mathsf{y}|\mathsf{a}}}[\mathsf{y}|\mathsf{a} = \widetilde{\boldsymbol{a}}^{(i)}].$$

Now, consider the p_u dimensional random vector u supported on $\mathcal{X}^{p_u}$ with distribution f_{u} parameterized by $\psi \in \mathbb{R}^{p_y}$ and $\Psi \in \mathbb{R}^{p_y \times p_y}$ as follows

$$f_{\mathsf{u}}(\boldsymbol{u}|\psi, \Psi) \propto \exp(\psi^\top \boldsymbol{u} + \boldsymbol{u}^\top \Psi \boldsymbol{u}). \tag{16}$$

Then, note that $\widehat{f}^{(i)}_{\mathsf{y}|\mathsf{a}}(\boldsymbol{y}|\boldsymbol{a})$ in (12) and $f^{(i)}_{\mathsf{y}|\mathsf{a}}(\boldsymbol{y}|\boldsymbol{a})$ in (15) belong to the set $\{f_{\mathsf{u}}(\cdot|\psi, \Psi) : \psi \in \mathbb{R}^{p_y}, \Psi \in \mathbb{R}^{p_y \times p_y}\}$ for some ψ and Ψ. Now, we consider any two distributions in this set, namely $f_{\mathsf{u}}(\boldsymbol{u}|\widehat{\psi}, \widehat{\Psi})$ and $f_{\mathsf{u}}(\boldsymbol{u}|\psi^\star, \Psi^\star)$. Then, we claim that the two norm of the difference of the mean vectors of these distributions is bounded as below. We provide a proof at the end.

Lemma 1 (Perturbation in the mean vector). *For any $\psi \in \mathbb{R}^{p_y}$ and $\Psi \in \mathbb{R}^{p_y \times p_y}$, let $\mu_{\psi,\Psi}(\mathsf{u}) \in \mathbb{R}^{p_u}$ and $\mathbb{C}ov_{\psi,\Psi}(\mathsf{u}, \mathsf{u}) \in \mathbb{R}^{p_u \times p_u}$ denote the mean vector and the covariance matrix of u respectively with respect to f_{u} in (16). Then, for any $\widehat{\psi}, \psi^\star \in \mathbb{R}^{p_y}$ and $\widehat{\Psi}, \Psi^\star \in \mathbb{R}^{p_y \times p_y}$, there exists some $t \in (0, 1)$, $\widetilde{\psi} \triangleq t\widehat{\psi} + (1-t)\psi^\star$ and $\widetilde{\Psi} \triangleq t\widehat{\psi} + (1-t)\widetilde{\psi}$ such that*

$$\|\mu_{\widehat{\psi},\widehat{\Psi}}(\mathsf{u}) - \mu_{\psi^\star,\Psi^\star}(\mathsf{u})\|_2 \leq \|\mathbb{C}ov_{\widetilde{\psi},\widetilde{\Psi}}(\mathsf{u}, \mathsf{u})\|_{\mathrm{op}} \|(\widehat{\psi} - \psi^\star)\|_2 + \sum_{t_3 \in [p]} \|\mathbb{C}ov_{\widetilde{\psi},\widetilde{\Psi}}(\mathsf{u}, u_{t_3}\mathsf{u})\|_{\mathrm{op}} \|(\widehat{\Psi}_{t_3} - \Psi^\star_{t_3})\|_2.$$

Given this lemma, we proceed with the proof. By applying this lemma to $\widehat{f}^{(i)}_{\mathsf{y}|\mathsf{a}}(\boldsymbol{y}|\boldsymbol{a})$ in (12) and $f^{(i)}_{\mathsf{y}|\mathsf{a}}(\boldsymbol{y}|\boldsymbol{a})$ in (15), we see that it is sufficient to show the following bound

$$\|(\theta^{\star(i,y)} - \widehat{\theta}^{(i,y)}) + 2\boldsymbol{v}^{(i)\top}(\Phi^{\star(v,y)} - \widehat{\Phi}^{(v,y)}) + 2\widetilde{\boldsymbol{a}}^{(i)\top}(\Phi^{\star(a,y)} - \widehat{\Phi}^{(a,y)})\|_2 + \sum_{t \in [p_y]} \|\Phi^{\star(y,y)}_t - \widehat{\Phi}^{(y,y)}_t\|_2 \leq R(\varepsilon, \delta/n) + p\varepsilon.$$

To that end, we have

$$\sum_{t \in [p_y]} \|\Phi^{\star(y,y)}_t - \widehat{\Phi}^{(y,y)}_t\|_2 \overset{(a)}{\leq} \sum_{t \in [p_y]} \|\Theta^\star_t - \widehat{\Theta}_t\|_2, \tag{17}$$

where (a) follows because ℓ_2 norm of any sub-vector is no more than ℓ_2 norm of the vector. Similarly, we have

$$\begin{aligned}
&\|(\theta^{\star(i,y)} - \widehat{\theta}^{(i,y)}) + 2\boldsymbol{v}^{(i)\top}(\Phi^{\star(v,y)} - \widehat{\Phi}^{(v,y)}) + 2\widetilde{\boldsymbol{a}}^{(i)\top}(\Phi^{\star(a,y)} - \widehat{\Phi}^{(a,y)})\|_2 \\
&\overset{(a)}{\leq} \|\theta^{\star(i,y)} - \widehat{\theta}^{(i,y)}\|_2 + 2\|\boldsymbol{v}^{(i)\top}(\Phi^{\star(v,y)} - \widehat{\Phi}^{(v,y)})\|_2 + 2\|\widetilde{\boldsymbol{a}}^{(i)\top}(\Phi^{\star(a,y)} - \widehat{\Phi}^{(a,y)})\|_2 \\
&\overset{(b)}{\leq} \|\theta^{\star(i,y)} - \widehat{\theta}^{(i,y)}\|_2 + 2\|\boldsymbol{v}^{(i)}\|_2 \|\Phi^{\star(v,y)} - \widehat{\Phi}^{(v,y)}\|_{\mathrm{op}} + 2\|\widetilde{\boldsymbol{a}}^{(i)}\|_2 \|(\Phi^{\star(a,y)} - \widehat{\Phi}^{(a,y)})\|_{\mathrm{op}} \\
&\overset{(c)}{\leq} \|\theta^{\star(i)} - \widehat{\theta}^{(i)}\|_2 + 2\Big(\|\boldsymbol{v}^{(i)}\|_2 + \|\widetilde{\boldsymbol{a}}^{(i)}\|_2\Big) \|\Theta^\star - \widehat{\Theta}\|_{\mathrm{op}} \\
&\overset{(d)}{\leq} \|\theta^{\star(i)} - \widehat{\theta}^{(i)}\|_2 + 2\Big(\|\boldsymbol{v}^{(i)}\|_2 + \|\widetilde{\boldsymbol{a}}^{(i)}\|_2\Big) \|\Theta^\star - \widehat{\Theta}\|_1 \\
&\overset{(e)}{\leq} \|\theta^{\star(i)} - \widehat{\theta}^{(i)}\|_2 + 2x_{\max}\big(\sqrt{p_v} + \sqrt{p_a}\big) \|\Theta^\star - \widehat{\Theta}\|_1,
\end{aligned} \tag{18}$$

where (a) follows from triangle inequality, (b) follows because induced matrix norms are submultiplicative, (c) follows because operator norm of any sub-matrix is no more than operator norm of the matrix and ℓ_2 norm of any sub-vector is no more than ℓ_2 norm of the vector, (d) follows because $\Theta^\star - \widehat{\Theta}$ is symmetric and because matrix operator norm is bounded by square root of the product of matrix one norm and matrix infinity norm, and (e) follows because $\max\{\|\boldsymbol{v}^{(i)}\|_\infty, \|\boldsymbol{a}^{(i)}\|_\infty\} \leq x_{\max}$ for all $i \in [n]$.

Now, combining (17) and (18), we have

$$\|(\theta^{\star(i,y)}-\widehat{\theta}^{(i,y)})+2\boldsymbol{v}^{(i)\top}(\Phi^{\star(v,y)}-\widehat{\Phi}^{(v,y)})+2\widetilde{\boldsymbol{a}}^{(i)\top}(\Phi^{\star(a,y)}-\widehat{\Phi}^{(a,y)})\|_2+\sum_{t\in[p_y]}\|\Phi_t^{\star(y,y)}-\widehat{\Phi}_t^{(y,y)}\|_2$$

$$\leq \|\theta^{\star(i)}-\widehat{\theta}^{(i)}\|_2 + 2x_{\max}\big(\sqrt{p_v}+\sqrt{p_a}\big)\|\Theta^\star-\widehat{\Theta}\|_1 + \sum_{t\in[p_y]}\|\Theta_t^\star-\widehat{\Theta}_t\|_2$$

$$\overset{(a)}{\leq} R(\varepsilon,\delta/n) + 2x_{\max}\big(\sqrt{p_v}+\sqrt{p_a}\big)\sqrt{p}\varepsilon + p_y\varepsilon,$$

and (a) follows from Thm. 1 by using the relationship between vector norms. The proof is complete by rescaling ε and absorbing the constants in c.

Proof of Lem. 1: Perturbation in the mean vector: Let $Z(\psi,\Psi)\in\mathbb{R}_+$ denote the log-partition function of $f_{\mathbf{u}}(\cdot|\psi,\Psi)$ in (16). Then, from [11, Theorem 1], we have

$$\|\mu_{\widehat{\psi},\widehat{\Psi}}(\mathbf{u})-\mu_{\psi^\star,\Psi^\star}(\mathbf{u})\|_2 = \|\nabla_{\widehat{\psi}}Z(\widehat{\psi},\widehat{\Psi})-\nabla_{\psi^\star}Z(\psi^\star,\Psi^\star)\|_2. \tag{19}$$

For $t_1,t_2,t_3\in[p]$, consider $\frac{\partial^2 Z(\psi,\Psi)}{\partial\psi_{t_1}\partial\psi_{t_2}}$ and $\frac{\partial^2 Z(\psi,\Psi)}{\partial\psi_{t_1}\partial\Psi_{t_2,t_3}}$. Using the fact that the Hessian of the log partition function of any regular exponential family is the covariance matrix of the associated sufficient statistic, we have

$$\frac{\partial^2 Z(\psi,\Psi)}{\partial\psi_{t_1}\partial\psi_{t_2}} = \mathbb{C}\mathrm{ov}_{\psi,\Psi}(u_{t_1},u_{t_2}) \quad\text{and}\quad \frac{\partial^2 Z(\psi,\Psi)}{\partial\psi_{t_1}\partial\Psi_{t_2,t_3}} = \mathbb{C}\mathrm{ov}_{\psi,\Psi}(u_{t_1},u_{t_2}u_{t_3}). \tag{20}$$

Now, for some $c\in(0,1)$, $\widetilde{\psi}\triangleq c\widehat{\psi}+(1-c)\psi^\star$ and $\widetilde{\Psi}\triangleq c\widehat{\psi}+(1-c)\widetilde{\psi}$, we have the following from the mean value theorem

$$\frac{\partial Z(\widehat{\psi},\widehat{\Psi})}{\partial\widehat{\psi}_{t_1}} - \frac{\partial Z(\psi^\star,\Psi^\star)}{\partial\psi^\star_{t_1}}$$

$$=\sum_{t_2\in[p]}\frac{\partial^2 Z(\widetilde{\psi},\widetilde{\Psi})}{\partial\widetilde{\psi}_{t_2}\partial\widetilde{\psi}_{t_1}}\cdot(\widehat{\psi}_{t_2}-\psi^\star_{t_2}) + \sum_{t_2\in[p]}\sum_{t_3\in[p]}\frac{\partial^2 Z(\widetilde{\psi},\widetilde{\Psi})}{\partial\widetilde{\Psi}_{t_2,t_3}\partial\widetilde{\psi}_{t_1}}\cdot(\widehat{\Psi}_{t_2,t_3}-\Psi^\star_{t_2,t_3})$$

$$\overset{(20)}{=}\sum_{t_2\in[p]}\mathbb{C}\mathrm{ov}_{\widetilde{\psi},\widetilde{\Psi}}(u_{t_1},u_{t_2})\cdot(\widehat{\psi}_{t_2}-\psi^\star_{t_2})+\sum_{t_3\in[p]}\sum_{t_2\in[p]}\mathbb{C}\mathrm{ov}_{\widetilde{\psi},\widetilde{\Psi}}(u_{t_1},u_{t_3}u_{t_2})\cdot(\widehat{\Psi}_{t_3,t_2}-\Psi^\star_{t_3,t_2}).$$

Now, using the triangle inequality and sub-multiplicativity of induced matrix norms, we have

$$\|\nabla_{\widehat{\psi}}Z(\widehat{\psi},\widehat{\Psi})-\nabla_{\psi^\star}Z(\psi^\star,\Psi^\star)\|_2 \leq \|\mathbb{C}\mathrm{ov}_{\widetilde{\psi},\widetilde{\Psi}}(\mathbf{u},\mathbf{u})\|_{\mathrm{op}}\|(\widehat{\psi}-\psi^\star)\|_2 + \sum_{t_3\in[p]}\|\mathbb{C}\mathrm{ov}_{\widetilde{\psi},\widetilde{\Psi}}(\mathbf{u},u_{t_3}\mathbf{u})\|_{\mathrm{op}}\|(\widehat{\Psi}_{t_3}-\Psi^\star_{t_3})\|_2. \tag{21}$$

Combining (19) and (21) completes the proof. □

B Proof of Prop. 1: Proper loss function

Fix any $\boldsymbol{z}\in\mathcal{Z}^{p_z}$. For every $t\in[p]$, define the following parametric distribution

$$u_{\mathbf{x}|\mathbf{z}}\big(\boldsymbol{x}|\boldsymbol{z};\theta_t(\boldsymbol{z}),\Theta_t\big) \propto \frac{f_{\mathbf{x}|\mathbf{z}}\big(\boldsymbol{x}|\boldsymbol{z};\theta^\star(\boldsymbol{z}),\Theta^\star\big)}{f_{\mathsf{x}_t|\mathbf{x}_{-t},\mathbf{z}}(x_t|\boldsymbol{x}_{-t},\boldsymbol{z};\theta_t(\boldsymbol{z}),\Theta_t)} \tag{22}$$

where $f_{\mathbf{x}|\mathbf{z}}\big(\boldsymbol{x}|\boldsymbol{z};\theta^\star(\boldsymbol{z}),\Theta^\star\big)$ is as defined in (4) and $f_{\mathsf{x}_t|\mathbf{x}_{-t},\mathbf{z}}(x_t|\boldsymbol{x}_{-t},\boldsymbol{z};\theta_t(\boldsymbol{z}),\Theta_t)$ is as defined in (8). Using (8), we can write $u_{\mathbf{x}|\mathbf{z}}\big(\boldsymbol{x}|\boldsymbol{z};\theta_t(\boldsymbol{z}),\Theta_t\big)$ in (22) as

$$u_{\mathbf{x}|\mathbf{z}}\big(\boldsymbol{x}|\boldsymbol{z};\theta_t(\boldsymbol{z}),\Theta_t\big) \propto f_{\mathbf{x}|\mathbf{z}}\big(\boldsymbol{x}|\boldsymbol{z};\theta^\star(\boldsymbol{z}),\Theta^\star\big)\exp\big(-[\theta_t(\boldsymbol{z})+2\Theta_t^\top\boldsymbol{x}]x_t\big).$$

Then, we have

$$u_{\mathbf{x}|\mathbf{z}}\big(\boldsymbol{x}|\boldsymbol{z};\theta_t(\boldsymbol{z}),\Theta_t\big) = \frac{f_{\mathbf{x}|\mathbf{z}}\big(\boldsymbol{x}|\boldsymbol{z};\theta^\star(\boldsymbol{z}),\Theta^\star\big)\exp\big(-[\theta_t(\boldsymbol{z})+2\Theta_t^\top\boldsymbol{x}]x_t\big)}{\int_{\boldsymbol{x}\in\mathcal{X}^p} f_{\mathbf{x}|\mathbf{z}}\big(\boldsymbol{x}|\boldsymbol{z};\theta^\star(\boldsymbol{z}),\Theta^\star\big)\exp\big(-[\theta_t(\boldsymbol{z})+2\Theta_t^\top\boldsymbol{x}]x_t\big)d\boldsymbol{x}}$$

$$= \frac{f_{\mathsf{x}|\mathsf{z}}\big(\boldsymbol{x}|\boldsymbol{z};\theta^\star(\boldsymbol{z}),\Theta^\star\big)\exp\big(-[\theta_t(\boldsymbol{z})+2\Theta_t^\top\boldsymbol{x}]x_t\big)}{\mathbb{E}_{\mathsf{x}|\mathsf{z}}\Big[\exp\big(-[\theta_t(\boldsymbol{z})+2\Theta_t^\top\boldsymbol{x}]x_t\big)\Big]}. \tag{23}$$

Further, for $\theta_t(\boldsymbol{z}) = \theta_t^\star(\boldsymbol{z})$, and $\Theta_t = \Theta_t^\star$, we can write an expression for $u_{\mathsf{x}|\mathsf{z}}\big(\boldsymbol{x}|\boldsymbol{z};\theta_t^\star(\boldsymbol{z}),\Theta_t^\star\big)$ which does not depend on x_t functionally. From (8), we have

$$u_{\mathsf{x}|\mathsf{z}}\big(\boldsymbol{x}|\boldsymbol{z};\theta_t^\star(\boldsymbol{z}),\Theta_t^\star\big) \propto f_{\mathsf{x}_{-t}|\mathsf{z}}\big(\boldsymbol{x}_{-t}|\boldsymbol{z};\theta^\star(\boldsymbol{z}),\Theta^\star\big). \tag{24}$$

Now, consider the difference between $\mathsf{KL}\left(u_{\mathsf{x}|\mathsf{z}}\big(\boldsymbol{x}|\boldsymbol{z};\theta_t^\star(\boldsymbol{z}),\Theta_t^\star\big)\,\middle\|\,u_{\mathsf{x}|\mathsf{z}}\big(\boldsymbol{x}|\boldsymbol{z};\theta_t(\boldsymbol{z}),\Theta_t\big)\right)$ and $\mathsf{KL}\left(u_{\mathsf{x}|\mathsf{z}}\big(\boldsymbol{x}|\boldsymbol{z};\theta_t^\star(\boldsymbol{z}),\Theta_t^\star\big)\,\middle\|\,f_{\mathsf{x}|\mathsf{z}}\big(\boldsymbol{x}|\boldsymbol{z};\theta^\star(\boldsymbol{z}),\Theta^\star\big)\right)$. We have

$$\begin{aligned}
&\mathsf{KL}\left(u_{\mathsf{x}|\mathsf{z}}\big(\cdot|\boldsymbol{z};\theta_t^\star(\boldsymbol{z}),\Theta_t^\star\big)\,\middle\|\,u_{\mathsf{x}|\mathsf{z}}\big(\cdot|\boldsymbol{z};\theta_t(\boldsymbol{z}),\Theta_t\big)\right) - \mathsf{KL}\left(u_{\mathsf{x}|\mathsf{z}}\big(\cdot|\boldsymbol{z};\theta_t^\star(\boldsymbol{z}),\Theta_t^\star\big)\,\middle\|\,f_{\mathsf{x}|\mathsf{z}}\big(\cdot|\boldsymbol{z};\theta^\star(\boldsymbol{z}),\Theta^\star\big)\right)\\
&\overset{(a)}{=} \int_{\boldsymbol{x}\in\mathcal{X}^p} u_{\mathsf{x}|\mathsf{z}}\big(\boldsymbol{x}|\boldsymbol{z};\theta_t^\star(\boldsymbol{z}),\Theta_t^\star\big)\log\frac{f_{\mathsf{x}|\mathsf{z}}\big(\boldsymbol{x}|\boldsymbol{z};\theta^\star(\boldsymbol{z}),\Theta^\star\big)}{u_{\mathsf{x}|\mathsf{z}}\big(\boldsymbol{x}|\boldsymbol{z};\theta_t(\boldsymbol{z}),\Theta_t\big)}d\boldsymbol{x}\\
&\overset{(23)}{=} \int_{\boldsymbol{x}\in\mathcal{X}^p} u_{\mathsf{x}|\mathsf{z}}\big(\boldsymbol{x}|\boldsymbol{z};\theta_t^\star(\boldsymbol{z}),\Theta_t^\star\big)\log\frac{\mathbb{E}_{\mathsf{x}|\mathsf{z}}\Big[\exp\big(-[\theta_t(\boldsymbol{z})+2\Theta_t^\top\boldsymbol{x}]x_t\big)\Big]}{\exp\big(-[\theta_t(\boldsymbol{z})+2\Theta_t^\top\boldsymbol{x}]x_t\big)}d\boldsymbol{x}\\
&= \log\mathbb{E}_{\mathsf{x}|\mathsf{z}}\Big[\exp\big(-[\theta_t(\boldsymbol{z})+2\Theta_t^\top\boldsymbol{x}]x_t\big)\Big] - \int_{\boldsymbol{x}\in\mathcal{X}^p} u_{\mathsf{x}|\mathsf{z}}\big(\boldsymbol{x}|\boldsymbol{z};\theta_t^\star(\boldsymbol{z}),\Theta_t^\star\big)\big([\theta_t(\boldsymbol{z})+2\Theta_t^\top\boldsymbol{x}]x_t\big)d\boldsymbol{x}\\
&\overset{(b)}{=} \log\mathbb{E}_{\mathsf{x}|\mathsf{z}}\Big[\exp\big(-[\theta_t(\boldsymbol{z})+2\Theta_t^\top\boldsymbol{x}]x_t\big)\Big],
\end{aligned} \tag{25}$$

where (a) follows from the definition of KL-divergence and (b) follows because integral is zero since (i) $u_{\mathsf{x}|\mathsf{z}}\big(\boldsymbol{x}|\boldsymbol{z};\theta_t^\star(\boldsymbol{z}),\Theta_t^\star\big)$ does not functionally depend on x_t as in (24), (ii) $\theta_t(\boldsymbol{z})+2\Theta_t^\top\boldsymbol{x}$ does not functionally depend on x_t as $\Theta_{tt} = 0$, and (iii) $\int_{x_t\in\mathcal{X}} x_t dx_t = 0$. Now, we can write

$$\begin{aligned}
\mathbb{E}\big[\mathcal{L}(\underline{\Theta})\big] &= \frac{1}{n}\sum_{t\in[p]}\sum_{i\in[n]}\mathbb{E}\Big[\exp\big(-[\theta_t(\boldsymbol{z}^{(i)})+2\Theta_t^\top\boldsymbol{x}^{(i)}]x_t^{(i)}\big)\Big]\\
&\overset{(25)}{=} \frac{1}{n}\sum_{t\in[p]}\sum_{i\in[n]}\mathbb{E}_{\mathsf{z}}\Big[\exp\Big(\mathsf{KL}\left(u_{\mathsf{x}|\mathsf{z}}\big(\cdot|\boldsymbol{z}^{(i)};\theta_t^\star(\boldsymbol{z}^{(i)}),\Theta_t^\star\big)\,\middle\|\,u_{\mathsf{x}|\mathsf{z}}\big(\cdot|\boldsymbol{z}^{(i)};\theta_t(\boldsymbol{z}^{(i)}),\Theta_t\big)\right)\\
&\qquad - \mathsf{KL}\left(u_{\mathsf{x}|\mathsf{z}}\big(\cdot|\boldsymbol{z}^{(i)};\theta_t^\star(\boldsymbol{z}^{(i)}),\Theta_t^\star\big)\,\middle\|\,f_{\mathsf{x}|\mathsf{z}}\big(\cdot|\boldsymbol{z}^{(i)};\theta^\star(\boldsymbol{z}^{(i)}),\Theta^\star\big)\right)\Big)\Big].
\end{aligned} \tag{26}$$

We note that the parameters only up in the first KL-divergence term in the right-hand-side of (26). Therefore, it is easy to see that $\mathbb{E}\big[\mathcal{L}(\underline{\Theta})\big]$ is minimized uniquely when $\theta_t(\boldsymbol{z}^{(i)}) = \theta_t^\star(\boldsymbol{z}^{(i)})$ and $\Theta_t = \Theta_t^\star$ for all $t \in [p]$ and all $i \in [n]$, i.e., when $\underline{\Theta} = \underline{\Theta}^\star$.

C Proof of Thm. 1 Part I: Recovering population-level parameter

To analyze our estimate of the population-level parameter, we note that the set Λ_Θ in (7) places independent constraints on the rows of Θ[5]. Therefore, we look at p independent convex optimization problems by decomposing the loss function $\mathcal{L}$ in (9) and the estimate $\widehat{\underline{\Theta}}$ in (10) as follows:

$$\mathcal{L}_t(\underline{\Theta}_t) \triangleq \frac{1}{n}\sum_{i\in[n]}\exp\Big(-[\theta_t^{(i)}+2\Theta_t^\top\boldsymbol{x}_{-t}]x_t^{(i)}\Big) \text{ and } \widehat{\underline{\Theta}}_t \in \operatorname*{arg\,min}_{\underline{\Theta}\in\Lambda_\theta^n\times\Lambda_\Theta} \mathcal{L}_t(\underline{\Theta}_t) \text{ for all } t\in[p], \tag{27}$$

where $\underline{\Theta}_t = \big\{\theta_t^{(1)},\cdots,\theta_t^{(n)},\Theta_t\big\}$ as defined in (9). Fix any $t \in [p]$. From (27), we have $\mathcal{L}_t(\widehat{\underline{\Theta}}_t) \le \mathcal{L}_t(\underline{\Theta}_t^\star)$. Using contraposition, to prove this part, it is sufficient to show that all points Θ_t that satisfy $\|\Theta_t - \Theta_t^\star\|_2 \ge \varepsilon$ also uniformly satisfy

$$\mathcal{L}_t(\underline{\Theta}_t) \ge \mathcal{L}_t(\underline{\Theta}_t^\star) + O(\varepsilon^2) \text{ for } n \ge \frac{c\exp(c\beta)\log\frac{p}{\delta}}{\varepsilon^4}, \tag{28}$$

[5] To ensure that the final estimate is symmetric, we can take the average of Θ and $\Theta^\top$.

with probability at least $1-\delta$, uniformly for all $t \in [p]$. The guarantee in Thm. 1 follows.

To that end, first, we claim that for any fixed $\underline{\Theta}_t$, if Θ_t is far from $\Theta_t^\star$, then with high probability $\mathcal{L}_t\big(\underline{\Theta}_t\big)$ will be significantly larger than $\mathcal{L}_t\big(\underline{\Theta}_t^\star\big)$. We provide a proof in App. C.1. First, we define the following constants that depend on model-parameters $\tau \triangleq (\alpha, \beta, x_{\max})$:

$$C_{1,\tau} \triangleq \alpha(1+4\beta x_{\max}) \quad \text{and} \quad C_{2,\tau} \triangleq \exp\left(\alpha x_{\max}(1+2\beta x_{\max})\right). \tag{29}$$

Lemma 2 (**Gap between the loss function for a fixed parameter**). *Consider any $\underline{\Theta} \in \Lambda_\theta^n \times \Lambda_\Theta$. Fix any $\delta \in (0,1)$. Then, we have uniformly for all $t \in [p]$*

$$\mathcal{L}_t\big(\underline{\Theta}_t\big) \geq \mathcal{L}_t\big(\underline{\Theta}_t^\star\big) + \frac{x_{\max}^2}{2\pi^2 e^2(\beta+1)C_{2,\tau}^5}\|\Theta_t - \Theta_t^\star\|_2^2 \quad \textit{for} \quad n \geq \frac{c\exp(c\beta)\log(p/\delta)}{\|\Omega_t\|_2^4},$$

with probability at least $1-\delta$ where $C_{2,\tau}$ was defined in (29).

Next, we claim that the loss function $\mathcal{L}_t$ is Lipschitz and capture this property via the following lemma. We provide a proof in App. C.2.

Lemma 3 (**Lipschitzness of the loss function**). *Consider any $\underline{\Theta}, \underline{\widetilde{\Theta}} \in \Lambda_\theta^n \times \Lambda_\Theta$ such that $\theta^{(i)} = \widetilde{\theta}^{(i)}$ for all $i \in [n]$. Fix any $t \in [p]$. Then, the loss function $\mathcal{L}_t$ is Lipschitz with respect to the ℓ_1 norm $\|\cdot\|_1$ and with Lipschitz constant $x_{\max}^2 C_{2,\tau}$, i.e.,*

$$\big|\mathcal{L}_t\big(\underline{\widetilde{\Theta}}_t\big) - \mathcal{L}_t\big(\underline{\Theta}_t\big)\big| \leq x_{\max}^2 C_{2,\tau}\|\widetilde{\Theta}_t - \Theta_t\|_1, \tag{30}$$

where the constant $C_{2,\tau}$ was defined in (29).

Given these lemmas, we now proceed with the proof.

Proof strategy: As mentioned earlier, the idea is to show that all points $\Theta_t \in \Lambda_\Theta$ that satisfy $\|\Theta_t - \Theta_t^\star\|_2 \geq \varepsilon$ also uniformly satisfy (28) with probability at least $1-\delta$. To do so, we consider the set of points $\Lambda_\Theta^{\varepsilon,t} \subset \Lambda_\Theta$ whose distance from $\Theta_t^\star$ is at least $\varepsilon > 0$ in ℓ_2 norm. Then, using an appropriate covering set of $\Lambda_\Theta^{\varepsilon,t}$ and the Lipschitzness of $\mathcal{L}_t$, we show that the value of $\mathcal{L}_t$ at all points in $\Lambda_\Theta^{\varepsilon,t}$ is uniformly $O(\varepsilon^2)$ larger than the value of $\mathcal{L}_t$ at $\Theta_t^\star$ with high probability. We ensure that the failure probability smaller than δ.

Gap between the loss function for all parameters in the covering set: Consider the set of elements $\Lambda_\Theta^{\varepsilon,t} \triangleq \{\Theta_t \in \Lambda_\Theta : \|\Theta_t^\star - \Theta_t\|_2 \geq \varepsilon\}$. Let $\mathcal{C}(\Lambda_\Theta^{\varepsilon,t}, \varepsilon')$ be the ε'-covering number of the set $\Lambda_\Theta^{\varepsilon,t}$ and let $\mathcal{U}(\Lambda_\Theta^{\varepsilon,t}, \varepsilon')$ be the associated ε'-cover (see Def. 2) where

$$\varepsilon' \triangleq \frac{\varepsilon^2}{8\pi^2 e^2 \beta(\beta+1)C_{2,\tau}^6} \tag{31}$$

Now, we apply Lem. 2 to each element in $\mathcal{U}(\Lambda_\Theta^{\varepsilon,t}, \varepsilon')$ and argue by a union bound that the value of $\mathcal{L}_t$ at all points in $\mathcal{U}(\Lambda_\Theta^{\varepsilon,t}, \varepsilon')$ is uniformly $O(\varepsilon^2)$ larger than the value of $\mathcal{L}_t$ at $\Theta_t^\star$ with high probability. We start by considering any $\Theta_t \in \mathcal{U}(\Lambda_\Theta^{\varepsilon,t}, \varepsilon')$. We have

$$\|\Theta_t^\star - \Theta_t\|_2 \overset{(a)}{\geq} \varepsilon, \tag{32}$$

where (a) follows because $\mathcal{U}(\Lambda_\Theta^{\varepsilon,t}, \varepsilon') \subseteq \Lambda_\Theta^{\varepsilon,t}$. Now, applying Lem. 2 with $\delta_0 = \delta/\mathcal{C}(\Lambda_\Theta^{\varepsilon,t}, \varepsilon')$, we have

$$\mathcal{L}_t\big(\Theta_t\big) \geq \mathcal{L}_t\big(\Theta_t^\star\big) + \frac{x_{\max}^2}{2\pi^2 e^2(\beta+1)C_{2,\tau}^5}\|\Theta_t^\star - \Theta_t\|_2^2 \overset{(32)}{\geq} \mathcal{L}_t\big(\Theta_t^\star\big) + \frac{x_{\max}^2\varepsilon^2}{2\pi^2 e^2(\beta+1)C_{2,\tau}^5},$$

with probability at least $1-\delta/\mathcal{C}(\Lambda_\Theta^{\varepsilon,t}, \varepsilon')$ whenever

$$n \geq \frac{c\exp(c\beta)\log\big(\mathcal{C}(\Lambda_\Theta^{\varepsilon,t}, \varepsilon')\cdot p/\delta\big)}{\|\Theta_t^\star - \Theta_t\|_2^4}.$$

Using (32), it suffices to choose n such that

$$n \geq \frac{c \exp(c\beta) \log\left(\mathcal{C}(\Lambda_\Theta^{\varepsilon,t}, \varepsilon') \cdot p/\delta\right)}{\varepsilon^4}. \tag{33}$$

By applying the union bound over $\mathcal{U}(\Lambda_\Theta^{\varepsilon,t}, \varepsilon')$, as long as n satisfies (33), we have

$$\mathcal{L}_t(\Theta_t) \geq \mathcal{L}_t(\Theta_t^\star) + \frac{x_{\max}^2 \varepsilon^2}{2\pi^2 e^2 (\beta+1) C_{2,\tau}^5} \text{ uniformly for every } \Theta_t \in \mathcal{U}(\Lambda_\Theta^{\varepsilon,t}, \varepsilon'), \tag{34}$$

with probability at least $1 - \delta$.

Generalize beyond the covering set: Next, we assume that (34) holds, and generalize from $\mathcal{U}(\Lambda_\Theta^{\varepsilon,t}, \varepsilon')$ to all of $\Lambda_\Theta^{\varepsilon,t}$. Consider any $\widetilde{\Theta}_t \in \Lambda_\Theta^{\varepsilon,t}$ and let Θ_t be any point in $\mathcal{U}(\Lambda_\Theta^{\varepsilon,t}, \varepsilon')$ that satisfies $\|\Theta_t - \widetilde{\Theta}_t\|_2 \leq \varepsilon'$ (see Def. 2). Then, from Lem. 3, we have

$$\begin{aligned}
\mathcal{L}_t(\underline{\widetilde{\Theta}}_t) \geq \mathcal{L}_t(\underline{\Theta}_t) - x_{\max}^2 C_{2,\tau} \|\Theta_t - \widetilde{\Theta}_t\|_1 &\overset{(a)}{\geq} \mathcal{L}_t(\Theta_t) - 2x_{\max}^2 C_{2,\tau} \beta \|\Theta_t - \widetilde{\Theta}_t\|_2 \\
&\overset{(b)}{\geq} \mathcal{L}_t(\Theta_t) - 2x_{\max}^2 C_{2,\tau} \beta \varepsilon' \\
&\overset{(31)}{\geq} \mathcal{L}_t(\Theta_t) - \frac{x_{\max}^2 \varepsilon^2}{4\pi^2 e^2 (\beta+1) C_{2,\tau}^5} \\
&\overset{(34)}{\geq} \mathcal{L}_t(\Theta_t^\star) + \frac{x_{\max}^2 \varepsilon^2}{4\pi^2 e^2 (\beta+1) C_{2,\tau}^5},
\end{aligned}$$

where (a) follows by using $\|\Theta_t - \widetilde{\Theta}_t\|_1 \leq 2\beta \|\Theta_t - \widetilde{\Theta}_t\|_2$ which follows from the order of norms on Euclidean space as well as $\Theta \in \Lambda_\Theta$ and $\widetilde{\Theta} \in \Lambda_\Theta$ and (b) follows because $\|\Theta_t - \widetilde{\Theta}_t\|_2 \leq \varepsilon'$. Recall that $\widetilde{\Theta}_t$ is any point in $\Lambda_\Theta^{\varepsilon,t}$, i.e., $\|\Theta_t^\star - \widetilde{\Theta}_t\|_2 \geq r$. Therefore, we have an inequality that looks like (28).

Bounding n: To bound n in (54), we bound the covering number $\mathcal{C}(\Lambda_\Theta^{\varepsilon,t}, \varepsilon')$ as follows

$$\mathcal{C}(\Lambda_\Theta^{\varepsilon,t}, \varepsilon') \overset{(a)}{\leq} \mathcal{C}(\Lambda_\Theta, \varepsilon'/2) \tag{35}$$

where (a) follows from the fact that for any sets $\mathcal{U} \subseteq \mathcal{V}$ and any ε, it holds that $\mathcal{C}(\mathcal{U}, \varepsilon) \leq \mathcal{C}(\mathcal{V}, \varepsilon/2)$. Then using (35) in (54) and observing that $\varepsilon' = \frac{\varepsilon^2}{c\exp(c\beta)}$, it is sufficient for

$$n \geq \frac{c\exp(c\beta)}{\varepsilon^4} \cdot \left(\log \frac{p}{\delta} + \mathcal{M}_\Theta\left(\frac{\varepsilon^2}{c\exp(c\beta)} \right) \right).$$

The proof is complete by noting that $\mathcal{M}_\Theta\left(\frac{\varepsilon^2}{c\exp(c\beta)}\right) = O(\beta \log p)$ since $\Theta_t \in \mathbb{R}^{p\times 1}$ is such that $\|\Theta_t\|_0 \leq \beta$.

C.1 Proof of Lem. 2: Gap between the loss function for a fixed parameter

Fix any $\varepsilon > 0$, any $\delta \in (0,1)$, and $t \in [p]$. Consider any direction $\underline{\Omega}_t \triangleq \{\omega_t^{(1)}, \cdots, \omega_t^{(n)}, \Omega_t\} \in \mathbb{R}^{n+p}$ along the parameter $\underline{\Theta}_t$, i.e.,

$$\underline{\Omega}_t = \underline{\Theta}_t - \underline{\Theta}_t^\star, \quad \text{and} \quad \Omega_t = \Theta_t - \Theta_t^\star. \tag{36}$$

Without loss of generality, we let $\Omega_{tt} = 0$ since $\Theta_{tt}^\star = 0$. We denote the first-order and the second-order directional derivatives of the loss function $\mathcal{L}_t$ in (27) along the direction $\underline{\Omega}_t$ evaluated at $\underline{\Theta}_t$ by $\partial_{\underline{\Omega}_t} \mathcal{L}_t(\underline{\Theta}_t)$ and $\partial^2_{\underline{\Omega}_t^2} \mathcal{L}_t(\underline{\Theta}_t)$ respectively. Below, we state a lemma (with proof divided across App. C.1.1 and App. C.1.2) that provides us a control on $\partial_{\underline{\Omega}_t} \mathcal{L}_t(\underline{\Theta}_t)$ and $\partial^2_{\underline{\Omega}_t^2} \mathcal{L}_t(\underline{\Theta}_t)$. The assumptions of Lem. 2 remain in force.

Lemma 4 (Control on first and second directional derivatives). *For any fixed $\varepsilon_1, \varepsilon_2 > 0$, $\delta_1, \delta_2 \in (0,1)$, $t \in [p]$, $\underline{\Theta} \in \Lambda_\theta^n \times \Lambda_\Theta$ defined in (9) with $\underline{\Omega}_t$ and Ω_t defined in (36), we have the following:*

(a) **Concentration of first directional derivative**: *with probability at least* $1-\delta_1$,

$$\left|\partial_{\underline{\Omega}_t}\mathcal{L}_t(\underline{\Theta}_t^\star)\right| \leq \varepsilon_1 \quad \text{for} \quad n \geq \frac{8C_{1,\tau}^2 C_{2,\tau}^2 x_{\max}^2 \cdot \log(2p/\delta_1)}{\varepsilon_1^2},$$

and uniformly for all $t \in [p]$.

(b) **Anti-concentration of second directional derivative**: *with probability at least* $1-\delta_2$,

$$\partial^2_{\underline{\Omega}_t^2}\mathcal{L}_t(\underline{\Theta}_t) \geq \frac{4x_{\max}^2\|\Omega_t\|_2^2}{\pi^2 e^2(\beta+1)C_{2,\tau}^5} - \varepsilon_2 \quad \text{for} \quad n \geq \frac{32C_{1,\tau}^4 x_{\max}^4 \cdot \log(2p/\delta_2)}{\varepsilon_2^2 C_{2,\tau}^2},$$

and uniformly for all $t \in [p]$.

Given this lemma, we now proceed with the proof. Define a function $g : [0,1] \to \mathbb{R}^{n+p}$

$$g(a) \triangleq \underline{\Theta}_t^\star + a(\underline{\Theta}_t - \underline{\Theta}_t^\star).$$

Notice that $g(0) = \underline{\Theta}_t^\star$ and $g(1) = \underline{\Theta}_t$ as well as

$$\frac{d\mathcal{L}_t(g(a))}{da} = \partial_{\underline{\Omega}_t}\mathcal{L}_t(\underline{\Theta}_t)\big|_{\underline{\Theta}_t = g(a)} \quad \text{and} \quad \frac{d^2\mathcal{L}_t(g(a))}{da^2} = \partial^2_{\underline{\Omega}_t^2}\mathcal{L}_t(\underline{\Theta}_t)\big|_{\underline{\Theta}_t = g(a)}. \tag{37}$$

By the fundamental theorem of calculus, we have

$$\frac{d\mathcal{L}_t(g(a))}{da} \geq \frac{d\mathcal{L}_t(g(a))}{da}\Big|_{a=0} + a \min_{a\in(0,1)} \frac{d^2\mathcal{L}_t(g(a))}{da^2}. \tag{38}$$

Integrating both sides of (38) with respect to a, we obtain

$$\begin{aligned}
\mathcal{L}_t(g(a)) - \mathcal{L}_t(g(0)) &\geq a\frac{d\mathcal{L}_t(g(a))}{da}\Big|_{a=0} + \frac{a^2}{2}\min_{a\in(0,1)}\frac{d^2\mathcal{L}_t(g(a))}{da^2} \\
&\overset{(37)}{=} a\partial_{\underline{\Omega}_t}\mathcal{L}_t(\underline{\Theta}_t)\big|_{\underline{\Theta}_t = g(0)} + \frac{a^2}{2}\min_{a\in(0,1)}\partial^2_{\underline{\Omega}_t^2}\mathcal{L}_t(\underline{\Theta}_t)\big|_{\underline{\Theta}_t = g(a)} \\
&\overset{(a)}{=} a\partial_{\underline{\Omega}_t}\mathcal{L}_t(\underline{\Theta}_t^\star) + \frac{a^2}{2}\min_{a\in(0,1)}\partial^2_{\underline{\Omega}_t^2}\mathcal{L}_t(\underline{\Theta}_t)\big|_{\underline{\Theta}_t = g(a)} \\
&\overset{(b)}{\geq} -a\left|\partial_{\underline{\Omega}_t}\mathcal{L}_t(\underline{\Theta}_t^\star)\right| + \frac{a^2}{2}\min_{a\in(0,1)}\partial^2_{\underline{\Omega}_t^2}\mathcal{L}_t(\underline{\Theta}_t)\big|_{\underline{\Theta}_t = g(a)},
\end{aligned} \tag{39}$$

where (a) follows because $g(0) = \underline{\Theta}_t^\star$ and (b) follows by the triangle inequality. Plugging in $a=1$ in (39) as well as using $g(0) = \underline{\Theta}_t^\star$ and $g(1) = \underline{\Theta}_t$, we find that

$$\mathcal{L}_t(\underline{\Theta}_t) - \mathcal{L}_t(\underline{\Theta}_t^\star) \geq -\left|\partial_{\underline{\Omega}_t}\mathcal{L}_t(\underline{\Theta}_t^\star)\right| + \frac{1}{2}\min_{a\in(0,1)}\partial^2_{\underline{\Omega}_t^2}\mathcal{L}_t(\underline{\Theta}_t)\big|_{\underline{\Theta}_t = g(a)}.$$

Now, we use Lem. 4 with

$$\varepsilon_1 = \frac{x_{\max}^2\|\Omega_t\|_2^2}{2\pi^2e^2(\beta+1)C_{2,\tau}^5}, \quad \delta_1 = \frac{\delta}{2}, \text{ and } \varepsilon_2 = \frac{2x_{\max}^2\|\Omega_t\|_2^2}{\pi^2e^2(\beta+1)C_{2,\tau}^5}, \quad \delta_2 = \frac{\delta}{2}.$$

Therefore, with probability at least $1-\delta$ as long as $n \geq O\left(\frac{\exp\left(O(\beta)\right)\log(p/\delta)}{\|\Omega_t\|_2^4}\right)$, we have uniformly for all $t \in [p]$

$$\begin{aligned}
\mathcal{L}_t(\underline{\Theta}_t) - \mathcal{L}_t(\underline{\Theta}_t^\star) &\geq -\frac{x_{\max}^2\|\Omega_t\|_2^2}{2\pi^2e^2(\beta+1)C_{2,\tau}^5} + \frac{1}{2}\left(\frac{4x_{\max}^2\|\Omega_t\|_2^2}{\pi^2e^2(\beta+1)C_{2,\tau}^5} - \frac{2x_{\max}^2\|\Omega_t\|_2^2}{\pi^2e^2(\beta+1)C_{2,\tau}^5}\right) \\
&= \frac{x_{\max}^2\|\Omega_t\|_2^2}{2\pi^2e^2(\beta+1)C_{2,\tau}^5}.
\end{aligned}$$

C.1.1 Proof of Lem. 4 (a): Concentration of first directional derivative

For every $t \in [p]$ with $\underline{\Omega}_t$ defined in (36), we claim that the first-order directional derivative of the loss function defined in (27) is given by

$$\partial_{\underline{\Omega}_t} \mathcal{L}_t(\underline{\Theta}_t) = -\frac{1}{n} \sum_{i \in [n]} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)} \Big) \exp\Big(-[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}] x_t^{(i)} \Big), \tag{40}$$

where $\Delta_t^{(i)} \triangleq \begin{bmatrix} \omega_t^{(i)} \\ 2\Omega_t \end{bmatrix} \in \mathbb{R}^{p+1}$ and $\widetilde{\boldsymbol{x}}^{(i)} \triangleq \begin{bmatrix} 1 \\ \boldsymbol{x}^{(i)} \end{bmatrix} \in \mathbb{R}^{p+1}$ for all $i \in [n]$. We provide a proof in App. C.1.1.

Next, we claim that the mean of the first-order directional derivative evaluated at the true parameter is zero. We provide a proof in App. C.1.1

Lemma 5 (Zero-meanness of first directional derivative). *For every $t \in [p]$ with $\underline{\Omega}_t$ defined in (36), we have $\mathbb{E}\big[\partial_{\underline{\Omega}_t} \mathcal{L}_t(\underline{\Theta}_t^\star)\big] = 0$.*

Given these, we proceed to show the concentration of the first-order directional derivative evaluated at the true parameter. Fix any $t \in [p]$. From (40), we have

$$\partial_{\underline{\Omega}_t} \mathcal{L}_t(\underline{\Theta}_t^\star) \overset{(40)}{=} -\frac{1}{n} \sum_{i \in [n]} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)} \Big) \exp\Big(-[\theta_t^{\star(i)} + 2\Theta_t^{\star\top} \boldsymbol{x}^{(i)}] x_t^{(i)} \Big).$$

Each term in the above summation is an independent random variable and is bounded as follows:

$$\begin{aligned}
&\Big| \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)} \Big) \times \exp\Big(-[\theta_t^{\star(i)} + 2\Theta_t^{\star\top} \boldsymbol{x}^{(i)}] x_t^{(i)} \Big) \Big| \\
&\overset{(a)}{=} \Big| \Big(\omega_t^{(i)} x_t^{(i)} + 2\Omega_t^\top \boldsymbol{x}^{(i)} x_t^{(i)} \Big) \times \exp\Big(-[\theta_t^{\star(i)} + 2\Theta_t^{\star\top} \boldsymbol{x}^{(i)}] x_t^{(i)} \Big) \Big| \\
&\overset{(b)}{\leq} \big| \omega_t^{(i)} + 2\Omega_t^\top \boldsymbol{x}^{(i)} \big| \times x_{\max} \times \exp\Big(|\theta_t^{\star(i)} + 2\Theta_t^{\star\top} \boldsymbol{x}^{(i)}| x_{\max} \Big) \\
&\overset{(c)}{\leq} \big| |\omega_t^{(i)}| + 2\|\Omega_t\|_1 \|\boldsymbol{x}^{(i)}\|_\infty \big| \times x_{\max} \times \exp\Big(\big(|\theta_t^{\star(i)}| + 2\|\Theta_t^\star\|_1 \|\boldsymbol{x}^{(i)}\|_\infty \big) x_{\max} \Big) \\
&\overset{(d)}{\leq} \big(2\alpha + 8\alpha\beta x_{\max} \big) \times x_{\max} \times \exp\Big((\alpha + 2\alpha\beta x_{\max}) x_{\max} \Big) \overset{(29)}{=} 2C_{1,\tau} C_{2,\tau} x_{\max},
\end{aligned}$$

where (a) follows by plugging in $\Delta_t^{(i)}$ and $\widetilde{\boldsymbol{x}}^{(i)}$, (b) follows because $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$, (c) follows by triangle inequality and Cauchy–Schwarz inequality, and (d) follows because $\theta^{\star(i)} \in \Lambda_\theta$ for all $i \in [n]$, $\Theta^\star \in \Lambda_\Theta$, $\omega^{(i)} \in 2\Lambda_\theta$ for all $i \in [n]$, $\Omega \in 2\Lambda_\Theta$, and $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$.

Further, from Lem. 5, we have $\mathbb{E}\big[\partial_{\underline{\Omega}_t} \mathcal{L}_t(\underline{\Theta}_t^\star)\big] = 0$. Therefore, using the Hoeffding's inequality results in

$$\mathbb{P}\Big(\big| \partial_{\underline{\Omega}_t} \mathcal{L}_t(\underline{\Theta}_t^\star) \big| > \varepsilon_1 \Big) < 2 \exp\Big(-\frac{n\varepsilon_1^2}{8C_{1,\tau}^2 C_{2,\tau}^2 x_{\max}^2} \Big).$$

The proof follows by using the union bound over all $t \in [p]$.

Proof of (40): Expression for first directional derivative: Fix any $t \in [p]$. The first-order partial derivatives of $\mathcal{L}_t$ with respect to entries of $\underline{\Theta}_t$ defined in (27) are given by

$$\frac{\partial \mathcal{L}_t(\underline{\Theta}_t)}{\partial \theta_t^{(i)}} = -\frac{1}{n} x_t^{(i)} \exp\Big(-[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}] x_t^{(i)} \Big) \quad \text{for all} \quad i \in [n], \quad \text{and}$$

$$\frac{\partial \mathcal{L}_t(\underline{\Theta}_t)}{\partial \Theta_{tu}} = \begin{cases} -\frac{2}{n} \sum_{i \in [n]} x_t^{(i)} x_u^{(i)} \exp\Big(-[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}] x_t^{(i)} \Big) & \text{for all} \quad u \in [p] \setminus \{t\}. \\ 0 \quad \text{for} \quad u = t \end{cases}$$

Now, we can write the first-order directional derivative of $\mathcal{L}_t$ as

$$\partial_{\underline{\Omega}_t} \mathcal{L}_t(\underline{\Theta}_t) \triangleq \lim_{h \to 0} \frac{\mathcal{L}_t(\underline{\Theta}_t + h\underline{\Omega}_t) - \mathcal{L}_t(\underline{\Theta}_t)}{h} = \sum_{i \in [n]} \omega_t^{(i)} \frac{\partial \mathcal{L}_t(\underline{\Theta}_t)}{\partial \theta_t^{(i)}} + \sum_{u \in [p]} \Omega_{tu} \frac{\partial \mathcal{L}_t(\underline{\Theta}_t)}{\partial \Theta_{tu}}$$

$$= -\frac{1}{n}\sum_{i\in[n]}\Big(\omega_t^{(i)}x_t^{(i)} + 2\sum_{u\in[p]\setminus\{t\}}\Omega_{tu}x_t^{(i)}x_u^{(i)}\Big)\exp\Big(-[\theta_t^{(i)}+2\Theta_t^\top \boldsymbol{x}^{(i)}]x_t^{(i)}\Big)$$
$$= -\frac{1}{n}\sum_{i\in[n]}\Big(\omega_t^{(i)}x_t^{(i)} + 2\Omega_t^\top \boldsymbol{x}^{(i)}x_t^{(i)}\Big)\exp\Big(-[\theta_t^{(i)}+2\Theta_t^\top \boldsymbol{x}^{(i)}]x_t^{(i)}\Big)$$
$$\stackrel{(a)}{=} -\frac{1}{n}\sum_{i\in[n]}\Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)}x_t^{(i)}\Big)\exp\Big(-[\theta_t^{(i)}+2\Theta_t^\top \boldsymbol{x}^{(i)}]x_t^{(i)}\Big),$$

where (a) follows from the definitions of $\Delta_t^{(i)}$ and $\widetilde{\boldsymbol{x}}^{(i)}$.

Proof of Lem. 5: Zero-meanness of first directional derivative: Fix any $t\in[p]$. From (40), we have

$$\mathbb{E}\big[\partial_{\underline{\Omega}_t}\mathcal{L}_t(\underline{\Theta}_t^\star)\big] \stackrel{(40)}{=} -\frac{1}{n}\sum_{i\in[n]}\mathbb{E}\Big[\Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)}x_t^{(i)}\Big)\exp\Big(-[\theta_t^{\star(i)}+2\Theta_t^{\star\top}\boldsymbol{x}^{(i)}]x_t^{(i)}\Big)\Big]$$
$$\stackrel{(a)}{=} -\frac{1}{n}\sum_{i\in[n]}\sum_{u\in[p+1]}\Delta_{tu}^{(i)}\mathbb{E}_{\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i)}}\Big[\widetilde{\boldsymbol{x}}_u^{(i)}x_t^{(i)}\exp\Big(-[\theta_t^\star(\boldsymbol{z}^{(i)})+2\Theta_t^{\star\top}\boldsymbol{x}^{(i)}]x_t^{(i)}\Big)\Big]$$
$$\stackrel{(b)}{=} -\frac{1}{n}\sum_{i\in[n]}\sum_{u\in[p+1]\setminus\{i+1\}}\Delta_{tu}^{(i)}\mathbb{E}_{\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i)}}\Big[\widetilde{\boldsymbol{x}}_u^{(i)}x_t^{(i)}\exp\Big(-[\theta_t^\star(\boldsymbol{z}^{(i)})+2\Theta_t^{\star\top}\boldsymbol{x}^{(i)}]x_t^{(i)}\Big)\Big],$$

where (a) follows by linearity of expectation and by plugging in $\theta_t^{\star(i)}=\theta_t^\star(\boldsymbol{z}^{(i)})$ and (b) follows because $\Delta_{tu}^{(i)}=\Omega_{tt}=0$ for $u=t+1$ for every $i\in[n]$. To complete the proof, we show

$$\mathbb{E}_{\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i)}}\Big[\widetilde{\boldsymbol{x}}_u^{(i)}x_t^{(i)}\exp\Big(-[\theta_t^\star(\boldsymbol{z}^{(i)})+2\Theta_t^{\star\top}\boldsymbol{x}^{(i)}]x_t^{(i)}\Big)\Big]=0,$$

for all $t\in[p], i\in[n], u\in[p+1]\setminus\{t+1\}$. To that end, fix any $t\in[p], i\in[n], u\in[p+1]\setminus\{t+1\}$. We have

$$\mathbb{E}_{\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i)}}\Big[\widetilde{\boldsymbol{x}}_u^{(i)}x_t^{(i)}\exp\Big(-[\theta_t^\star(\boldsymbol{z}^{(i)})+2\Theta_t^{\star\top}\boldsymbol{x}^{(i)}]x_t^{(i)}\Big)\Big]$$
$$=\int_{\mathcal{X}^p\times\mathcal{Z}^{p_z}}\widetilde{x}_u^{(i)}x_t^{(i)}\exp\Big(-[\theta_t^\star(\boldsymbol{z}^{(i)})+2\Theta_t^{\star\top}\boldsymbol{x}^{(i)}]x_t^{(i)}\Big)f_{\mathbf{x},\mathbf{z}}\big(\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i)}\big)d\boldsymbol{x}^{(i)}d\boldsymbol{z}^{(i)}$$
$$=\int_{\mathcal{X}^p\times\mathcal{Z}^{p_z}}\widetilde{x}_u^{(i)}x_t^{(i)}\exp\Big(-[\theta_t^\star(\boldsymbol{z}^{(i)})+2\Theta_t^{\star\top}\boldsymbol{x}^{(i)}]x_t^{(i)}\Big)f_{\mathbf{x}_{-t},\mathbf{z}}\big(\boldsymbol{x}_{-t}^{(i)},\boldsymbol{z}^{(i)}\big)$$
$$\times f_{\mathbf{x}_t|\mathbf{x}_{-t},\mathbf{z}}\big(x_t^{(i)}|\boldsymbol{x}_{-t}^{(i)},\boldsymbol{z}^{(i)};\theta_t^\star(\boldsymbol{z}^{(i)}),\Theta_t^\star\big)d\boldsymbol{x}^{(i)}d\boldsymbol{z}^{(i)}$$
$$\stackrel{(a)}{=}\int_{\mathcal{X}^p\times\mathcal{Z}^{p_z}}\frac{\widetilde{x}_u^{(i)}x_t^{(i)}f_{\mathbf{x}_{-t},\mathbf{z}}\big(\boldsymbol{x}_{-t}^{(i)},\boldsymbol{z}^{(i)}\big)d\boldsymbol{x}^{(i)}d\boldsymbol{z}^{(i)}}{\int_{\mathcal{X}}\exp\Big([\theta_t^\star(\boldsymbol{z}^{(i)})+2\Theta_t^{\star\top}\boldsymbol{x}^{(i)}]x_t^{(i)}\Big)dx_t^{(i)}}$$
$$\stackrel{(b)}{=}\int_{\mathcal{X}^{p-1}\times\mathcal{Z}^{p_z}}\Big[\int_{\mathcal{X}}x_t^{(i)}dx_t^{(i)}\Big]\frac{\widetilde{x}_u^{(i)}f_{\mathbf{x}_{-t},\mathbf{z}}\big(\boldsymbol{x}_{-t}^{(i)},\boldsymbol{z}^{(i)}\big)d\boldsymbol{x}_{-t}^{(i)}d\boldsymbol{z}^{(i)}}{\int_{\mathcal{X}}\exp\Big([\theta_t^\star(\boldsymbol{z}^{(i)})+2\Theta_t^{\star\top}\boldsymbol{x}^{(i)}]x_t^{(i)}\Big)dx_t^{(i)}}$$
$$\stackrel{(c)}{=}\ 0.$$

where (a) follows by plugging in $f_{\mathbf{x}_t|\mathbf{x}_{-t},\mathbf{z}}\big(x_t^{(i)}|\boldsymbol{x}_{-t}^{(i)},\boldsymbol{z}^{(i)};\theta_t^\star(\boldsymbol{z}^{(i)}),\Theta_t^\star\big)$ from (8), (b) follows by re-arranging and observing that $\widetilde{x}_u^{(i)}\neq x_t^{(i)}$ for any $u\in[p+1]\setminus\{t+1\}$, and (c) follows because $\int_{\mathcal{X}}x_t^{(i)}dx_t^{(i)}=0$ when $\mathcal{X}$ is symmetric around 0.

C.1.2 Proof of Lem. 4 (b): Anti-concentration of second directional derivative

We start by claiming that the second-order directional derivative can be lower bounded by a quadratic form. We provide a proof in App. C.1.2.

Lemma 6 (Lower bound on the second directional derivative). *For every $t \in [p]$ with $\underline{\Omega}_t$ defined in (36), we have*

$$\partial^2_{\underline{\Omega}_t^2} \mathcal{L}_t(\underline{\Theta}_t) \geq \frac{1}{nC_{2,\tau}} \sum_{i \in [n]} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)} \Big)^2,$$

where $\Delta_t^{(i)} \triangleq \begin{bmatrix} \omega_t^{(i)} \\ 2\Omega_t \end{bmatrix} \in \mathbb{R}^{p+1}$ and $\widetilde{\boldsymbol{x}}^{(i)} \triangleq \begin{bmatrix} 1 \\ \boldsymbol{x}^{(i)} \end{bmatrix} \in \mathbb{R}^{p+1}$ for all $i \in [n]$, and the constant $C_{2,\tau}$ was defined in (29).

Next, we claim that the conditional variance of $x_t^{(i)}$ conditioned on $\mathbf{x}_{-t} = \boldsymbol{x}_{-t}^{(i)}$ and $\mathbf{z} = \boldsymbol{z}^{(i)}$ is lower bounded by a constant for every $t \in [p]$ and $i \in [n]$. We provide a proof in App. C.1.2.

Lemma 7 (Lower bound on the conditional variance). *For every $t \in [p]$ and $i \in [n]$, we have*

$$\mathbb{V}ar\big(x_t^{(i)} \big| \boldsymbol{x}_{-t}^{(i)}, \boldsymbol{z}^{(i)}\big) \geq \frac{x_{\max}}{\pi e C_{2,\tau}^2},$$

where the constant $C_{2,\tau}$ was defined in (29).

Given these, we proceed to show the anti-concentration of the second-order directional derivative. Fix any $t \in [p]$ and any $\underline{\Theta} \in \Lambda_\theta^n \times \Lambda_\Theta$. From Lem. 6, we have

$$\partial^2_{\underline{\Omega}_t^2} \mathcal{L}_t(\underline{\Theta}_t) \geq \frac{1}{nC_{2,\tau}} \sum_{i \in [n]} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)} \Big)^2. \tag{41}$$

First, using the Hoeffding's inequality, let us show concentration of $\frac{1}{n}\sum_{i\in[n]} \big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)}\big)^2$ around its mean. We observe that each term in the summation is an independent random variable and is bounded as follows

$$\begin{aligned}
\big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)}\big)^2 \overset{(a)}{=} \big(\omega_t^{(i)} x_t^{(i)} + 2\Omega_t^\top \boldsymbol{x}^{(i)} x_t^{(i)}\big)^2 &\overset{(b)}{\leq} \big(\omega_t^{(i)} + 2\Omega_t^\top \boldsymbol{x}^{(i)}\big)^2 x_{\max}^2 \\
&\overset{(c)}{\leq} \big(|\omega_t^{(i)}| + 2\|\Omega_t\|_1 \|\boldsymbol{x}^{(i)}\|_\infty\big)^2 x_{\max}^2 \\
&\overset{(d)}{\leq} \big(2\alpha + 8\alpha\beta x_{\max}\big)^2 x_{\max}^2 \overset{(29)}{=} 4C_{1,\tau}^2 x_{\max}^2,
\end{aligned}$$

where (a) follows by plugging in $\Delta_t^{(i)}$ and $\widetilde{\boldsymbol{x}}^{(i)}$, (b) follows because $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$, (c) follows by triangle inequality and Cauchy–Schwarz inequality, and (d) follows because $\Omega \in 2\Lambda_\Theta$, $\omega^{(i)} \in 2\Lambda_\theta$, and $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$. Then, from the Hoeffding's inequality, we have

$$\mathbb{P}\bigg(\bigg| \frac{1}{n} \sum_{i \in [n]} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)} \Big)^2 - \frac{1}{n} \sum_{i \in [n]} \mathbb{E}\bigg[\Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)} \Big)^2 \bigg] \bigg| > \varepsilon \bigg) < 2\exp\bigg(-\frac{n\varepsilon^2}{32C_{1,\tau}^4 x_{\max}^4} \bigg).$$

Applying the union bound over all $t \in [p]$, for any $\delta \in (0,1)$ and uniformly for all $t \in [p]$, we have

$$\frac{1}{n} \sum_{i \in [n]} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)} \Big)^2 \geq \frac{1}{n} \sum_{i \in [n]} \mathbb{E}\bigg[\Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)} \Big)^2 \bigg] - \varepsilon \tag{42}$$

with probability at least $1 - \delta$ as long as

$$n \geq \frac{32C_{1,\tau}^4 x_{\max}^4}{\varepsilon^2} \log\bigg(\frac{2p}{\delta}\bigg).$$

Now, we lower bound $\mathbb{E}\Big[\big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)}\big)^2\Big]$ for every $t \in [p]$ and every $i \in [n]$. Fix any $t \in [p]$ and $i \in [n]$. We have

$$\mathbb{E}\bigg[\Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)} \Big)^2 \bigg] \overset{(a)}{\geq} \mathbb{V}\mathrm{ar}\Big[[\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)} \Big] = \mathbb{V}\mathrm{ar}\Big[\omega_t^{(i)} x_t^{(i)} + 2\Omega_t^\top \boldsymbol{x}^{(i)} x_t^{(i)} \Big], \tag{43}$$

where (a) follows from the fact that for any random variable a, $\mathbb{E}[a^2] \geq \mathbb{V}\mathrm{ar}[a]$. We define the following set to lower bound $\mathbb{V}\mathrm{ar}\big[\omega_t^{(i)} x_t^{(i)} + 2\Omega_t^\top \boldsymbol{x}^{(i)} x_t^{(i)}\big]$:

$$\mathcal{E}(\Theta^\star) \triangleq \left\{(t, u) \in [p]^2 : t < u, \Theta^\star_{tu} \neq 0\right\}, \tag{44}$$

and consider the graph $\mathcal{G}(\Theta^\star) = ([p], \mathcal{E}(\Theta^\star))$ with $[p]$ as nodes and $\mathcal{E}(\Theta^\star)$ as edges such that $f_{\mathsf{x}|\mathsf{z}}\big(\boldsymbol{x}|\boldsymbol{z}; \theta^\star(\boldsymbol{z}), \Theta^\star\big)$ is Markov with respect to $\mathcal{G}(\Theta^\star)$. We claim that there exists a non-empty set $\mathcal{R}_t \subset [p] \setminus \{t\}$ such that

(i) $\mathcal{R}_t$ is an independent set of $\mathcal{G}(\Theta^\star)$, i.e., there are no edges between any pair of nodes in $\mathcal{R}_t$, and

(ii) the row vector Ω_t satisfies $\sum_{u \in \mathcal{R}_t} |\Omega_{tu}|^2 \geq \frac{1}{\beta+1} \|\Omega_t\|_2^2$.

Taking this claim as given at the moment, we continue our proof. Denoting $\mathcal{R}_t^c \triangleq [p] \setminus \mathcal{R}_t$, and using the law of total variance, the variance term in (43) can be lower bounded as

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\bigg[\omega_t^{(i)} x_t^{(i)} + 2\Omega_t^\top \boldsymbol{x}^{(i)} x_t^{(i)}\bigg] &\geq \mathbb{E}\bigg[\mathbb{V}\mathrm{ar}\bigg[\omega_t^{(i)} x_t^{(i)} + 2\Omega_t^\top \boldsymbol{x}^{(i)} x_t^{(i)} \bigg| \boldsymbol{x}^{(i)}_{\mathcal{R}_t^c}, \boldsymbol{z}^{(i)}\bigg]\bigg] \\
&\stackrel{(a)}{=} 4\mathbb{E}\bigg[\big(x_t^{(i)}\big)^2 \mathbb{V}\mathrm{ar}\Big(\sum_{u \in \mathcal{R}_t} \Omega_{tu} x_u^{(i)} \Big| \boldsymbol{x}^{(i)}_{\mathcal{R}_t^c}, \boldsymbol{z}^{(i)}\Big)\bigg] \\
&\stackrel{(b)}{=} 4\mathbb{E}\bigg[\big(x_t^{(i)}\big)^2 \sum_{u \in \mathcal{R}_t} \Omega_{tu}^2 \mathbb{V}\mathrm{ar}\Big(x_u^{(i)} \Big| \boldsymbol{x}^{(i)}_{\mathcal{R}_t^c}, \boldsymbol{z}^{(i)}\Big)\bigg] \\
&\stackrel{(c)}{=} 4\mathbb{E}\bigg[\big(x_t^{(i)}\big)^2 \sum_{u \in \mathcal{R}_t} \Omega_{tu}^2 \mathbb{V}\mathrm{ar}\Big(x_u^{(i)} \Big| \boldsymbol{x}^{(i)}_{-u}, \boldsymbol{z}^{(i)}\Big)\bigg] \\
&\stackrel{(d)}{\geq} \frac{4x_{\max}}{\pi e C_{2,\tau}^2} \sum_{u \in \mathcal{R}_t} \Omega_{tu}^2 \mathbb{E}\Big[\big(x_t^{(i)}\big)^2\Big] \\
&\stackrel{(e)}{\geq} \frac{4x_{\max}}{\pi e C_{2,\tau}^2} \sum_{u \in \mathcal{R}_t} \Omega_{tu}^2 \mathbb{V}\mathrm{ar}\Big(x_t^{(i)} \Big| \boldsymbol{x}^{(i)}_{-t}, \boldsymbol{z}^{(i)}\Big) \\
&\stackrel{(f)}{\geq} \frac{4x_{\max}^2}{\pi^2 e^2 C_{2,\tau}^4} \sum_{u \in \mathcal{R}_t} \Omega_{tu}^2 \stackrel{(ii)}{\geq} \frac{4x_{\max}^2 \|\Omega_t\|_2^2}{\pi^2 e^2 (\beta+1) C_{2,\tau}^4},
\end{aligned} \tag{45}$$

where (a) follows because $(x_u^{(i)})_{u \in \mathcal{R}_t^c}$ are deterministic when conditioned on themselves, and $t \in \mathcal{R}_t^c$, (b) follows because $(x_u^{(i)})_{u \in \mathcal{R}_t}$ are conditionally independent given $\boldsymbol{x}^{(i)}_{\mathcal{R}_t^c}$ and $\boldsymbol{z}^{(i)}$ which is a direct consequence of (i), (c) follows because of the local Markov property (as the conditioning set includes all the neighbors in $\mathcal{G}(\Theta^\star)$ of each node in $\mathcal{R}_t$), (d) and (f) follow from Lem. 7, and (e) follows because $\mathbb{E}\Big[\big(x_t^{(i)}\big)^2\Big] = \mathbb{E}\bigg[\mathbb{E}\Big[\big(x_t^{(i)}\big)^2 \Big| \boldsymbol{x}^{(i)}_{-t}, \boldsymbol{z}^{(i)}\Big]\bigg] \geq \mathbb{V}\mathrm{ar}\Big(x_t^{(i)} \Big| \boldsymbol{x}^{(i)}_{-t}, \boldsymbol{z}^{(i)}\Big)$.

Combining (41) to (43) and (45), for any $\delta \in (0, 1)$ and uniformly for all $t \in [p]$, we have

$$\partial^2_{\underline{\Omega}_t^2} \mathcal{L}_t(\underline{\Theta}_t) \geq \frac{1}{C_{2,\tau}} \left(\frac{4x_{\max}^2 \|\Omega_t\|_2^2}{\pi^2 e^2 (\beta+1) C_{2,\tau}^4} - \varepsilon\right)$$

with probability at least $1 - \delta$ as long as

$$n \geq \frac{32 C_{1,\tau}^4 x_{\max}^4}{\varepsilon^2} \log\left(\frac{2p}{\delta}\right).$$

Choosing $\varepsilon_2 = \varepsilon / C_{2,\tau}$ and $\delta_2 = \delta$ yields the claim.

It remains to construct the set $\mathcal{R}_t$ that is an independent set of $\mathcal{G}(\Theta^\star)$ and satisifies (ii).

Construction of the set $\mathcal{R}_t$: We select $r_1 \in [p] \setminus \{t\}$ such that

$$|\Omega_{tr_1}| \geq |\Omega_{tu}| \quad \text{for all} \quad u \in [p] \setminus \{t, r_1\}\,.$$

Next, we identify $r_2 \in [p] \setminus \{t, r_1, \mathcal{N}(r_1)\}$ such that

$$|\Omega_{tr_2}| \geq |\Omega_{tu}| \quad \text{for all} \quad u \in [p] \setminus \{t, r_1, \mathcal{N}(r_1), r_2\}\,.$$

We continue identifying $r_3, \ldots, r_s$ in such a manner till no more nodes are left, where s denotes the total number of nodes selected. Now we define $\mathcal{R}_t \triangleq \{r_1, \cdots, r_s\}$. Further, for any $u \in [p]$, let $\mathcal{N}(u)$ denote the set of neighbors of u in $\mathcal{G}(\Theta^\star)$. We have $|\mathcal{N}(u)| \leq \|\Theta_u^\star\|_0 \leq \beta$ from (44) and Assum. 1(b). We can now see that $\mathcal{R}_t$ is an independent set of $\mathcal{G}(\Theta^\star)$ as claimed in (i) such that it satisfies (ii) by construction.

Proof of Lem. 6: Lower bound on the second directional derivative: For every $t \in [p]$ with $\underline{\Omega}_t$ defined in (36), we claim that the second-order directional derivative of the loss function defined in (27) is given by

$$\partial^2_{\underline{\Omega}_t^2} \mathcal{L}_t(\underline{\Theta}_t) = \frac{1}{n} \sum_{i \in [n]} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)} \Big)^2 \exp\Big(-[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}] x_t^{(i)} \Big), \tag{46}$$

where $\Delta_t^{(i)} \triangleq \begin{bmatrix} \omega_t^{(i)} \\ 2\Omega_t \end{bmatrix} \in \mathbb{R}^{p+1}$ and $\widetilde{\boldsymbol{x}}^{(i)} \triangleq \begin{bmatrix} 1 \\ \boldsymbol{x}^{(i)} \end{bmatrix} \in \mathbb{R}^{p+1}$ for all $i \in [n]$. We provide a proof at the end.

Given this, we proceed to prove the lower bound on the second directional derivative. Fix any $t \in [p]$. From (46), we have

$$\begin{aligned}
\partial^2_{\underline{\Omega}_t^2} \mathcal{L}_t(\underline{\Theta}_t) &= \frac{1}{n} \sum_{i \in [n]} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)} \Big)^2 \times \exp\Big(-[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}] x_t^{(i)} \Big) \\
&\overset{(a)}{\geq} \frac{1}{n} \sum_{i \in [n]} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)} \Big)^2 \times \exp\Big(-|\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}| x_{\max} \Big) \\
&\overset{(b)}{\geq} \frac{1}{n} \sum_{i \in [n]} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)} \Big)^2 \times \exp\Big(-\big(|\theta_t^{(i)}| + 2\|\Theta_t\|_1 \|\boldsymbol{x}^{(i)}\|_\infty\big) x_{\max} \Big) \\
&\overset{(c)}{\geq} \frac{1}{n} \sum_{i \in [n]} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)} \Big)^2 \times \exp\Big(-(\alpha + 2\alpha\beta x_{\max}) x_{\max} \Big) \\
&\overset{(29)}{=} \frac{1}{C_{2,\tau} n} \sum_{i \in [n]} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)} \Big)^2,
\end{aligned}$$

where (a) follows because $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$, (b) follows by triangle inequality and Cauchy–Schwarz inequality, and (c) follows because $\theta^{(i)} \in \Lambda_\theta$ for all $i \in [n]$, $\Theta \in \Lambda_\Theta$, and $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$.

Proof of (46): Expression for second directional derivative: Fix any $t \in [p]$. The second-order partial derivatives of $\mathcal{L}_t$ with respect to entries of $\underline{\Theta}_t$ defined in (9) are given by

$$\frac{\partial^2 \mathcal{L}_t(\underline{\Theta}_t)}{\partial [\theta_t^{(i)}]^2} = \frac{1}{n} [x_t^{(i)}]^2 \exp\Big(-[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}] x_t^{(i)} \Big) \text{ for all } i \in [n],$$

$$\frac{\partial^2 \mathcal{L}_t(\underline{\Theta}_t)}{\partial \Theta_{tu} \Theta_{tv}} = \frac{4}{n} \sum_{i \in [n]} [x_t^{(i)}]^2 x_u^{(i)} x_v^{(i)} \exp\Big(-[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}] x_t^{(i)} \Big) \text{ for all } u, v \in [p] \setminus \{i\}\,, \text{ and}$$

$$\frac{\partial^2 \mathcal{L}_t(\underline{\Theta}_t)}{\partial \Theta_{tu} \theta_t^{(i)}} = \frac{\partial^2 \mathcal{L}_t(\underline{\Theta}_t)}{\partial \theta_t^{(i)} \Theta_{tu}} = \frac{2}{n} [x_t^{(i)}]^2 x_u^{(i)} \exp\Big(-[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}] x_t^{(i)} \Big) \text{ for all } i \in [n], u \in [p] \setminus \{t\}\,.$$

Now, we can write the second-order directional derivative of $\mathcal{L}_t$ as

$$\partial^2_{\underline{\Omega}_t^2} \mathcal{L}_t(\underline{\Theta}_t) \triangleq \lim_{h \to 0} \frac{\partial_{\underline{\Omega}_t} \mathcal{L}_t(\underline{\Theta}_t + h\underline{\Omega}_t) - \partial_{\underline{\Omega}_t} \mathcal{L}_t(\underline{\Theta}_t)}{h}$$

$$
\begin{aligned}
&= \sum_{i\in[n]} \big[\omega_t^{(i)}\big]^2 \frac{\partial^2 \mathcal{L}_t(\underline{\Theta}_t)}{\partial\big[\theta_t^{(i)}\big]^2} + \sum_{u\in[p]}\sum_{v\in[p]} \Omega_{tu}\Omega_{tv} \frac{\partial^2 \mathcal{L}_t(\underline{\Theta}_t)}{\partial\Theta_{tu}\Theta_{tv}} + 2\sum_{i\in[n]}\sum_{u\in[p]} \omega_t^{(i)}\Omega_{tu} \frac{\partial^2 \mathcal{L}_t(\underline{\Theta}_t)}{\partial\Theta_{tu}\theta_t^{(i)}} \\
&= \frac{1}{n}\sum_{i\in[n]} \Big(\big[\omega_t^{(i)} x_t^{(i)}\big]^2 + 4\sum_{u\in[p]} \Omega_{tu} x_t^{(i)} x_u^{(i)} \sum_{v\in[p]} \Omega_{tv} x_t^{(i)} x_v^{(i)} + 4\omega_t^{(i)} x_t^{(i)} \sum_{u\in[p]} \Omega_{tu} x_t^{(i)} x_u^{(i)} \Big) \\
&\qquad\qquad \times \exp\Big(-[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}] x_t^{(i)} \Big) \\
&= \frac{1}{n}\sum_{i\in[n]} \Big(\omega_t^{(i)} x_t^{(i)} + 2\Omega_t^\top \boldsymbol{x}^{(i)} x_t^{(i)} \Big)^2 \exp\Big(-[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}] x_t^{(i)} \Big) \\
&\overset{(a)}{=} \frac{1}{n}\sum_{i\in[n]} \Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)} \Big)^2 \exp\Big(-[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}] x_t^{(i)} \Big),
\end{aligned}
$$

where (a) follows from the definitions of $\Delta_t^{(i)}$ and $\widetilde{\boldsymbol{x}}^{(i)}$.

Proof of Lem. 7: Lower bound on the conditional variance: For any random variable x, let $h(\mathsf{x})$ denote the differential entropy of x. Fix any $t \in [p]$ and $i \in [n]$. Then, we have

$$
\begin{aligned}
&2\pi e \mathbb{V}\mathrm{ar}\big(x_t^{(i)} \big| \boldsymbol{x}_{-t}^{(i)}, \boldsymbol{z}^{(i)}\big) \overset{(a)}{\geq} \exp\Big(h\big[x_t^{(i)} \big| \boldsymbol{x}_{-t}^{(i)}, \boldsymbol{z}^{(i)}\big] \Big) \\
&= \exp\Big(-\int_{\mathcal{X}^p \times \mathcal{Z}^{p_z}} f_{\mathsf{x},\mathsf{z}}(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}) \log\Big(f_{\mathsf{x}_t|\mathsf{x}_{-t},\mathsf{z}}\big(x_t^{(i)} \big| \boldsymbol{x}_{-t}^{(i)}, \boldsymbol{z}^{(i)}; \theta_t^\star(\boldsymbol{z}^{(i)}), \Theta_t^\star\big) \Big) d\boldsymbol{x}^{(i)} d\boldsymbol{z}^{(i)} \Big) \\
&= \exp\left(-\int_{\mathcal{X}^p \times \mathcal{Z}^{p_z}} f_{\mathsf{x},\mathsf{z}}(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}) \log\left(\frac{\exp\Big([\theta_t^\star(\boldsymbol{z}^{(i)}) + 2\Theta_t^{\star\top} \boldsymbol{x}^{(i)}] x_t^{(i)} \Big)}{\int_{\mathcal{X}} \exp\Big([\theta_t^\star(\boldsymbol{z}^{(i)}) + 2\Theta_t^{\star\top} \boldsymbol{x}^{(i)}] x_t^{(i)} \Big) dx_t^{(i)}} \right) d\boldsymbol{x}^{(i)} d\boldsymbol{z}^{(i)} \right) \\
&\overset{(b)}{\geq} \exp\left(-\int_{\mathcal{X}^p \times \mathcal{Z}^{p_z}} f_{\mathsf{x},\mathsf{z}}(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}) \log\left(\frac{\exp\Big(|\theta_t^\star(\boldsymbol{z}^{(i)}) + 2\Theta_t^{\star\top} \boldsymbol{x}^{(i)}| x_{\max} \Big)}{\int_{\mathcal{X}} \exp\Big(-|\theta_t^\star(\boldsymbol{z}^{(i)}) + 2\Theta_t^{\star\top} \boldsymbol{x}^{(i)}| x_{\max} \Big) dx_t^{(i)}} \right) d\boldsymbol{x}^{(i)} d\boldsymbol{z}^{(i)} \right) \\
&\overset{(c)}{\geq} \exp\left(-\int_{\mathcal{X}^p \times \mathcal{Z}^{p_z}} f_{\mathsf{x},\mathsf{z}}(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}) \log\left(\frac{\exp\Big(\big(|\theta_t^\star(\boldsymbol{z}^{(i)})| + 2\|\Theta_t^\star\|_1 \|\boldsymbol{x}^{(i)}\|_\infty\big) x_{\max} \Big)}{\int_{\mathcal{X}} \exp\Big(-\big(|\theta_t^\star(\boldsymbol{z}^{(i)})| + 2\|\Theta_t^\star\|_1 \|\boldsymbol{x}^{(i)}\|_\infty\big) x_{\max} \Big) dx_t^{(i)}} \right) d\boldsymbol{x}^{(i)} d\boldsymbol{z}^{(i)} \right) \\
&\overset{(d)}{\geq} \exp\left(-\int_{\mathcal{X}^p \times \mathcal{Z}^{p_z}} f_{\mathsf{x},\mathsf{z}}(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}) \log\left(\frac{\exp\Big((\alpha + 2\alpha\beta x_{\max}) x_{\max} \Big)}{\int_{\mathcal{X}} \exp\Big(-(\alpha + 2\alpha\beta x_{\max}) x_{\max} \Big) dx_t^{(i)}} \right) d\boldsymbol{x}^{(i)} d\boldsymbol{z}^{(i)} \right) \\
&\overset{(e)}{=} \exp\left(-\int_{\mathcal{X}^p \times \mathcal{Z}^{p_z}} f_{\mathsf{x},\mathsf{z}}(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}) \log\left(\frac{C_{3,\tau}^2}{2x_{\max}} \right) d\boldsymbol{x}^{(i)} d\boldsymbol{z}^{(i)} \right) = \frac{2x_{\max}}{C_{3,\tau}^2},
\end{aligned}
$$

where (a) follows from Shannon's entropy inequality $(h(\cdot) \leq \log \sqrt{2\pi \mathbb{V}\mathrm{ar}(\cdot)})$, (b) follows because $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$, (c) follows by triangle inequality and Cauchy–Schwarz inequality, and (d) follows because $\theta^\star(\boldsymbol{z}^{(i)}) \in \Lambda_\theta$ for all $i \in [n]$, $\Theta^\star \in \Lambda_\Theta$, $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$, and (e) follows because $\int_{\mathcal{X}} dx_t^{(i)} = 2x_{\max}$.

C.2 Proof of Lem. 3: Lipschitzness of the loss function

Fix any $t \in [p]$. Consider the direction $\underline{\Omega}_t = \widetilde{\underline{\Theta}}_t - \underline{\Theta}_t$ and note that $\omega_t^{(i)} = 0$ for all $i \in [n]$. Now, define the function $q : [0, 1] \to \mathbb{R}$ as follows:

$$
q(a) = \mathcal{L}_t\big(\underline{\Theta}_t + a(\widetilde{\underline{\Theta}}_t - \underline{\Theta}_t)\big). \tag{47}
$$

Then, the desired inequality in (30) is equivalent to $|q(1) - q(0)| \leq x_{\max}^2 C_{2,\tau} \|\Omega_t\|_1$. From the mean value theorem, there exists $a' \in (0, 1)$ such that

$$
|q(1) - q(0)| = \left| \frac{dq(a')}{da} \right|. \tag{48}
$$

Therefore, we have

$$
\begin{aligned}
&\big|q(1)-q(0)\big| \\
&\overset{(48)}{=} \left|\frac{dq(a')}{da}\right| \\
&\overset{(47)}{=} \left|\frac{d\mathcal{L}_t\big(\underline{\Theta}_t + a'(\widetilde{\underline{\Theta}}_t - \underline{\Theta}_t)\big)}{da}\right| \\
&\overset{(37)}{=} \left|\partial_{\underline{\Omega}_t}\mathcal{L}_t(\underline{\Theta}_t)\big|_{\underline{\Theta}_t = \underline{\Theta}_t + a'(\widetilde{\underline{\Theta}}_t - \underline{\Theta}_t)}\right| \\
&\overset{(40)}{=} \frac{1}{n}\left|\sum_{i\in[n]}\Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)}\Big)\exp\Big(-\Big[\big(\theta_t^{(i)}+a'(\widetilde{\theta}_t^{(i)}-\theta_t^{(i)})\big)+2\big(\Theta_t+a'(\widetilde{\Theta}_t-\Theta_t)\big)^\top \boldsymbol{x}^{(i)}\Big]x_t^{(i)}\Big)\right| \\
&\overset{(a)}{\leq} \exp\Big(\big([(1-a')\alpha+a'\alpha]+2[(1-a')\alpha\beta+a'\alpha\beta]x_{\max}\big)x_{\max}\Big)\frac{1}{n}\left|\sum_{i\in[n]}\Big([\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)} x_t^{(i)}\Big)\right| \\
&\overset{(b)}{\leq} \frac{x_{\max}C_{2,\tau}}{n}\sum_{i\in[n]}\left|[\Delta_t^{(i)}]^\top \widetilde{\boldsymbol{x}}^{(i)}\right| \overset{(c)}{\leq} x_{\max}C_{2,\tau}\left|\Omega_t^\top \boldsymbol{x}^{(i)}\right| \overset{(d)}{\leq} x_{\max}^2 C_{2,\tau}\|\Omega_t\|_1,
\end{aligned}
$$

where (a) follows from triangle inequality, Cauchy–Schwarz inequality, $\theta^{(i)}, \widetilde{\theta}^{(i)} \in \Lambda_\theta$, $\Theta, \widetilde{\Theta} \in \Lambda_\Theta$, and $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$, (b) follows from (29), the triangle inequality, and because $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$, (c) follows from the definitions of $\Delta_t^{(i)}$ and $\widetilde{\boldsymbol{x}}^{(i)}$ because $\omega_t^{(i)} = 0$ for all $i \in [n]$, and (d) follows from the triangle inequality, and because $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$.

D Proof of Thm. 1 Part II: Recovering unit-level parameters

Throughout the proof, c denotes either a universal constant or a constant depending on the model-parameters α and $x_{\max}$, and may change appearance line by line.

To analyze our estimate of the unit-level parameters, we use the estimate $\widehat{\Theta}$ of the population-level parameter $\Theta^\star$ along with the associated guarantee provided in Thm. 1. We note that the constraints on the unit-level parameters in (10) are independent across units, i.e., $\theta^{(i)} \in \Lambda_\theta$ independently for all $i \in [n]$. Therefore, we look at n independent convex optimization problems by decomposing the loss function $\mathcal{L}$ in (9) and the estimate $\widehat{\underline{\Theta}}$ in (10) as follows:

$$\mathcal{L}^{(i)}\big(\theta^{(i)}\big) \triangleq \sum_{t\in[p]}\exp\Big(-[\theta_t^{(i)} + \widehat{\Theta}_t^\top \boldsymbol{x}^{(i)}]x_t^{(i)}\Big) \quad\text{and}\quad \widehat{\theta}^{(i)} = \operatorname*{arg\,min}_{\theta^{(i)}\in\Lambda_\theta}\mathcal{L}^{(i)}\big(\theta^{(i)}\big) \quad \text{for all} \quad i\in[n], \tag{49}$$

Now, fix any $i \in [n]$. From (49), we have $\mathcal{L}^{(i)}\big(\widehat{\theta}^{(i)}\big) \leq \mathcal{L}^{(i)}\big(\theta^{\star(i)}\big)$. Using contraposition, to prove this part, it is sufficient to show that all points $\theta^{(i)} \in \Lambda_\theta$ that satisfy $\|\theta^{(i)} - \theta^{\star(i)}\|_2 \geq R(\varepsilon,\delta)$ also uniformly satisfy

$$\mathcal{L}^{(i)}\big(\theta^{(i)}\big) \geq \mathcal{L}^{(i)}\big(\theta^{\star(i)}\big) + R^2(\varepsilon,\delta) \text{ for } n \geq \frac{c\exp(c\beta)}{\varepsilon^4}\cdot\bigg(\log\frac{p}{\delta} + \mathcal{M}_\theta(\eta)\bigg), \tag{50}$$

with probability at least $1-\delta$ where $R(\varepsilon,\delta)$ and η were defined in (11), and the metric entropy $\mathcal{M}$ was defined in Def. 2. Then, the guarantee in Thm. 1 follows by applying a union bound over all $i \in [n]$.

To that end, first, we claim that for any fixed $\theta^{(i)} \in \Lambda_\theta$, if $\theta^{(i)}$ is far from $\theta^{\star(i)}$, then with high probability $\mathcal{L}^{(i)}\big(\theta^{(i)}\big)$ will be significantly larger than $\mathcal{L}^{(i)}\big(\theta^{\star(i)}\big)$. We provide a proof in App. D.1.

Lemma 8 (**Gap between the loss function for a fixed parameter**). *Fix any $\varepsilon_1 > 0$, $\delta_1 \in (0,1)$, and $i \in [n]$. Then, for any $\theta^{(i)} \in \Lambda_\theta$ such that $\|\theta^{(i)} - \theta^{\star(i)}\|_2 \geq \varepsilon_1\gamma$ (see (11)), we have*

$$\mathcal{L}^{(i)}\big(\theta^{(i)}\big) \geq \mathcal{L}^{(i)}\big(\theta^{\star(i)}\big) + \frac{2\sqrt{2}\alpha\beta x_{\max}^3}{\pi e C_{2,\tau}^4}\|\theta^{(i)} - \theta^{\star(i)}\|_2^2 \quad \textit{for} \quad n \geq \frac{c\exp(c\beta)\log(p/\delta_1)}{\varepsilon_1^4},$$

with probability at least $1-\delta_1-c\beta^2\log p\cdot\exp(-\exp(-c\beta)\|\theta^{(i)}-\theta^{\star(i)}\|_2^2)$ *where* $C_{2,\tau}$ *was defined in* (29).

Next, we claim that the loss function $\mathcal{L}^{(i)}$ is Lipschitz and capture this property via the following lemma. We provide a proof in App. D.2.

Lemma 9 (**Lipschitzness of the loss function**). *Consider any* $i\in[n]$. *Then, the loss function* $\mathcal{L}^{(i)}$ *is Lipschitz with respect to the* ℓ_1 *norm* $\|\cdot\|_1$ *and with Lipschitz constant* $x_{\max}C_{2,\tau}$, *i.e.,*

$$\left|\mathcal{L}^{(i)}\big(\widetilde{\theta}^{(i)}\big)-\mathcal{L}^{(i)}\big(\theta^{(i)}\big)\right|\leq x_{\max}C_{2,\tau}\|\widetilde{\theta}^{(i)}-\theta^{(i)}\|_1\quad\textit{for all}\quad\theta^{(i)},\widetilde{\theta}^{(i)}\in\Lambda_\theta,\tag{51}$$

where the constant $C_{2,\tau}$ *was defined in* (29).

Given these lemmas, we now proceed with the proof.

Proof strategy: As mentioned earlier, the idea is to show that all points $\theta^{(i)}\in\Lambda_\theta$ that satisfy $\|\theta^{(i)}-\theta^{\star(i)}\|_2\geq R(\varepsilon,\delta)$ also uniformly satisfy (50) with probability at least $1-\delta$. To do so, we consider the set of points $\Lambda_\theta^{r,i}\subset\Lambda_\theta$ whose distance from $\theta^{\star(i)}$ is at least $r>0$ in ℓ_2 norm. Then, using an appropriate covering set of $\Lambda_\theta^{r,i}$ and the Lipschitzness of $\mathcal{L}^{(i)}$, we show that the value of $\mathcal{L}^{(i)}$ at all points in $\Lambda_\theta^{r,i}$ is uniformly $O(r^2)$ larger than the value of $\mathcal{L}^{(i)}$ at $\theta^{\star(i)}$ with high probability. Finally, we choose an r small enough to make the failure probability smaller than δ.

Gap between the loss function for all parameters in the covering set: Consider any $r\geq\varepsilon\gamma$ (where γ is defined in (11)) and the set of elements $\Lambda_\theta^{r,i}\triangleq\big\{\theta^{(i)}\in\Lambda_\theta:\|\theta^{\star(i)}-\theta^{(i)}\|_2\geq r\big\}$. Let $\mathcal{C}(\Lambda_\theta^{r,i},\varepsilon')$ be the ε'-covering number of the set $\Lambda_\theta^{r,i}$ and let $\mathcal{U}(\Lambda_\theta^{r,i},\varepsilon')$ be the associated ε'-cover (see Def. 2) where

$$\varepsilon'\triangleq\frac{\sqrt{2}\alpha\beta x_{\max}^2r^2}{\pi eC_{2,\tau}^5\gamma}.\tag{52}$$

Now, we apply Lem. 8 to each element in $\mathcal{U}(\Lambda_\theta^{r,i},\varepsilon')$ and argue by a union bound that the value of $\mathcal{L}^{(i)}$ at all points in $\mathcal{U}(\Lambda_\theta^{r,i},\varepsilon')$ is uniformly $O(r^2)$ larger than the value of $\mathcal{L}^{(i)}$ at $\theta^{\star(i)}$ with high probability. We start by considering any $\theta^{(i)}\in\mathcal{U}(\Lambda_\theta^{r,i},\varepsilon')$. We have

$$\|\theta^{\star(i)}-\theta^{(i)}\|_2\overset{(a)}{\geq}r,\tag{53}$$

where (a) follows because $\mathcal{U}(\Lambda_\theta^{r,i},\varepsilon')\subseteq\Lambda_\theta^{r,i}$. Now, applying Lem. 8 with $\varepsilon_1=\varepsilon$ and $\delta_1=\delta/2\mathcal{C}(\Lambda_\theta^{r,i},\varepsilon')$, we have

$$\mathcal{L}^{(i)}\big(\theta^{(i)}\big)\geq\mathcal{L}^{(i)}\big(\theta^{\star(i)}\big)+\frac{2\sqrt{2}\alpha\beta x_{\max}^3}{\pi eC_{2,\tau}^4}\|\theta^{\star(i)}-\theta^{(i)}\|_2^2\overset{(53)}{\geq}\mathcal{L}^{(i)}\big(\theta^{\star(i)}\big)+\frac{2\sqrt{2}\alpha\beta x_{\max}^3r^2}{\pi eC_{2,\tau}^4},$$

with probability at least $1-\delta/2\mathcal{C}(\Lambda_\theta^{r,i},\varepsilon')-c\beta^2\log p\cdot\exp(-\exp(-c\beta)\|\theta^{(i)}-\theta^{\star(i)}\|_2^2)$ whenever

$$n\geq\frac{c\exp(c\beta)\log\big(2\mathcal{C}(\Lambda_\theta^{r,i},\varepsilon')\cdot p/\delta\big)}{\varepsilon^4}.\tag{54}$$

By applying the union bound over $\mathcal{U}(\Lambda_\theta^{r,i},\varepsilon')$, as long as n satisfies (54), we have

$$\mathcal{L}^{(i)}\big(\theta^{(i)}\big)\geq\mathcal{L}^{(i)}\big(\theta^{\star(i)}\big)+\frac{2\sqrt{2}\alpha\beta x_{\max}^3r^2}{\pi eC_{2,\tau}^4}\text{ uniformly for every }\theta^{(i)}\in\mathcal{U}(\Lambda_\theta^{r,i},\varepsilon'),\tag{55}$$

with probability at least $1-\delta/2-c\beta^2\mathcal{C}(\Lambda_\theta^{r,i},\varepsilon')\log p\cdot\exp(-\exp(-c\beta)\|\theta^{(i)}-\theta^{\star(i)}\|_2^2)$ which can lower bounded by $1-\delta/2-c\beta^2\mathcal{C}(\Lambda_\theta^{r,i},\varepsilon')\log p\cdot\exp(-\exp(-c\beta)r^2)$ using (53).

Generalize beyond the covering set: Next, we assume that (55) holds, and generalize from $\mathcal{U}(\Lambda_\theta^{r,i}, \varepsilon')$ to all of $\Lambda_\theta^{r,i}$. Consider any $\widetilde{\theta}^{(i)} \in \Lambda_\theta^{r,i}$ and let $\theta^{(i)}$ be any point in $\mathcal{U}(\Lambda_\theta^{r,i}, \varepsilon')$ that satisfies $\|\theta^{(i)} - \widetilde{\theta}^{(i)}\|_2 \le \varepsilon'$ (see Def. 2). Then, from Lem. 9, we have

$$\begin{aligned}
\mathcal{L}^{(i)}\big(\widetilde{\theta}^{(i)}\big) \ge \mathcal{L}^{(i)}\big(\theta^{(i)}\big) - x_{\max} C_{2,\tau} \|\theta^{(i)} - \widetilde{\theta}^{(i)}\|_1 &\overset{(11)}{\ge} \mathcal{L}^{(i)}\big(\theta^{(i)}\big) - x_{\max} C_{2,\tau}\gamma \|\theta^{(i)} - \widetilde{\theta}^{(i)}\|_2 \\
&\overset{(a)}{\ge} \mathcal{L}^{(i)}\big(\theta^{(i)}\big) - x_{\max} C_{2,\tau}\gamma\varepsilon' \\
&\overset{(52)}{\ge} \mathcal{L}^{(i)}\big(\theta^{(i)}\big) - \frac{\sqrt{2}\alpha\beta x_{\max}^3 r^2}{\pi e C_{2,\tau}^4} \\
&\overset{(55)}{\ge} \mathcal{L}^{(i)}\big(\theta^{\star(i)}\big) + \frac{\sqrt{2}\alpha\beta x_{\max}^3 r^2}{\pi e C_{2,\tau}^4},
\end{aligned}$$

where (a) follows because $\|\theta^{(i)} - \widetilde{\theta}^{(i)}\|_2 \le \varepsilon'$. Recall that $\widetilde{\theta}^{(i)}$ is any point in $\Lambda_\theta^{r,i}$, i.e., $\|\theta^{\star(i)} - \widetilde{\theta}^{(i)}\|_2 \ge r$. Therefore, we have an inequality that looks like (50). It remains to bound n and the failure probability.

Bounding n: To bound n in (54), we bound the covering number $\mathcal{C}(\Lambda_\theta^{r,i}, \varepsilon')$ as follows

$$\mathcal{C}(\Lambda_\theta^{r,i}, \varepsilon') \overset{(a)}{\le} \mathcal{C}(\Lambda_\theta, \varepsilon'/2) \tag{56}$$

where (a) follows from the fact that for any sets $\mathcal{U} \subseteq \mathcal{V}$ and any ε, it holds that $\mathcal{C}(\mathcal{U}, \varepsilon) \le \mathcal{C}(\mathcal{V}, \varepsilon/2)$. Then using (56) in (54) and observing that $\varepsilon' = \frac{r^2}{c\exp(c\beta)\gamma}$, it is sufficient for

$$n \ge \frac{c\exp(c\beta)}{\varepsilon^4} \cdot \bigg(\log\frac{p}{\delta} + \mathcal{M}_\theta(r^2\eta)\bigg).$$

Bounding the failure probability: To bound the failure probability by δ, it is sufficient to chose r such that

$$\begin{aligned}
\delta &\ge \delta/2 + c\beta^2 \mathcal{C}(\Lambda_\theta^{r,i}, \varepsilon') \log p \cdot \exp(-\exp(-c\beta) r^2) \\
&\overset{(56)}{\ge} \delta/2 + c\beta^2 \mathcal{C}(\Lambda_\theta, \varepsilon'/2) \log p \cdot \exp(-\exp(-c\beta) r^2).
\end{aligned} \tag{57}$$

Re-arranging and taking logarithm on both sides of (57), and observing that $\varepsilon' = \frac{r^2}{c\exp(c\beta)\gamma}$, we have

$$\log\delta \ge c\bigg[\log\big(\beta^2\log p\big) + \mathcal{M}_\theta(r^2\eta) - \exp(-c\beta) r^2\bigg]. \tag{58}$$

Finally, (58) holds whenever

$$r \ge c\exp\big(c\beta\big)\sqrt{\log\frac{\beta^2\log p}{\delta} + \mathcal{M}_\theta(r^2\eta)}.$$

Recalling that the choice of r was such that $r \ge \varepsilon\gamma$ completes the proof.

D.1 Proof of Lem. 8: Gap between the loss function for a fixed parameter

Fix any $\varepsilon_1 > 0$, any $\delta_1 \in (0,1)$, and any $i \in [n]$. Consider any direction $\omega^{(i)} \in \mathbb{R}^p$ along the parameter $\theta^{(i)}$, i.e.,

$$\omega^{(i)} = \theta^{(i)} - \theta^{\star(i)}. \tag{59}$$

We denote the first-order and the second-order directional derivatives of the loss function $\mathcal{L}^{(i)}$ in (49) along the direction $\omega^{(i)}$ evaluated at $\theta^{(i)}$ by $\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{(i)}))$ and $\partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\theta^{(i)})$ respectively. Below, we state a lemma (with proof divided across App. D.1.1 and App. D.1.2) that provides us a control on $\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{\star(i)}))$ and $\partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\theta^{(i)})$. The assumptions of Lem. 8 remain in force.

Lemma 10 (Control on first and second directional derivatives). *For any fixed $\varepsilon_2, \varepsilon_3 > 0$, $\delta_2 \in (0,1)$, $i \in [n]$, $\theta^{(i)} \in \Lambda_\theta$ with $\omega^{(i)}$ defined in (59), we have the following:*

(a) **Concentration of first directional derivative***:*

$$\left|\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{\star(i)}))\right| \leq \varepsilon_2\|\omega^{(i)}\|_1 + \varepsilon_3\|\omega^{(i)}\|_2^2 \quad \textit{for} \quad n \geq O\bigg(\frac{\exp\big(O(\beta)\big)\log(p/\delta_2)}{\varepsilon_2^4}\bigg),$$

with probability at least $1 - \delta_2 - O\bigg(\beta^2 \log p \exp\bigg(\dfrac{-\varepsilon_3^2\|\omega^{(i)}\|_2^2}{\exp\big(O(\beta)\big)}\bigg)\bigg)$.

(b) **Anti-concentration of second directional derivative***:*

$$\partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\theta^{(i)}) \geq \frac{16\sqrt{2}\alpha\beta x_{\max}^3}{\pi e C_{2,\tau}^3}\|\omega^{(i)}\|_2^2,$$

with probability at least $1 - O\bigg(\beta^2 \log p \exp\bigg(\dfrac{-\|\omega^{(i)}\|_2^2}{\exp\big(O(\beta)\big)}\bigg)\bigg)$ *where* $C_{2,\tau}$ *was defined in* (29).

Given this lemma, we now proceed with the proof. Define a function $g : [0,1] \to \mathbb{R}^p$ as follows:

$$g(a) = \theta^{\star(i)} + a(\theta^{(i)} - \theta^{\star(i)}).$$

Notice that $g(0) = \theta^{\star(i)}$ and $g(1) = \theta^{(i)}$ as well as

$$\frac{d\mathcal{L}^{(i)}(g(a))}{da} = \partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{(i)}))\big|_{\theta^{(i)}=g(a)} \quad \text{and} \quad \frac{d^2\mathcal{L}^{(i)}(g(a))}{da^2} = \partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\theta^{(i)})\big|_{\theta^{(i)}=g(a)}. \tag{60}$$

By the fundamental theorem of calculus, we have

$$\frac{d\mathcal{L}^{(i)}(g(a))}{da} \geq \frac{d\mathcal{L}^{(i)}(g(a))}{da}\Big|_{a=0} + a \min_{a\in(0,1)} \frac{d^2\mathcal{L}^{(i)}(g(a))}{da^2} \tag{61}$$

Integrating both sides of (61) with respect to a, we obtain

$$\begin{aligned}
\mathcal{L}^{(i)}(g(a)) - \mathcal{L}^{(i)}(g(0)) &\geq a\frac{d\mathcal{L}^{(i)}(g(a))}{da}\Big|_{a=0} + \frac{a^2}{2}\min_{a\in(0,1)}\frac{d^2\mathcal{L}^{(i)}(g(a))}{da^2} \\
&\overset{(60)}{=} a\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{(i)}))\big|_{\theta^{(i)}=g(0)} + \frac{a^2}{2}\min_{a\in(0,1)}\partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\theta^{(i)})\big|_{\theta^{(i)}=g(a)} \\
&\overset{(a)}{=} a\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{\star(i)})) + \frac{a^2}{2}\min_{a\in(0,1)}\partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\theta^{(i)})\big|_{\theta^{(i)}=g(a)} \\
&\overset{(b)}{\geq} -a\big|\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{\star(i)}))\big| + \frac{a^2}{2}\min_{a\in(0,1)}\partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\theta^{(i)})\big|_{\theta^{(i)}=g(a)},
\end{aligned} \tag{62}$$

where (a) follows because $g(0) = \theta^{\star(i)}$, and (b) follows by the triangle inequality. Plugging in $a = 1$ in (62) as well as using $g(0) = \theta^{\star(i)}$ and $g(1) = \theta^{(i)}$, we find that

$$\mathcal{L}^{(i)}(\theta^{(i)}) - \mathcal{L}^{(i)}(\theta^{\star(i)}) \geq -\big|\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{\star(i)}))\big| + \frac{1}{2}\min_{a\in(0,1)}\partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\theta^{(i)})\big|_{\theta^{(i)}=g(a)}$$

Now, we use Lem. 10 with $\varepsilon_2 = 2\sqrt{2}\alpha\beta x_{\max}^3\varepsilon_1/\pi e C_{2,\tau}^4$, $\varepsilon_3 = 4\sqrt{2}\alpha\beta x_{\max}^3/\pi e C_{2,\tau}^4$, and $\delta_2 = \delta_1$. Therefore, with probability at least $1 - \delta_1 - O\bigg(\beta^2 \log p \exp\bigg(\dfrac{-\|\omega^{(i)}\|_2^2}{\exp\big(O(\beta)\big)}\bigg)\bigg)$ and as long as $n \geq O\bigg(\dfrac{\exp\big(O(\beta)\big)\log(p/\delta_1)}{\varepsilon_1^4}\bigg)$, we have

$$\begin{aligned}
\mathcal{L}^{(i)}(\theta^{(i)}) - \mathcal{L}^{(i)}(\theta^{\star(i)}) &\geq -\frac{2\sqrt{2}\alpha\beta x_{\max}^3\varepsilon_1}{\pi e C_{2,\tau}^4}\|\omega^{(i)}\|_1 - \frac{4\sqrt{2}\alpha\beta x_{\max}^3}{\pi e C_{2,\tau}^4}\|\omega^{(i)}\|_2^2 + \frac{8\sqrt{2}\alpha\beta x_{\max}^3}{\pi e C_{2,\tau}^4}\|\omega^{(i)}\|_2^2 \\
&= -\frac{2\sqrt{2}\alpha\beta x_{\max}^3\varepsilon_1}{\pi e C_{2,\tau}^4}\|\omega^{(i)}\|_1 + \frac{4\sqrt{2}\alpha\beta x_{\max}^3}{\pi e C_{2,\tau}^4}\|\omega^{(i)}\|_2^2
\end{aligned}$$

$$\overset{(11)}{\geq} -\frac{2\sqrt{2}\alpha\beta x_{\max}^3 \varepsilon_1 \gamma}{\pi e C_{2,\tau}^4}\|\omega^{(i)}\|_2 + \frac{4\sqrt{2}\alpha\beta x_{\max}^3}{\pi e C_{2,\tau}^4}\|\omega^{(i)}\|_2^2$$
$$\overset{(a)}{\geq} -\frac{2\sqrt{2}\alpha\beta x_{\max}^3}{\pi e C_{2,\tau}^4}\|\omega^{(i)}\|_2^2 + \frac{4\sqrt{2}\alpha\beta x_{\max}^3}{\pi e C_{2,\tau}^4}\|\omega^{(i)}\|_2^2 = \frac{2\sqrt{2}\alpha\beta x_{\max}^3}{\pi e C_{2,\tau}^4}\|\omega^{(i)}\|_2^2,$$

where (a) follows because $\|\omega^{(i)}\|_2 = \|\theta^{(i)} - \theta^{\star(i)}\|_2 \geq \varepsilon_1\gamma$ according to the lemma statement.

D.1.1 Proof of Lem. 10 (a): Concentration of first directional derivative

Fix some $i \in [n]$ and some $\theta^{(i)} \in \Lambda_\theta$. Let $\omega^{(i)}$ defined in (59). We claim that the first-order directional derivative of $\mathcal{L}^{(i)}$ defined in (49) is given by

$$\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{(i)})) = -\sum_{t\in[p]} \omega_t^{(i)} x_t^{(i)} \exp\Big(-[\theta_t^{(i)} + 2\widehat{\Theta}_t^\top \boldsymbol{x}^{(i)}]x_t^{(i)}\Big). \tag{63}$$

We provide a proof at the end. For now, we assume the claim and proceed.

We note that the pair $\{\mathbf{x}, \mathbf{z}\}$ corresponds to a τ-SGM (see Def. 8) with $\tau \triangleq (\alpha, \alpha\beta, x_{\max}, \Theta)$. To show the concentration, we decompose $\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{\star(i)}))$ as a sum of $L = 1024\alpha^2\beta^2 x_{\max}^4 \log 4p$ terms using Prop. 4 (see App. F) with $\lambda = \frac{1}{4\sqrt{2}x_{\max}^2}$ and focus on these L terms. Consider the L subsets $S_1, \cdots, S_L \in [p]$ obtained from Prop. 4 with $\lambda = \frac{1}{4\sqrt{2}x_{\max}^2}$ and define

$$\psi_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) \triangleq \sum_{t\in S_u} \omega_t^{(i)} x_t^{(i)} \exp\Big(-[\theta_t^{\star(i)} + 2\widehat{\Theta}_t^\top \boldsymbol{x}^{(i)}]x_t^{(i)}\Big) \quad \text{for every} \quad u \in L. \tag{64}$$

Now, we decompose $\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{\star(i)}))$ as a sum of the L terms defined above. More precisely, we have

$$\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{\star(i)})) \overset{(63)}{=} -\sum_{t\in[p]} \omega_t^{(i)} x_t^{(i)} \exp\Big(-[\theta_t^{\star(i)} + 2\widehat{\Theta}_t^\top \boldsymbol{x}^{(i)}]x_t^{(i)}\Big)$$
$$\overset{(a)}{=} -\frac{1}{L'}\sum_{u\in[L]}\sum_{t\in S_u} \omega_t^{(i)} x_t^{(i)} \exp\Big(-[\theta_t^{\star(i)} + 2\widehat{\Theta}_t^\top \boldsymbol{x}^{(i)}]x_t^{(i)}\Big)$$
$$\overset{(64)}{=} -\frac{1}{L'}\sum_{u\in[L]} \psi_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) \tag{65}$$

where (a) follows because each $t \in [p]$ appears in exactly $L' = \lceil L/32\sqrt{2}\alpha\beta x_{\max}^2 \rceil$ of the sets $S_1, \cdots, S_L$ according to Prop. 4(a) (with $\lambda = \frac{1}{4\sqrt{2}x_{\max}^2}$). Now, we focus on the L terms in (65).

Consider any $u \in [L]$. We claim that conditioned on $\boldsymbol{x}_{-S_u}^{(i)}$ and $\boldsymbol{z}^{(i)}$, the expected value of $\psi_u(\boldsymbol{x}^{(i)}; \omega^{(i)})$ can be upper bounded uniformly across all $u \in [L]$. We provide a proof at the end.

Lemma 11 (**Upper bound on expected** ψ_u)**.** *Fix $\varepsilon_5 > 0$, $\delta_5 \in (0,1)$, $i \in [n]$ and $\theta^{(i)} \in \Lambda_\theta$. Then, with $\omega^{(i)}$ defined in (59) and given $\boldsymbol{z}^{(i)}$ and $\boldsymbol{x}_{-S_u}^{(i)}$ for all $u \in [L]$, we have*

$$\max_{u\in[L]} \mathbb{E}\Big[\psi_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) \mid \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)}\Big] \leq \varepsilon_5 \|\omega^{(i)}\|_1 \quad \textit{for} \quad n \geq \frac{O(\exp(\beta))\log(4p/\delta_5)}{\varepsilon_5^4}$$

with probability at least $1 - \delta_5$.

Consider again any $u \in [L]$. Now, we claim that conditioned on $\boldsymbol{x}_{-S_u}^{(i)}$ and $\boldsymbol{z}^{(i)}$, $\psi_u(\boldsymbol{x}^{(i)}; \omega^{(i)})$ concentrates around its conditional expected value. We provide a proof at the end.

Lemma 12 (**Concentration of** ψ_u)**.** *Fix $\varepsilon_6 > 0$, $i \in [n]$, $u \in [L]$, and $\theta^{(i)} \in \Lambda_\theta$. Then, with $\omega^{(i)}$ defined in (59) and given $\boldsymbol{z}^{(i)}$ and $\boldsymbol{x}_{-S_u}^{(i)}$, we have*

$$\Big|\psi_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) - \mathbb{E}\big[\psi_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) \mid \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)}\big]\Big| \leq \varepsilon_6,$$

with probability at least $1 - \exp\left(\dfrac{-\varepsilon_6^2}{\exp\big(O(\beta)\big)\|\omega^{(i)}\|_2^2}\right)$.

Given these lemmas, we proceed to show the concentration of $\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{\star(i)}))$. To that end, for any $u \in [L]$, given $\boldsymbol{x}^{(i)}_{-S_u}$ and $\boldsymbol{z}^{(i)}$, let E_u denote the event that

$$\psi_u(\boldsymbol{x}^{(i)};\omega^{(i)}) \leq \mathbb{E}\big[\psi_u(\boldsymbol{x}^{(i)};\omega^{(i)})|\boldsymbol{x}^{(i)}_{-S_u},\boldsymbol{z}^{(i)}\big] + \frac{1}{32\sqrt{2}\alpha\beta x_{\max}^2}\varepsilon_3\|\omega^{(i)}\|_2^2. \tag{66}$$

Since E_u in an indicator event, using the law of total expectation results in

$$\mathbb{P}(E_u) = \mathbb{E}\Big[\mathbb{P}(E_u|\boldsymbol{x}^{(i)}_{-S_u},\boldsymbol{z}^{(i)})\Big] \overset{(a)}{\geq} 1 - \exp\bigg(\frac{-\varepsilon_3^2\|\omega^{(i)}\|_2^2}{\exp\big(O(\beta)\big)}\bigg).$$

where (a) follows from Lem. 12 with $\varepsilon_6 = \dfrac{\varepsilon_3\|\omega^{(i)}\|_2^2}{32\sqrt{2}\alpha\beta x_{\max}^2}$. Now, by applying the union bound over all $u \in [L]$ where $L = 1024\alpha^2\beta^2 x_{\max}^4 \log 4p$, we have

$$\mathbb{P}\Big(\bigcap_{u\in L} E_u\Big) \geq 1 - O\bigg(\beta^2 \log p \exp\bigg(\frac{-\varepsilon_3^2\|\omega^{(i)}\|_2^2}{\exp\big(O(\beta)\big)}\bigg)\bigg).$$

Now, assume the event $\cap_{u\in L} E_u$ holds. Whenever this holds, we also have

$$\begin{aligned}\big|\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{\star(i)}))\big| &\overset{(65)}{\leq} \frac{1}{L'}\sum_{u\in[L]}\big|\psi_u(\boldsymbol{x}^{(i)};\omega^{(i)})\big| \\ &\overset{(66)}{\leq} \frac{1}{L'}\sum_{u\in[L]}\Big|\mathbb{E}\big[\psi_u(\boldsymbol{x}^{(i)};\omega^{(i)})|\boldsymbol{x}^{(i)}_{-S_u},\boldsymbol{z}^{(i)}\big] + \frac{1}{32\sqrt{2}\alpha\beta x_{\max}^2}\varepsilon_3\|\omega^{(i)}\|_2^2\Big| \end{aligned}\tag{67}$$

where $L' = \lceil L/32\sqrt{2}\alpha\beta x_{\max}^2\rceil$. Further, using Lem. 11 in (67) with $\varepsilon_5 = \dfrac{\varepsilon_2}{32\sqrt{2}\alpha\beta x_{\max}^2}$ and $\delta_5 = \delta_2$, whenever

$$n \geq O\bigg(\frac{\exp\big(O(\beta)\big)\log(p/\delta_2)}{\varepsilon_2^4}\bigg),$$

with probability at least $1-\delta_2$, we have,

$$\begin{aligned}\big|\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{\star(i)}))\big| &\leq \frac{1}{L'}\sum_{u\in[L]}\Big(\frac{1}{32\sqrt{2}\alpha\beta x_{\max}^2}\varepsilon_2\|\omega^{(i)}\|_1 + \frac{1}{32\sqrt{2}\alpha\beta x_{\max}^2}\varepsilon_3\|\omega^{(i)}\|_2^2\Big) \\ &= \frac{L}{32\sqrt{2}\alpha\beta x_{\max}^2 L'}\Big(\varepsilon_2\|\omega^{(i)}\|_1 + \varepsilon_3\|\omega^{(i)}\|_2^2\Big) \overset{(a)}{\leq} \varepsilon_2\|\omega^{(i)}\|_1 + \varepsilon_3\|\omega^{(i)}\|_2^2,\end{aligned}$$

where (a) follows because $L' = \lceil L/32\sqrt{2}\alpha\beta x_{\max}^2\rceil$.

Proof of (63): Expression for first directional derivative: Fix any $i \in [n]$. The first-order partial derivatives of $\mathcal{L}^{(i)}$ (defined in (49)) with respect to the entries of the parameter vector $\theta^{(i)}$ are given by

$$\frac{\partial\mathcal{L}^{(i)}(\theta^{(i)})}{\partial\theta_t^{(i)}} = -x_t^{(i)}\exp\Big(-[\theta_t^{(i)} + 2\widehat{\Theta}_t^\top\boldsymbol{x}^{(i)}]x_t^{(i)}\Big) \quad \text{for all} \quad t\in[p].$$

Now, we can write the first-order directional derivative of $\mathcal{L}^{(i)}$ as

$$\begin{aligned}\partial_{\omega^{(i)}}(\mathcal{L}^{(i)}(\theta^{(i)})) \triangleq \lim_{h\to 0}\frac{\mathcal{L}^{(i)}(\theta^{(i)}+h\omega^{(i)}) - \mathcal{L}^{(i)}(\theta^{(i)})}{h} &= \sum_{t\in[p]}\omega_t^{(i)}\frac{\partial\mathcal{L}^{(i)}(\theta^{(i)})}{\partial\theta_t^{(i)}} \\ &= -\sum_{t\in[p]}\omega_t^{(i)}x_t^{(i)}\exp\Big(-[\theta_t^{(i)} + 2\widehat{\Theta}_t^\top\boldsymbol{x}^{(i)}]x_t^{(i)}\Big).\end{aligned}$$

Proof of Lem. 11: Upper bound on expected ψ_u: Fix any $i \in [n]$, $u \in [L]$, and $\theta^{(i)} \in \Lambda_\theta$. Then, given $\boldsymbol{x}^{(i)}_{-S_u}$ and $\boldsymbol{z}^{(i)}$, we have

$$\begin{aligned}
&\mathbb{E}\Big[\psi_u(\boldsymbol{x}^{(i)};\omega^{(i)}) \;\Big|\; \boldsymbol{x}^{(i)}_{-S_u}, \boldsymbol{z}^{(i)}\Big] \\
&\overset{(a)}{=} \mathbb{E}\Big[\sum_{t\in S_u} \omega_t^{(i)} x_t^{(i)} \exp\big(-[\theta_t^{\star(i)} + 2\widehat{\Theta}_t^\top \boldsymbol{x}^{(i)}]x_t^{(i)}\big) \;\Big|\; \boldsymbol{x}^{(i)}_{-S_u}, \boldsymbol{z}^{(i)}\Big] \\
&\overset{(b)}{=} \sum_{t\in S_u} \omega_t^{(i)} \mathbb{E}\Big[x_t^{(i)} \exp\big(-[\theta_t^{\star(i)} + 2\widehat{\Theta}_t^\top \boldsymbol{x}^{(i)}]x_t^{(i)}\big) \;\Big|\; \boldsymbol{x}^{(i)}_{-S_u}, \boldsymbol{z}^{(i)}\Big] \\
&\overset{(c)}{=} \sum_{t\in S_u} \omega_t^{(i)} \mathbb{E}\Big[\mathbb{E}\Big[x_t^{(i)} \exp\big(-[\theta_t^{\star(i)} + 2\widehat{\Theta}_t^\top \boldsymbol{x}^{(i)}]x_t^{(i)}\big) \;\Big|\; \boldsymbol{x}^{(i)}_{-t}, \boldsymbol{z}^{(i)}\Big] \;\Big|\; \boldsymbol{x}^{(i)}_{-S_u}, \boldsymbol{z}^{(i)}\Big],
\end{aligned} \tag{68}$$

where (a) follows from the definition of $\psi_u(\boldsymbol{x}^{(i)};\omega^{(i)})$ in (64), (b) follows from linearity of expectation, and (c) follows from the law of total expectation, i.e., $\mathbb{E}[\mathbb{E}[Y|X,Z]|Z] = \mathbb{E}[Y|Z]$ since $\boldsymbol{x}^{(i)}_{-S_u} \subseteq \boldsymbol{x}^{(i)}_{-t}$.

Now, for every $t \in S_u$, we will bound $\mathbb{E}\Big[x_t^{(i)} \exp\big(-[\theta_t^{\star(i)} + 2\widehat{\Theta}_t^\top \boldsymbol{x}^{(i)}]x_t^{(i)}, \boldsymbol{z}^{(i)}\big) \;\Big|\; \boldsymbol{x}^{(i)}_{-t}, \boldsymbol{z}^{(i)}\Big]$. We have

$$\begin{aligned}
&\mathbb{E}\Big[x_t^{(i)} \exp\big(-[\theta_t^{\star(i)} + 2\widehat{\Theta}_t^\top \boldsymbol{x}^{(i)}]x_t^{(i)}\big) \;\Big|\; \boldsymbol{x}^{(i)}_{-t}, \boldsymbol{z}^{(i)}\Big] \\
&= \int_{\mathcal{X}} x_t^{(i)} \exp\big(-[\theta_t^{\star(i)} + 2\widehat{\Theta}_t^\top \boldsymbol{x}^{(i)}]x_t^{(i)}\big) f_{\mathsf{x}_t|\mathsf{x}_{-t},\mathsf{z}}\big(x_t^{(i)}|\boldsymbol{x}^{(i)}_{-t}, \boldsymbol{z}^{(i)}; \theta_t^\star(\boldsymbol{z}^{(i)}), \Theta_t^\star\big) dx_t^{(i)} \\
&\overset{(8)}{=} \frac{\int_{\mathcal{X}} x_t^{(i)} \exp\big(-[\theta_t^{\star(i)} + 2\widehat{\Theta}_t^\top \boldsymbol{x}^{(i)}]x_t^{(i)}\big) \exp\big([\theta_t^\star(\boldsymbol{z}^{(i)}) + 2\Theta_t^{\star\top} \boldsymbol{x}^{(i)}]x_t^{(i)}\big) dx_t^{(i)}}{\int_{\mathcal{X}} \exp\big([\theta_t^\star(\boldsymbol{z}^{(i)}) + 2\Theta_t^{\star\top} \boldsymbol{x}^{(i)}]x_t^{(i)}\big) dx_t^{(i)}} \\
&\overset{(a)}{=} \frac{\int_{\mathcal{X}} x_t^{(i)} \exp\big(2[\Theta_t^\star - \widehat{\Theta}_t]^\top \boldsymbol{x}^{(i)} x_t^{(i)}\big) dx_t^{(i)}}{\int_{\mathcal{X}} \exp\big([\theta_t^\star(\boldsymbol{z}^{(i)}) + 2\Theta_t^{\star\top} \boldsymbol{x}^{(i)}]x_t^{(i)}\big) dx_t^{(i)}} \\
&\overset{(b)}{=} \frac{\int_{\mathcal{X}} x_t^{(i)} \Big[1+2\big([\Theta_t^\star-\widehat{\Theta}_t]^\top \boldsymbol{x}^{(i)} x_t^{(i)}\big)+4\big([\Theta_t^\star-\widehat{\Theta}_t]^\top \boldsymbol{x}^{(i)} x_t^{(i)}\big)^2+o\big([\Theta_t^\star-\widehat{\Theta}_t]^\top \boldsymbol{x}^{(i)} x_t^{(i)}\big)^3\Big] dx_t^{(i)}}{\int_{\mathcal{X}} \exp\big([\theta_t^\star(\boldsymbol{z}^{(i)}) + 2\Theta_t^{\star\top} \boldsymbol{x}^{(i)}]x_t^{(i)}\big) dx_t^{(i)}} \\
&\overset{(c)}{=} \frac{4x_{\max}^3[\Theta_t^\star - \widehat{\Theta}_t]^\top \boldsymbol{x}^{(i)}}{3\int_{\mathcal{X}} \exp\Big([\theta_t^\star(\boldsymbol{z}^{(i)}) + 2\Theta_t^{\star\top} \boldsymbol{x}^{(i)}]x_t^{(i)}\Big) dx_t^{(i)}} + \frac{x_{\max}^5\big([\Theta_t^\star - \widehat{\Theta}_t]^\top \boldsymbol{x}^{(i)}\big)^3 o(1)}{\int_{\mathcal{X}} \exp\Big([\theta_t^\star(\boldsymbol{z}^{(i)}) + 2\Theta_t^{\star\top} \boldsymbol{x}^{(i)}]x_t^{(i)}\Big) dx_t^{(i)}},
\end{aligned} \tag{69}$$

where (a) follows because $\theta^{\star(i)} = \theta^\star(\boldsymbol{z}^{(i)})\ \forall i \in [n]$, (b) follows by using the Taylor series expansion $\exp(y) = 1 + y + y^2 + o(y^3)$ around zero, (c) follows because $\int_{\mathcal{X}} x_t^{(i)} dx_t^{(i)} = 0$, $\int_{\mathcal{X}} \big(x_t^{(i)}\big)^2 dx_t^{(i)} = 2x_{\max}^3/3$, $\int_{\mathcal{X}} \big(x_t^{(i)}\big)^3 dx_t^{(i)} = 0$, and $\int_{\mathcal{X}} \big(x_t^{(i)}\big)^4 dx_t^{(i)} = 2x_{\max}^5/5$.

Now, we bound the numerators in (69) by using $\|\Theta_t^\star - \widehat{\Theta}_t\|_1 \leq 2\beta\|\Theta_t^\star - \widehat{\Theta}_t\|_\infty \leq 2\beta\|\Theta_t^\star - \widehat{\Theta}_t\|_2$ which follows from the order of norms on Euclidean space as well as $\Theta^\star \in \Lambda_\Theta$ and $\widehat{\Theta} \in \Lambda_\Theta$. Then, we invoke Thm. 1 to bound $\|\Theta_t^\star - \widehat{\Theta}_t\|_2$ by $\varepsilon = \dfrac{3\varepsilon_5}{4\beta C_{2,\tau} x_{\max}^3}$. Therefore, we subsume the second term by the first term resulting in the following bound:

$$\mathbb{E}\Big[x_t^{(i)} \exp\big(-[\theta_t^{\star(i)} + 2\widehat{\Theta}_t^\top \boldsymbol{x}^{(i)}]x_t^{(i)}\big) \;\Big|\; \boldsymbol{x}^{(i)}_{-t}, \boldsymbol{z}^{(i)}\Big] \leq \frac{4\beta C_{2,\tau} x_{\max}^3 \|\Theta_t^\star - \widehat{\Theta}_t\|_2}{3}, \tag{70}$$

where we have used the triangle inequality, $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$ as well as $\|\Theta_t^\star - \widehat{\Theta}_t\|_1 \leq 2\beta\|\Theta_t^\star - \widehat{\Theta}_t\|_2$ to upper bound the numerator, and the arguments used in the proof of Lem. 7 as well as $\int_{\mathcal{X}} dx_t^{(i)} = 2x_{\max}$ to lower bound the denominator.

Using Thm. 1 in (70) with $\varepsilon = \dfrac{3\varepsilon_5}{4\beta C_{2,\tau} x_{\max}^3}$ and $\delta = \delta_5$, we have

$$\mathbb{E}\Big[x_t^{(i)} \exp\Big(- [\theta_t^{\star(i)} + 2\widehat{\Theta}_t^\top \boldsymbol{x}^{(i)}] x_t^{(i)}\Big) \;\Big|\; \boldsymbol{x}_{-t}^{(i)}, \boldsymbol{z}^{(i)}\Big] \le \varepsilon_5, \tag{71}$$

with probability at least $1 - \delta_5$ as long as

$$n \ge \frac{O(\exp(\beta)) \log(4p/\delta_5)}{\varepsilon_5^4}. \tag{72}$$

Using (71) and triangle inequality in (68), we have

$$\mathbb{E}\Big[\psi_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) \mid \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)}\Big] \le \varepsilon_5 \sum_{t \in S_u} \big|\omega_t^{(i)}\big| \le \varepsilon_5 \|\omega^{(i)}\|_1,$$

with probability at least $1 - \delta_5$ as long as n satisfies (72).

Proof of Lem. 12: Concentration of ψ_u: To show this concentration result, we use Cor. 2 (140) for the function q_2. To that end, we note that the pair $\{\mathbf{x}, \mathbf{z}\}$ corresponds to a τ-SGM (Def. 8) with $\tau \triangleq (\alpha, \alpha\beta, x_{\max}, \Theta)$. However, the random vector $\mathbf{x}$ conditioned on $\mathbf{z}$ need not satisfy the Dobrushin's uniqueness condition (Def. 4). Therefore, we cannot apply Cor. 2 (140) as is. To resolve this, we resort to Prop. 4 with $\lambda = \frac{1}{4\sqrt{2}x_{\max}^2}$ to reduce the random vector $\mathbf{x}$ conditioned on $\mathbf{z}$ to Dobrushin's regime.

Fix any $u \in [L]$. Then, from Prop. 4(b), (i) the pair of random vectors $\{\mathbf{x}_{S_u}, (\mathbf{x}_{-S_u}, \mathbf{z})\}$ corresponds to a τ_1-SGM with $\tau_1 \triangleq (\alpha + 2\alpha\beta x_{\max}, \frac{1}{4\sqrt{2}x_{\max}^2}, x_{\max}, \Theta_{S_u})$, and (ii) the random vector $\mathbf{x}_{S_u}$ conditioned on $(\mathbf{x}_{-S_u}, \mathbf{z})$ satisfies the Dobrushin's uniqueness condition (Def. 4) with coupling matrix $2\sqrt{2}x_{\max}^2 \Theta_{S_u}$ with $2\sqrt{2}x_{\max}^2 \|\Theta_{S_u}\|_{\mathrm{op}} \le 2\sqrt{2}x_{\max}^2 \lambda \le 1/2$. Now, for any fixed $i \in [n]$, we apply Cor. 2 (140) for the function q_2 with $\varepsilon = \varepsilon_6$ for a given $\boldsymbol{x}_{-S_u}^{(i)}$ and $\boldsymbol{z}^{(i)}$, to obtain

$$\mathbb{P}\bigg(\Big|\psi_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) - \mathbb{E}\Big[\psi_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) \;\Big|\; \boldsymbol{x}_{-S_u}^{(i)}, \mathbf{z}\Big]\Big| \ge \varepsilon_6 \;\Big|\; \boldsymbol{x}_{-S_u}^{(i)}, \mathbf{z}\bigg) \le \exp\bigg(\frac{-\varepsilon_6^2}{\exp\big(O(\beta)\big)\|\omega^{(i)}\|_2^2}\bigg).$$

D.1.2 Proof of Lem. 10 (b): Anti-concentration of second directional derivative

Fix some $i \in [n]$ and some $\theta^{(i)} \in \Lambda_\theta$. Let $\omega^{(i)}$ defined in (59). We claim that the second-order directional derivative of $\mathcal{L}^{(i)}$ defined in (49) is given by

$$\partial^2_{[\omega^{(i)}]^2} \mathcal{L}^{(i)}(\theta^{(i)}) = \sum_{t \in [p]} \big(\omega_t^{(i)} x_t^{(i)}\big)^2 \exp\Big(- [\theta_t^{(i)} + 2\widehat{\Theta}_t^\top \boldsymbol{x}^{(i)}] x_t^{(i)}\Big). \tag{73}$$

We provide a proof at the end. For now, we assume the claim and proceed. Now, we lower bound $\partial^2_{[\omega^{(i)}]^2} \mathcal{L}^{(i)}(\theta^{(i)})$ by a quadratic form as follows:

$$\begin{aligned}
\partial^2_{[\omega^{(i)}]^2} \mathcal{L}^{(i)}(\theta^{(i)}) &\overset{(a)}{\ge} \sum_{t \in [p]} \big(\omega_t^{(i)} x_t^{(i)}\big)^2 \times \exp\Big(- |\theta_t^{(i)} + 2\widehat{\Theta}_t^\top \boldsymbol{x}^{(i)}| x_{\max}\Big) \\
&\overset{(b)}{\ge} \sum_{t \in [p]} \big(\omega_t^{(i)} x_t^{(i)}\big)^2 \times \exp\Big(- \big(|\theta_t^{(i)}| + 2\|\widehat{\Theta}_t\|_1 \|\boldsymbol{x}^{(i)}\|_\infty\big) x_{\max}\Big) \\
&\overset{(c)}{\ge} \sum_{t \in [p]} \big(\omega_t^{(i)} x_t^{(i)}\big)^2 \times \exp\Big(- (\alpha + 2\alpha\beta x_{\max}) x_{\max}\Big) \overset{(29)}{=} \frac{1}{C_{2,\tau}} \sum_{t \in [p]} \big(\omega_t^{(i)} x_t^{(i)}\big)^2,
\end{aligned} \tag{74}$$

where (a) follows from (73) because $\|\boldsymbol{x}^{(i)}\|_\infty \le x_{\max}$ for all $i \in [n]$, (b) follows by triangle inequality and Cauchy–Schwarz inequality, and (c) follows because $\widehat{\Theta} \in \Lambda_\Theta$, $\theta^{(i)} \in \Lambda_\theta$, and $\|\boldsymbol{x}^{(i)}\|_\infty \le x_{\max}$ for all $i \in [n]$.

Now, to show the anti-concentration of $\partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\theta^{(i)})$, we show the anti-concentration of the quadratic form in (74). To that end, we note that the pair $\{\mathbf{x}, \mathbf{z}\}$ corresponds to a τ-SGM (Def. 8) with $\tau \triangleq (\alpha, \alpha\beta, x_{\max}, \Theta)$. Then, we decompose the quadratic form in (74) as a sum of $L = 1024\alpha^2\beta^2 x_{\max}^4 \log 4p$ terms using Prop. 4 (see App. F) with $\lambda = \frac{1}{4\sqrt{2}x_{\max}^2}$ and focus on these L terms. Consider the L subsets $S_1, \cdots, S_L \in [p]$ obtained from Prop. 4 and define

$$\overline{\psi}_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) \triangleq \sum_{t \in S_u} \big(\omega_t^{(i)} x_t^{(i)}\big)^2 \quad \text{for every} \quad u \in L. \tag{75}$$

Then, we have

$$\sum_{t\in[p]} \big(\omega_t^{(i)} x_t^{(i)}\big)^2 \overset{(a)}{=} \frac{1}{L'} \sum_{u\in[L]} \sum_{t\in S_u} \big(\omega_t^{(i)} x_t^{(i)}\big)^2 \overset{(75)}{=} \frac{1}{L'} \sum_{u\in[L]} \overline{\psi}_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) \tag{76}$$

where (a) follows because each $t \in [p]$ appears in exactly $L' = \lceil L/32\sqrt{2}\alpha\beta x_{\max}^2 \rceil$ of the sets $S_1, \cdots, S_L$ according to Prop. 4(a) (with $\lambda = \frac{1}{4\sqrt{2}x_{\max}^2}$). Now, we focus on the L terms in (76).

Consider any $u \in [L]$. We claim that conditioned on $\boldsymbol{x}_{-S_u}^{(i)}$ and $\boldsymbol{z}^{(i)}$, the expected value of $\overline{\psi}_u(\boldsymbol{x}^{(i)}; \omega^{(i)})$ can be upper bounded uniformly across all $u \in [L]$. We provide a proof at the end.

Lemma 13 (Lower bound on expected $\overline{\psi}_u$). *Fix $i \in [n]$ and $\theta^{(i)} \in \Lambda_\theta$. Then, with $\omega^{(i)}$ defined in (59) and given $\boldsymbol{z}^{(i)}$ and $\boldsymbol{x}_{-S_u}^{(i)}$, we have given $\boldsymbol{x}_{-S_u}^{(i)}$ and $\boldsymbol{z}^{(i)}$,*

$$\min_{u\in[L]} \mathbb{E}\Big[\overline{\psi}_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) \mid \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)}\Big] \geq \frac{x_{\max}}{\pi e C_{2,\tau}^2} \|\omega^{(i)}\|_2^2,$$

where the constant $C_{2,\tau}$ was defined in (29).

Consider again any $u \in [L]$. Now, we claim that conditioned on $\boldsymbol{x}_{-S_u}^{(i)}$ and $\boldsymbol{z}^{(i)}$, $\overline{\psi}_u(\boldsymbol{x}^{(i)}; \omega^{(i)})$ concentrates around its conditional expected value. We provide a proof at the end.

Lemma 14 (Concentration of $\overline{\psi}_u$). *Fix $\varepsilon_7 > 0$, $i \in [n]$, $u \in [L]$, and $\theta^{(i)} \in \Lambda_\theta$. Then, with $\omega^{(i)}$ defined in (59) and given $\boldsymbol{z}^{(i)}$ and $\boldsymbol{x}_{-S_u}^{(i)}$, we have*

$$\Big|\overline{\psi}_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) - \mathbb{E}\big[\overline{\psi}_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) \mid \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)}\big]\Big| \leq \varepsilon_7,$$

with probability at least $1 - \exp\left(\dfrac{-\varepsilon_7^2}{\exp\big(O(\beta)\big)\|\omega^{(i)}\|_2^2}\right)$.

Given these lemmas, we proceed to show the anti-concentration of the quadratic form in (74) implying the anti-concentration of $\partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\theta^{(i)})$. To that end, for any $u \in [L]$, given $\boldsymbol{x}_{-S_u}^{(i)}$ and $\boldsymbol{z}^{(i)}$, let E_u denote the event that

$$\overline{\psi}_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) \geq \mathbb{E}\big[\overline{\psi}_u(\boldsymbol{x}^{(i)}; \omega^{(i)}) | \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)}\big] - \frac{x_{\max}}{2\pi e C_{2,\tau}^2} \|\omega^{(i)}\|_2^2. \tag{77}$$

Since E_u in an indicator event, using the law of total expectation results in

$$\mathbb{P}(E_u) = \mathbb{E}\Big[\mathbb{P}(E_u | \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)})\Big] \overset{(a)}{\geq} 1 - \exp\left(\frac{\|\omega^{(i)}\|_2^2}{\exp\big(O(\beta)\big)}\right),$$

where (a) follows from Lem. 14 with $\varepsilon_7 = \dfrac{x_{\max}}{2\pi e C_{2,\tau}^2}\|\omega^{(i)}\|_2^2$. Now, by applying the union bound over all $u \in [L]$ where $L = 1024\alpha^2\beta^2 x_{\max}^4 \log 4p$, we have

$$\mathbb{P}\Big(\bigcap_{u\in L} E_u\Big) \geq 1 - O\left(\beta^2 \log p \exp\left(\frac{\|\omega^{(i)}\|_2^2}{\exp\big(O(\beta)\big)}\right)\right).$$

Now, assume the event $\cap_{u\in L} E_u$ holds. Whenever this holds, we also have

$$\sum_{t\in[p]} \big(\omega_t^{(i)} x_t^{(i)}\big)^2 \overset{(76)}{=} \frac{1}{L'} \sum_{u\in[L]} \overline{\psi}_u(\boldsymbol{x}^{(i)};\omega^{(i)})$$
$$\overset{(77)}{\geq} \frac{1}{L'} \sum_{u\in[L]} \bigg(\mathbb{E}\big[\overline{\psi}_u(\boldsymbol{x}^{(i)};\omega^{(i)})|\boldsymbol{x}_{-S_u}^{(i)},\boldsymbol{z}^{(i)}\big] - \frac{x_{\max}}{2\pi e C_{2,\tau}^2}\|\omega^{(i)}\|_2^2\bigg)$$
$$\overset{(a)}{\geq} \frac{1}{L'} \sum_{u\in[L]} \frac{x_{\max}}{2\pi e C_{2,\tau}^2}\|\omega^{(i)}\|_2^2 = \frac{x_{\max} L}{2\pi e L' C_{2,\tau}^2}\|\omega^{(i)}\|_2^2 \tag{78}$$

where $L' = \lceil L/32\sqrt{2}\alpha\beta x_{\max}^2 \rceil$ and (a) follows from Lem. 13. Finally, approximating $L' = L/32\sqrt{2}\alpha\beta x_{\max}^2$ and using (74), we have

$$\partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\theta^{(i)}) \geq \frac{1}{C_{2,\tau}} \sum_{t\in[p]} \big(\omega_t^{(i)} x_t^{(i)}\big)^2 \overset{(78)}{\geq} \frac{16\sqrt{2}\alpha\beta x_{\max}^3}{\pi e C_{2,\tau}^3}\|\omega^{(i)}\|_2^2,$$

which completes the proof.

Proof of (73): Expression for second directional derivative: Fix any $i \in [n]$. The second-order partial derivatives of $\mathcal{L}^{(i)}$ (defined in (49)) with respect to the entries of the parameter vector $\theta^{(i)}$ are given by

$$\frac{\partial^2 \mathcal{L}^{(i)}(\theta^{(i)})}{\partial\big[\theta_t^{(i)}\big]^2} = \big[x_t^{(i)}\big]^2 \exp\Big(-[\theta_t^{(i)} + 2\widehat{\Theta}_t^\top \boldsymbol{x}^{(i)}]x_t^{(i)}\Big) \quad \text{for all} \quad t\in[p].$$

Now, we can write the second-order directional derivative of $\mathcal{L}^{(i)}$ as

$$\partial^2_{[\omega^{(i)}]^2}\mathcal{L}^{(i)}(\theta^{(i)}) \triangleq \lim_{h\to 0} \frac{\partial_{\omega^{(i)}}\mathcal{L}^{(i)}(\theta^{(i)} + h\omega^{(i)}) - \partial_{\omega^{(i)}}\mathcal{L}^{(i)}(\theta^{(i)})}{h} = \sum_{t\in[p]} \big[\omega_t^{(i)}\big]^2 \frac{\partial^2 \mathcal{L}^{(i)}(\theta^{(i)})}{\partial\big[\theta_t^{(i)}\big]^2}$$
$$= \sum_{t\in[p]} \big(\omega_t^{(i)} x_t^{(i)}\big)^2 \exp\Big(-[\theta_t^{(i)} + 2\widehat{\Theta}_t^\top \boldsymbol{x}^{(i)}]x_t^{(i)}\Big).$$

Proof of Lem. 13: Lower bound on expected $\overline{\psi}_u$: Fix any $i \in [n]$, $u \in [L]$, and $\theta^{(i)} \in \Lambda_\theta$. Then, given $\boldsymbol{x}_{-S_u}^{(i)}$ and $\boldsymbol{z}^{(i)}$, we have

$$\mathbb{E}\Big[\overline{\psi}_u(\boldsymbol{x}^{(i)};\omega^{(i)}) \mid \boldsymbol{x}_{-S_u}^{(i)},\boldsymbol{z}^{(i)}\Big] \overset{(75)}{=} \mathbb{E}\Big[\sum_{t\in S_u} \big(\omega_t^{(i)} x_t^{(i)}\big)^2 \mid \boldsymbol{x}_{-S_u}^{(i)},\boldsymbol{z}^{(i)}\Big]$$
$$\overset{(a)}{=} \sum_{t\in S_u} \mathbb{E}\Big[\big(\omega_t^{(i)} x_t^{(i)}\big)^2 \mid \boldsymbol{x}_{-S_u}^{(i)},\boldsymbol{z}^{(i)}\Big]$$
$$\overset{(b)}{=} \sum_{t\in S_u} \mathbb{E}\Big[\mathbb{E}\Big[\big(\omega_t^{(i)} x_t^{(i)}\big)^2 \Big| \boldsymbol{x}_{-t}^{(i)},\boldsymbol{z}^{(i)}\Big] \Bigm| \boldsymbol{x}_{-S_u}^{(i)},\boldsymbol{z}^{(i)}\Big]$$
$$\overset{(c)}{\geq} \sum_{t\in S_u} \mathbb{E}\Big[\mathbb{V}\mathrm{ar}\Big(\omega_t^{(i)} x_t^{(i)} \Big| \boldsymbol{x}_{-t}^{(i)},\boldsymbol{z}^{(i)}\Big) \Bigm| \boldsymbol{x}_{-S_u}^{(i)},\boldsymbol{z}^{(i)}\Big]$$
$$\overset{(d)}{\geq} \frac{x_{\max}}{\pi e C_{2,\tau}^2}\|\omega^{(i)}\|_2^2,$$

where (a) follows from linearity of expectation, (b) follows from the law of total expectation i.e., $\mathbb{E}[\mathbb{E}[Y|X,Z]|Z] = \mathbb{E}[Y|Z]$ since $\boldsymbol{x}_{-S_u}^{(i)} \subseteq \boldsymbol{x}_{-t}^{(i)}$, (c) follows follows from the fact that for any random variable a, $\mathbb{E}[a^2] \geq \mathbb{V}\mathrm{ar}[a]$, and (d) follows from Lem. 7.

Proof of Lem. 14: Concentration of $\overline{\psi}_u$: To show this concentration result, we use Cor. 2 (140) for the function q_1. To that end, we note that the pair $\{\mathbf{x},\mathbf{z}\}$ corresponds to a τ-SGM (Def. 8) with $\tau \triangleq (\alpha, \alpha\beta, x_{\max}, \Theta)$. However, the random vector $\mathbf{x}$ conditioned on $\mathbf{z}$ need not satisfy the Dobrushin's uniqueness condition (Def. 4). Therefore, we cannot apply Cor. 2 (140) as is. To resolve this, we resort to Prop. 4 with $\lambda = \frac{1}{4\sqrt{2}x_{\max}^2}$ to reduce the random vector $\mathbf{x}$ conditioned on $\mathbf{z}$ to Dobrushin's regime.

Fix any $u \in [L]$. Then, from Prop. 4(b), (i) the pair of random vectors $\{\mathbf{x}_{S_u}, (\mathbf{x}_{-S_u}, \mathbf{z})\}$ corresponds to a τ_1-SGM with $\tau_1 \triangleq (\alpha + 2\alpha\beta x_{\max}, \frac{1}{4\sqrt{2}x_{\max}^2}, x_{\max}, \Theta_{S_u})$, and (ii) the random vector $\mathbf{x}_{S_u}$ conditioned on $(\mathbf{x}_{-S_u}, \mathbf{z})$ satisfies the Dobrushin's uniqueness condition (Def. 4) with coupling matrix $2\sqrt{2}x_{\max}^2\Theta_{S_u}$ with $2\sqrt{2}x_{\max}^2 \|\Theta_{S_u}\|_{\text{op}} \leq 2\sqrt{2}x_{\max}^2\lambda \leq 1/2$. Now, for any fixed $i \in [n]$, we apply Cor. 2 (140) for the function q_1 with $\varepsilon = \varepsilon_7$ for a given $\boldsymbol{x}_{-S_u}^{(i)}$ and $\boldsymbol{z}^{(i)}$, to obtain

$$\mathbb{P}\biggl(\Bigl|\overline{\psi}_u(\boldsymbol{x}^{(i)};\omega^{(i)}) - \mathbb{E}\Bigl[\overline{\psi}_u(\boldsymbol{x}^{(i)};\omega^{(i)})\Bigm|\boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)}\Bigr]\Bigr| \geq \varepsilon_7 \Bigm| \boldsymbol{x}_{-S_u}^{(i)}, \boldsymbol{z}^{(i)}\biggr) \leq \exp\left(\frac{-\varepsilon_7^2}{\exp\bigl(O(\beta)\bigr)\|\omega^{(i)}\|_2^2}\right).$$

D.2 Proof of Lem. 9: Lipschitzness of the loss function

Fix any $i \in [n]$, any $\theta^{(i)}, \widetilde{\theta}^{(i)} \in \Lambda_\theta$. Consider the direction $\omega^{(i)} = \widetilde{\theta}^{(i)} - \theta^{(i)}$, and define the function $q : [0,1] \to \mathbb{R}$ as follows:

$$q(a) = \mathcal{L}^{(i)}\bigl(\theta^{(i)} + a(\widetilde{\theta}^{(i)} - \theta^{(i)})\bigr). \tag{79}$$

Then, the desired inequality in (51) is equivalent to $|q(1) - q(0)| \leq x_{\max}C_{2,\tau}\|\omega^{(i)}\|_1$. From the mean value theorem, there exists $a' \in (0,1)$ such that

$$|q(1) - q(0)| = \left|\frac{dq(a')}{da}\right|. \tag{80}$$

Therefore, we have

$$\begin{aligned}
\bigl|q(1) - q(0)\bigr| &\overset{(80)}{=} \left|\frac{dq(a')}{da}\right| \\
&\overset{(79)}{=} \left|\frac{d\mathcal{L}^{(i)}\bigl(\theta^{(i)} + a'(\widetilde{\theta}^{(i)} - \theta^{(i)})\bigr)}{da}\right| \\
&\overset{(60)}{=} \left|\partial_{\omega^{(i)}}\bigl(\mathcal{L}^{(i)}(\theta^{(i)})\bigr)\big|_{\theta^{(i)} = \theta^{(i)} + a'(\widetilde{\theta}^{(i)} - \theta^{(i)})}\right| \\
&\overset{(63)}{=} \Bigl|\sum_{t\in[p]} \omega_t^{(i)} x_t^{(i)} \exp\Bigl(-[\theta_t^{(i)} + a'(\widetilde{\theta}_t^{(i)} - \theta_t^{(i)}) + 2\widehat{\Theta}_t^\top \boldsymbol{x}^{(i)}]x_t^{(i)}\Bigr)\Bigr| \\
&\overset{(a)}{\leq} x_{\max} \sum_{t\in[p]} \bigl|\omega_t^{(i)}\bigr| \exp\Bigl(\Bigl[\bigl|(1-a')\theta_t^{(i)}\bigr| + \bigl|a'\widetilde{\theta}_t^{(i)}\bigr| + 2\|\widehat{\Theta}_t\|_1\|\boldsymbol{x}^{(i)}\|_\infty\Bigr]x_{\max}\Bigr) \\
&\overset{(b)}{\leq} x_{\max} \exp\Bigl(\bigl((1-a')\alpha + a'\alpha + 2\alpha\beta x_{\max}\bigr)x_{\max}\Bigr) \sum_{t\in[p]} \bigl|\omega_t^{(i)}\bigr| \overset{(29)}{=} x_{\max}C_{2,\tau}\|\omega^{(i)}\|_1,
\end{aligned}$$

where (a) follows from triangle inequality, Cauchy–Schwarz inequality, and because $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$ and (b) follows because $\theta^{(i)}, \widetilde{\theta}^{(i)} \in \Lambda_\theta$, $\widehat{\Theta} \in \Lambda_\Theta$, and $\|\boldsymbol{x}^{(i)}\|_\infty \leq x_{\max}$ for all $i \in [n]$.

E Logarithmic Sobolev inequality and tail bounds

In this section, we present two results which may be of independent interest. First, we show that a random vector supported on a compact set satisfies the logarithmic Sobolev inequality (to be defined) if it satisfies the Dobrushin's uniqueness condition (to be defined). This result is a generalization of the result in [20] for discrete random vectors to continuous random vectors supported on a compact set. Next, we show that if a random vector satisfies the logarithmic Sobolev inequality, then any arbitrary function of the random vector concentrates around its mean. This result is a generalization of the result in [12] for discrete random vectors to continuous random vectors.

Throughout this section, we consider a p-dimensional random vector $\mathbf{x}$ supported on $\mathcal{X}^p$ with distribution $f_{\mathbf{x}}$ where $p \geq 1$. We start by defining the logarithmic Sobolev inequality (LSI). We use the convention $0 \log 0 = 0$.

Definition 3 (Logarithmic Sobolev inequality). *A random vector $\mathbf{x}$ satisfies the logarithmic Sobolev inequality with constant $\sigma^2 > 0$ (abbreviated as $\mathrm{LSI}_{\mathbf{x}}(\sigma^2)$) if*

$$\mathsf{Ent}_{\mathbf{x}}\left(q^2\right) \leq \sigma^2 \mathbb{E}_{\mathbf{x}}\Big[\|\nabla_{\mathbf{x}} q(\mathbf{x})\|_2^2\Big] \quad \textit{for all} \quad q: \mathcal{X}^p \to \mathbb{R}, \tag{81}$$

where $\mathsf{Ent}_{\mathbf{x}}(g) \triangleq \mathbb{E}_{\mathbf{x}}[g(\mathbf{x}) \log g(\mathbf{x})] - \mathbb{E}_{\mathbf{x}}[g(\mathbf{x})] \log \mathbb{E}_{\mathbf{x}}[g(\mathbf{x})]$ denotes the entropy of the function $g: \mathcal{X}^p \to \mathbb{R}_+$.

Next, we restate the Dobrushin's uniqueness condition [20].

Definition 4. *[20, **Dobrushin's uniqueness condition**] A random vector $\mathbf{x}$ satisfies the Dobrushin's uniqueness condition with coupling matrix $\Theta \in \mathbb{R}^{p\times p}$ with $\Theta_{tt} = 0$ for all $t \in [p]$, if $\|\Theta\|_{\mathrm{op}} < 1$, and for every $t \in [p], u \in [p]\setminus\{t\}$, and $\boldsymbol{x}_{-t}, \widetilde{\boldsymbol{x}}_{-t} \in \mathcal{X}^{p-1}$ differing only in the u^{th} coordinate,*

$$\|f_{\mathrm{x}_t|\mathbf{x}_{-t}=\boldsymbol{x}_{-t}} - f_{\mathrm{x}_t|\mathbf{x}_{-t}=\widetilde{\boldsymbol{x}}_{-t}}\|_{\mathsf{TV}} \leq \Theta_{tu}. \tag{82}$$

From hereon, we let $\mathcal{X}^p$ be compact unless otherwise specified. Moreover, we define

$$f_{\min} \triangleq \min_{t\in[p], \boldsymbol{x}\in\mathcal{X}^p} f_{\mathrm{x}_t|\mathbf{x}_{-t}}(x_t|\boldsymbol{x}_{-t}). \tag{83}$$

Now, we provide the first main result of this section with a proof in App. E.1.

Proposition 2 (Logarithmic Sobolev inequality). *If a random vector $\mathbf{x}$ with $f_{\min} > 0$ (see (83)) satisfies (a) the Dobrushin's uniqueness condition (Def. 4) with coupling matrix $\Theta \in \mathbb{R}^{p\times p}$, and (b) $\mathrm{x}_t|\mathbf{x}_{-t}$ satisfies $\mathrm{LSI}_{\mathrm{x}_t|\mathbf{x}_{-t}=\boldsymbol{x}_{-t}}(\sigma^2)$ for all $t \in [p]$ and $\boldsymbol{x}_{-t} \in \mathcal{X}^{p-1}$, then it satisfies $\mathrm{LSI}_{\mathbf{x}}(2\sigma^2/(f_{\min}(1-\|\Theta\|_{\mathrm{op}})^2))$.*

Next, we define the notion of pseudo derivative and pseudo Hessian that come in handy in our proofs for providing upper bounds on the norm of the derivative and the Hessian.

Definition 5 (Pseudo derivative and Hessian). *For a function $q: \mathcal{X}^p \to \mathbb{R}$, the functions $\widetilde{\nabla} q: \mathcal{X}^p \to \mathbb{R}^{p_1}$ and $\widetilde{\nabla}^2 q: \mathcal{X}^p \to \mathbb{R}^{p_1\times p_2}$ ($p_1, p_2 \geq 1$) are respectively called a pseudo derivative and a pseudo Hessian for q if for all $\boldsymbol{y} \in \mathcal{X}^p$ and $\rho \in \mathbb{R}^{p_1\times 1}$, we have*

$$\|\widetilde{\nabla} q(\boldsymbol{y})\|_2 \geq \|\nabla q(\boldsymbol{y})\|_2 \quad \textit{and} \quad \|\rho^\top \widetilde{\nabla}^2 q(\boldsymbol{y})\|_2 \geq \|\nabla\big[\rho^\top \widetilde{\nabla} q(\boldsymbol{y})\big]\|_2. \tag{84}$$

Finally, we provide the second main result of this section with a proof in App. E.2.

Proposition 3 (Tail bounds for arbitrary functions under LSI). *Given a random vector $\mathbf{x}$ satisfying $\mathrm{LSI}_{\mathbf{x}}(\sigma^2)$, any function $q: \mathcal{X}^p \to \mathbb{R}$ with a pseudo derivative $\widetilde{\nabla} q$, and pseudo Hessian $\widetilde{\nabla}^2 q$ (see Def. 5) satisfies a tail bound, namely for any fixed $\varepsilon > 0$, we have*

$$\mathbb{P}\Big[\big|q(\mathbf{x}) - \mathbb{E}\big[q(\mathbf{x})\big]\big| \geq \varepsilon\Big] \leq \exp\left(\frac{-c}{\sigma^4} \min\left(\frac{\varepsilon^2}{\mathbb{E}\big[\|\widetilde{\nabla} q(\mathbf{x})\|_2\big]^2 + \max_{\boldsymbol{x}\in\mathcal{X}^p} \|\widetilde{\nabla}^2 q(\boldsymbol{x})\|_{\mathrm{F}}^2}, \frac{\varepsilon}{\max_{\boldsymbol{x}\in\mathcal{X}^p} \|\widetilde{\nabla}^2 q(\boldsymbol{x})\|_{\mathrm{op}}}\right)\right),$$

where c is a universal constant.

E.1 Proof of Prop. 2: Logarithmic Sobolev inequality

We start by defining the notion of W_2 distance [20] which is useful in the proof. We note that W_2 distance is a metric on the space of probability measures and satisfies triangle inequality.

Definition 6. *[20, W_2 **distance**] For random vectors $\mathbf{x}$ and $\mathbf{y}$ supported on $\mathcal{X}^p$ with distributions f and g respectively, the W_2 distance is given by $W_2^2(g_{\mathbf{y}}, f_{\mathbf{x}}) \triangleq \inf_\pi \sum_{t\in[p]} \Big[\mathbb{P}_\pi(\mathrm{x}_t \neq \mathrm{y}_t)\Big]^2$, where the infimum is taken over all couplings $\pi(\mathbf{x}, \mathbf{y})$ such that $\pi(\mathbf{x}) = f(\mathbf{x})$ and $\pi(\mathbf{y}) = g(\mathbf{y})$.*

Given Def. 6, our next lemma states that if appropriate W_2 distances are bounded, then the KL divergence and the entropy approximately tensorize. We provide a proof in App. E.1.1.

Lemma 15 (**Approximate tensorization of KL divergence and entropy**). *Given random vectors* $\mathbf{x}$ *and* $\mathbf{y}$ *supported on* $\mathcal{X}^p$ *with distributions* f *and* g *respectively such that* $f_{\min} > 0$ *(see (83)), if for all subsets* $S \subseteq [p]$ *(with* $S^C \triangleq [p] \setminus S$*) and all* $\boldsymbol{y}_{S^C} \in \mathcal{X}^{p-|S|}$*,*

$$W_2^2\left(g_{\mathbf{y}_S|\mathbf{y}_{S^C}=\boldsymbol{y}_{S^C}}, f_{\mathbf{x}_S|\mathbf{x}_{S^C}=\boldsymbol{y}_{S^C}}\right) \le C \sum_{t\in S} \mathbb{E}\left[\|g_{y_t|\mathbf{y}_{-t}=\boldsymbol{y}_{-t}} - f_{x_t|\mathbf{x}_{-t}=\boldsymbol{y}_{-t}}\|_{\mathsf{TV}}^2 \Big| \mathbf{y}_{S^C} = \boldsymbol{y}_{S^C}\right], \quad (85)$$

almost surely for some constant $C \ge 1$*, then*

$$\mathsf{KL}\left(g_{\mathbf{y}} \,\|\, f_{\mathbf{x}}\right) \le \frac{2C}{f_{\min}} \sum_{t\in[p]} \mathbb{E}\left[\mathsf{KL}\left(g_{y_t|\mathbf{y}_{-t}=\boldsymbol{y}_{-t}} \,\Big\|\, f_{x_t|\mathbf{x}_{-t}=\boldsymbol{y}_{-t}}\right)\right], \quad \textit{and} \quad (86)$$

$$\mathsf{Ent}_{\mathbf{x}}\left(q\right) \le \frac{2C}{f_{\min}} \sum_{t\in[p]} \mathbb{E}_{\mathbf{x}_{-t}}\left[\mathsf{Ent}_{x_t|\mathbf{x}_{-t}}\left(q\right)\right] \quad \textit{for any function } q : \mathcal{X}^p \to \mathbb{R}_+. \quad (87)$$

Next, we claim that if the random vector $\mathbf{x}$ satisfies Dobrushin's uniqueness condition, then the condition (85) of Lem. 15 is naturally satisfied. We provide a proof in App. E.1.2.

Lemma 16 (**Dobrushin's uniqueness implies approximate tensorization**). *Given random vectors* $\mathbf{x}$ *and* $\mathbf{y}$ *supported on* $\mathcal{X}^p$ *with distributions* f *and* g *respectively, if* $\mathbf{x}$ *satisfies Dobrushin's uniqueness condition (see Def. 4) with coupling matrix* $\Theta \in \mathbb{R}^{p\times p}$*, then for all subsets* $S \subseteq [p]$ *(with* $S^C \triangleq [p]\setminus S$*) and all* $\boldsymbol{y}_{S^C} \in \mathcal{X}^{p-|S|}$*,*

$$W_2^2\left(g_{\mathbf{y}_S|\mathbf{y}_{S^C}=\boldsymbol{y}_{S^C}}, f_{\mathbf{x}_S|\mathbf{x}_{S^C}=\boldsymbol{y}_{S^C}}\right) \le \frac{1}{\left(1-\|\Theta\|_{\mathrm{op}}\right)^2} \sum_{t\in S} \mathbb{E}\left[\|g_{y_t|\mathbf{y}_{-t}=\boldsymbol{y}_{-t}} - f_{x_t|\mathbf{x}_{-t}=\boldsymbol{y}_{-t}}\|_{\mathsf{TV}}^2 \Big| \mathbf{y}_{S^C} = \boldsymbol{y}_{S^C}\right], \quad (88)$$

almost surely.

Now to prove Prop. 2, applying Lem. 15 and 16 for an arbitrary function $f : \mathcal{X}^p \to \mathbb{R}$, we find that

$$\begin{aligned}
\mathsf{Ent}_{\mathbf{x}}\left(q^2\right) &\le \frac{2}{f_{\min}\left(1-\|\Theta\|_{\mathrm{op}}\right)^2} \sum_{t\in[p]} \mathbb{E}_{\mathbf{x}_{-t}}\left[\mathsf{Ent}_{x_t|\mathbf{x}_{-t}}\left(q^2\right)\right] \\
&\overset{(a)}{\le} \frac{2\sigma^2}{f_{\min}\left(1-\|\Theta\|_{\mathrm{op}}\right)^2} \sum_{t\in[p]} \mathbb{E}_{\mathbf{x}_{-t}}\left[\mathbb{E}_{x_t|\mathbf{x}_{-t}}\left[\|\nabla_{x_t} q(x_t;\mathbf{x}_{-t})\|_2^2\right]\right] \\
&\overset{(b)}{=} \frac{2\sigma^2}{f_{\min}\left(1-\|\Theta\|_{\mathrm{op}}\right)^2} \mathbb{E}_{\mathbf{x}_{-t}}\left[\mathbb{E}_{x_t|\mathbf{x}_{-t}}\left[\sum_{t\in[p]} \|\nabla_{x_t} q(x_t;\mathbf{x}_{-t})\|_2^2\right]\right] \\
&\overset{(c)}{=} \frac{2\sigma^2}{f_{\min}\left(1-\|\Theta\|_{\mathrm{op}}\right)^2} \mathbb{E}_{\mathbf{x}}\left[\|\nabla_{\mathbf{x}} q(\mathbf{x})\|_2^2\right],
\end{aligned}$$

where (a) follows because $x_t|\mathbf{x}_{-t}$ satisfies $\mathrm{LSI}_{x_t|\mathbf{x}_{-t}=\boldsymbol{x}_{-t}}(\sigma^2)$ for all $t \in [p]$ and $\boldsymbol{x}_{-t} \in \mathcal{X}^{p-1}$, (b) follows by the linearity of expectation and (b) follows by the law of total expectation. The claim follows.

E.1.1 Proof of Lem. 15: Approximate tensorization of KL divergence and entropy

We start by establishing a reverse-Pinsker style inequality for distributions with compact support to bound their KL divergence by their total variation distance. We provide a proof at the end..

Lemma 17 (**Reverse-Pinsker inequality**). *For any distributions* f *and* g *supported on* $\mathcal{X} \subset \mathbb{R}$ *such that* $\min_{x\in\mathcal{X}} f(x) > 0$*, we have* $\mathsf{KL}\left(g \,\|\, f\right) \le \frac{4}{\min_{x\in\mathcal{X}} f(x)} \|g - f\|_{\mathsf{TV}}^2$.

Given Lem. 17, we proceed to prove Lem. 15.

Proof of bound (86): To prove (86), we show that the following inequality holds using the technique of mathematical induction on p:

$$\mathsf{KL}\left(g_{\mathbf{y}} \,\|\, f_{\mathbf{x}}\right) \leq \frac{4C}{f_{\min}} \sum_{t\in[p]} \mathbb{E}\Big[\|g_{y_t|\mathbf{y}_{-t}=\boldsymbol{y}_{-t}} - f_{x_t|\mathbf{x}_{-t}=\boldsymbol{y}_{-t}}\|_{\mathsf{TV}}^2\Big]. \tag{89}$$

Then, (86) follows by using Pinsker's inequality to bound the right hand side of (89).

Base case: $p = 1$**:** For the base case, we need to establish that the claim holds for all distributions supported on $\mathcal{X}$ that satisfy the required conditions. In other words, we need to show that

$$\mathsf{KL}\left(g_{y} \,\|\, f_{x}\right) \leq \frac{4C}{f_{\min}} \|g_y - f_x\|_{\mathsf{TV}}^2 \quad \text{for every} \quad t \in [p].$$

for all random variables x and y supported on $\mathcal{X}$ such that $f_{\min} = \min_{x\in\mathcal{X}} f_{x}(x) > 0$. This follows from Lem. 17 by observing that $C \geq 1$.

Inductive step: Now, we assume that the claim holds for all distributions supported on $\mathcal{X}^{p-1}$ that satisfy the required conditions, and establish it for distributions supported on $\mathcal{X}^p$. From the chain rule of KL divergence, we have

$$\mathsf{KL}\left(g_{\mathbf{y}} \,\|\, f_{\mathbf{x}}\right) = \mathsf{KL}\left(g_{y_t} \,\|\, f_{x_t}\right) + \mathbb{E}\big[\mathsf{KL}\left(g_{\mathbf{y}_{-t}|y_t} \,\big\|\, f_{\mathbf{x}_{-t}|x_t}\right)\big] \quad \text{for every} \quad t \in [p].$$

Taking an average over all $t \in [p]$, we have

$$\mathsf{KL}\left(g_{\mathbf{y}} \,\|\, f_{\mathbf{x}}\right) = \frac{1}{p}\sum_{t\in[p]} \mathsf{KL}\left(g_{y_t} \,\|\, f_{x_t}\right) + \frac{1}{p}\sum_{t\in[p]} \mathbb{E}\big[\mathsf{KL}\left(g_{\mathbf{y}_{-t}|y_t} \,\big\|\, f_{\mathbf{x}_{-t}|x_t}\right)\big]. \tag{90}$$

Now, we bound the first term in (90). Let π^* be such that

$$\pi^* = \underset{\pi:\pi(\mathbf{x})=f(\mathbf{x}),\pi(\mathbf{y})=g(\mathbf{y})}{\arg\min} \sum_{t\in[p]} \Big[\mathbb{P}_{\pi}(x_t \neq y_t)\Big]^2. \tag{91}$$

Then, we have

$$\begin{aligned}
\frac{1}{p}\sum_{t\in[p]} \mathsf{KL}\left(g_{y_t} \,\|\, f_{x_t}\right) &\overset{(a)}{\leq} \frac{1}{p}\sum_{t\in[p]} \frac{4}{f_{\min}}\|g_{y_t} - f_{x_t}\|_{\mathsf{TV}}^2 \overset{(b)}{\leq} \frac{4}{pf_{\min}} \sum_{t\in[p]} \Big[\mathbb{P}_{\pi^*}(x_t \neq y_t)\Big]^2 \\
&\overset{(c)}{=} \frac{4}{pf_{\min}} W_2^2(g_{\mathbf{y}}, f_{\mathbf{x}}) \\
&\overset{(85)}{\leq} \frac{4C}{pf_{\min}} \sum_{t\in[p]} \mathbb{E}\Big[\|g_{y_t|\mathbf{y}_{-t}=\boldsymbol{y}_{-t}} - f_{x_t|\mathbf{x}_{-t}=\boldsymbol{y}_{-t}}\|_{\mathsf{TV}}^2\Big]
\end{aligned} \tag{92}$$

where (a) follows from Lem. 17 by observing that the marginals are lower bounded if the conditional are lower bounded, i.e., $\min_{t\in[p],x_t\in\mathcal{X}} f_{x_t}(x_t) = \min_{t\in[p],x_t\in\mathcal{X}} \int_{\boldsymbol{x}_{-t}\in\mathcal{X}^{p-1}} f_{x_t|\mathbf{x}_{-t}}(x_t|\boldsymbol{x}_{-t}) f_{\mathbf{x}_{-t}}(\boldsymbol{x}_{-t}) d\boldsymbol{x}_{-t} > f_{\min}$, (b) follows from the connections of total variation distance to optimal transportation cost, i.e., $\|g_y - f_x\|_{\mathsf{TV}} = \inf_{\pi:\pi(x)=f(x),\pi(y)=g(y)} \mathbb{P}_{\pi}(x \neq y)$, and (c) follows from Def. 6 and (91).

Next, we bound the second term in (90). We have

$$\begin{aligned}
\sum_{t\in[p]} \mathbb{E}\big[\mathsf{KL}\left(g_{\mathbf{y}_{-t}|y_t} \,\big\|\, f_{\mathbf{x}_{-t}|x_t}\right)\big] &\overset{(a)}{\leq} \sum_{t\in[p]} \mathbb{E}\Big[\frac{4C}{f_{\min}} \sum_{u\in[p]\setminus\{t\}} \mathbb{E}\Big[\|g_{y_u|\mathbf{y}_{-u}=\boldsymbol{y}_{-u}} - f_{x_u|\mathbf{x}_{-u}=\boldsymbol{y}_{-u}}\|_{\mathsf{TV}}^2 \Big| y_t = y_t\Big]\Big] \\
&\overset{(b)}{=} \frac{4C}{f_{\min}} \sum_{t\in[p]} \sum_{u\in[p]\setminus\{t\}} \mathbb{E}\Big[\|g_{y_u|\mathbf{y}_{-u}=\boldsymbol{y}_{-u}} - f_{x_u|\mathbf{x}_{-u}=\boldsymbol{y}_{-u}}\|_{\mathsf{TV}}^2\Big] \\
&= \frac{4C(p-1)}{f_{\min}} \sum_{u\in[p]} \mathbb{E}\Big[\|g_{y_u|\mathbf{y}_{-u}=\boldsymbol{y}_{-u}} - f_{x_u|\mathbf{x}_{-u}=\boldsymbol{y}_{-u}}\|_{\mathsf{TV}}^2\Big],
\end{aligned} \tag{93}$$

where (a) follows from the inductive hypothesis and (b) follows from the law of total expectation. Then, (89) follows by putting (90), (92), and (93) together.

Proof of bound (87): To prove (87), we note that (86) holds for any random vector $\mathbf{y}$ supported on $\mathcal{X}^p$. Consider $\mathbf{y}$ be such that $q(\mathbf{x})/\mathbb{E}_{\mathbf{x}}[q(\mathbf{x})]$ is the Radon-Nikodym derivative of $g_{\mathbf{y}}$ with respect to $f_{\mathbf{x}}$. Then, we have

$$\frac{dg_{\mathbf{y}_{-t}}}{df_{\mathbf{x}_{-t}}} = \frac{\mathbb{E}_{\mathbf{x}_t|\mathbf{x}_{-t}}\big[q(\mathbf{x})\big]}{\mathbb{E}_{\mathbf{x}}\big[q(\mathbf{x})\big]} \quad \text{and} \quad \frac{dg_{\mathbf{y}_t|\mathbf{y}_{-t}}}{df_{\mathbf{x}_t|\mathbf{x}_{-t}}} = \frac{q(\mathbf{x})}{\mathbb{E}_{\mathbf{x}_t|\mathbf{x}_{-t}}\big[q(\mathbf{x})\big]} \quad \text{for all} \quad t \in [p]. \tag{94}$$

We have

$$\begin{aligned}
\mathsf{KL}\left(g_{\mathbf{y}} \,\|\, f_{\mathbf{x}}\right) &\overset{(a)}{=} \mathbb{E}_{\mathbf{x}}\left[\frac{dg_{\mathbf{y}}}{df_{\mathbf{x}}} \log \frac{dg_{\mathbf{y}}}{df_{\mathbf{x}}}\right] \\
&\overset{(b)}{=} \mathbb{E}_{\mathbf{x}}\left[\frac{q(\mathbf{x})}{\mathbb{E}_{\mathbf{x}}\big[q(\mathbf{x})\big]} \log \frac{q(\mathbf{x})}{\mathbb{E}_{\mathbf{x}}\big[q(\mathbf{x})\big]}\right] \\
&= \frac{1}{\mathbb{E}_{\mathbf{x}}\big[q(\mathbf{x})\big]}\Big(\mathbb{E}_{\mathbf{x}}\big[q(\mathbf{x}) \log q(\mathbf{x})\big] - \mathbb{E}_{\mathbf{x}}\big[q(\mathbf{x})\big] \log \mathbb{E}_{\mathbf{x}}\big[q(\mathbf{x})\big]\Big) = \frac{\mathsf{Ent}_{\mathbf{x}}\left(q\right)}{\mathbb{E}_{\mathbf{x}}\big[q(\mathbf{x})\big]},
\end{aligned} \tag{95}$$

where (a) follows from the definition of KL divergence and (b) follows from the choice of $\mathbf{y}$. Similarly, for every $t \in [p]$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{y}_{-t}}\Big[\mathsf{KL}\left(g_{y_t|\mathbf{y}_{-t}=\boldsymbol{y}_{-t}} \,\big\|\, f_{x_t|\mathbf{x}_{-t}=\boldsymbol{y}_{-t}}\right)\Big] \\
&\overset{(a)}{=} \mathbb{E}_{\mathbf{y}_{-t}}\left[\mathbb{E}_{y_t|\mathbf{y}_{-t}}\left[\log \frac{dg_{y_t|\mathbf{y}_{-t}}}{df_{x_t|\mathbf{x}_{-t}}}\right]\right] \\
&\overset{(b)}{=} \mathbb{E}_{\mathbf{y}}\left[\log \frac{dg_{y_t|\mathbf{y}_{-t}}}{df_{x_t|\mathbf{x}_{-t}}}\right] \\
&\overset{(c)}{=} \mathbb{E}_{\mathbf{x}}\left[\frac{dg_{\mathbf{y}}}{df_{\mathbf{x}}} \log \frac{dg_{y_t|\mathbf{y}_{-t}}}{df_{x_t|\mathbf{x}_{-t}}}\right] \\
&\overset{(d)}{=} \mathbb{E}_{\mathbf{x}}\left[\frac{q(\mathbf{x})}{\mathbb{E}_{\mathbf{x}}\big[q(\mathbf{x})\big]} \log \frac{q(\mathbf{x})}{\mathbb{E}_{x_t|\mathbf{x}_{-t}}\big[q(\mathbf{x})\big]}\right] \\
&\overset{(e)}{=} \frac{\mathbb{E}_{\mathbf{x}_{-t}}\Big[\mathbb{E}_{x_t|\mathbf{x}_{-t}}\big[q(\mathbf{x}) \log q(\mathbf{x})\big] - \mathbb{E}_{x_t|\mathbf{x}_{-t}}\big[q(\mathbf{x}) \log \mathbb{E}_{x_t|\mathbf{x}_{-t}}\big[q(\mathbf{x})\big]\big]\Big]}{\mathbb{E}_{\mathbf{x}}\big[q(\mathbf{x})\big]} \overset{(f)}{=} \frac{\mathbb{E}_{\mathbf{x}_{-t}}\big[\mathsf{Ent}_{x_t|\mathbf{x}_{-t}}\left(q\right)\big]}{\mathbb{E}\big[q(\mathbf{x})\big]}.
\end{aligned} \tag{96}$$

where (a) follows from the definition of KL divergence, (b) follows from the law of total expectation, (c) follows from the definition of Radon-Nikodym derivative, (d) follows from the choice of $\mathbf{y}$ and (94), (e) follows from the law of total expectation, (f) follows from the definition of entropy. Then, (87) follows by putting (86), (95), and (96) together.

Proof of Lem. 17: Reverse-Pinsker inequality: Using the facts (a) $\log a \geq 1 - \frac{1}{a}$ for all $a > 0$, and (b) $\min_{x \in \mathcal{X}} f(x) > 0$, we find that

$$\log \frac{f(x)}{g(x)} \geq 1 - \frac{g(x)}{f(x)} \quad \text{for every} \quad x \in \mathcal{X}. \tag{97}$$

Multiplying both sides of (97) by $g(x) \geq 0$ and rearranging terms yields that

$$g(x) \log \frac{g(x)}{f(x)} \leq \frac{g^2(x)}{f(x)} - g(x) \quad \text{for every} \quad x \in \mathcal{X}. \tag{98}$$

Now, we have

$$\begin{aligned}
\mathsf{KL}\left(g \,\|\, f\right) = \int_{x \in \mathcal{X}} g(x) \log \frac{g(x)}{f(x)} dx &\overset{(98)}{\leq} \int_{x \in \mathcal{X}} \left(\frac{g^2(x)}{f(x)} - g(x)\right) dx \\
&\overset{(a)}{=} \int_{x \in \mathcal{X}} \frac{\big(g(x) - f(x)\big)^2}{f(x)} dx \\
&\leq \frac{1}{\min_{x \in \mathcal{X}} f(x)} \int_{x \in \mathcal{X}} \big(g(x) - f(x)\big)^2 dx
\end{aligned}$$

$$\overset{(b)}{\leq} \frac{1}{\min_{x\in\mathcal{X}} f(x)} \Big(\int_{x\in\mathcal{X}} \big|g(x) - f(x)\big| dx \Big)^2$$
$$\overset{(c)}{=} \frac{1}{\min_{x\in\mathcal{X}} f(x)} \Big(2\|g - f\|_{\mathsf{TV}} \Big)^2 = \frac{4}{\min_{x\in\mathcal{X}} f(x)} \|g - f\|_{\mathsf{TV}}^2,$$

where (a) follows by simple manipulations, (b) follows by using the order of norms on Euclidean space, and (c) follows by the definition of the total variation distance.

E.1.2 Proof of Lem. 16: Dobrushin's uniqueness implies approximate tensorization

We start by defining the notion of Gibbs sampler which is useful in the proof.

Definition 7. *[20, **Gibbs Sampler**] For a random vector $\mathbf{x}$ with distribution f, define the Markov kernels and the Gibbs sampler as follows:*

$$\Gamma_t(\boldsymbol{x}|\boldsymbol{x}') \triangleq \mathbb{1}(\boldsymbol{x}_{-t} = \boldsymbol{x}'_{-t}) f_{\mathsf{x}_t|\mathbf{x}_{-t}}(x_t|\boldsymbol{x}'_{-t}) \quad \textit{and} \quad \Gamma(\boldsymbol{x}|\boldsymbol{x}') \triangleq p^{-1} \sum_{t\in[p]} \Gamma_t(\boldsymbol{x}|\boldsymbol{x}'), \tag{99}$$

for all $t \in [p]$ and $x, x' \in \mathcal{X}^p$. That is, the kernel Γ_t leaves all but the t^{th} coordinate unchanged, and updates the t^{th} coordinate according to $f_{\mathsf{x}_t|\mathbf{x}_{-t}}$, and the sampler Γ selects an index $t \in [p]$ at random, and applies Γ_t. Further, for a random vector $\mathbf{y}$ with distribution g supported on $\mathcal{X}^p$, we also define

$$g_{\mathbf{y}}\Gamma_t(\boldsymbol{y}) \triangleq \int g_{\mathbf{y}}(\boldsymbol{y}')\Gamma_t(\boldsymbol{y}|\boldsymbol{y}')d\boldsymbol{y}' \;\; \textit{for} \;\; t \in [p], \; \textit{and} \; g_{\mathbf{y}}\Gamma(\boldsymbol{y}) \triangleq \int g_{\mathbf{y}}(\boldsymbol{y}')\Gamma(\boldsymbol{y}|\boldsymbol{y}')d\boldsymbol{y}' \quad \textit{for} \quad \boldsymbol{y} \in \mathcal{X}^p. \tag{100}$$

We now proceed to prove Lem. 16 and split it in two cases: (i) $S = [p]$, and (ii) $S \subset [p]$.

Case (i) ($S = [p]$): Let Γ be the Gibbs sampler associated with the distribution f. Then,

$$W_2\big(g_{\mathbf{y}_S|\mathbf{y}_{S^C}}, f_{\mathbf{x}_S|\mathbf{x}_{S^C}}\big) = W_2(g_{\mathbf{y}}, f_{\mathbf{x}}) \overset{(a)}{\leq} W_2(g_{\mathbf{y}}, g_{\mathbf{y}}\Gamma) + W_2(g_{\mathbf{y}}\Gamma, f_{\mathbf{x}}), \tag{101}$$

where (a) follows from the triangle inequality. We claim that

$$W_2(g_{\mathbf{y}}, g_{\mathbf{y}}\Gamma) \leq \frac{1}{p}\sqrt{\sum_{t\in[p]} \mathbb{E}_{\mathbf{y}_{-t}}\Big[\|g_{\mathsf{y}_t|\mathbf{y}_{-t}=\boldsymbol{y}_{-t}} - f_{\mathsf{x}_t|\mathbf{x}_{-t}=\boldsymbol{y}_{-t}}\|_{\mathsf{TV}}^2\Big]}, \quad \text{and} \tag{102}$$

$$W_2(g_{\mathbf{y}}\Gamma, f_{\mathbf{x}}) \leq \bigg(1 - \frac{(1 - \|\Theta\|_{\mathrm{op}})}{p}\bigg) W_2(g_{\mathbf{y}}, f_{\mathbf{x}}). \tag{103}$$

Putting (101) to (103) together, we have

$$W_2(g_{\mathbf{y}}, f_{\mathbf{x}}) \leq \frac{1}{p}\sqrt{\sum_{t\in[p]} \mathbb{E}_{\mathbf{y}_{-t}}\Big[\|g_{\mathsf{y}_t|\mathbf{y}_{-t}=\boldsymbol{y}_{-t}} - f_{\mathsf{x}_t|\mathbf{x}_{-t}=\boldsymbol{y}_{-t}}\|_{\mathsf{TV}}^2\Big]} + \bigg(1 - \frac{(1 - \|\Theta\|_{\mathrm{op}})}{p}\bigg) W_2(g_{\mathbf{y}}, f_{\mathbf{x}}). \tag{104}$$

Rearranging (104) results in (88) for $S = [p]$ as desired. It remains to prove our earlier claims (102) and (103) which we now do one-by-one.

Proof of bound (102) on $W_2(g_{\mathbf{y}}, g_{\mathbf{y}}\Gamma)$: To bound $W_2(g_{\mathbf{y}}, g_{\mathbf{y}}\Gamma)$, we construct a random vector $\mathbf{y}^\Gamma$ such that it is coupled with the random vector $\mathbf{y}$. We select an index $b \in [p]$ at random, and define

$$y_v^\Gamma \triangleq y_v \quad \text{for all} \quad v \in [p] \setminus \{b\}.$$

Then, given b and $\mathbf{y}_{-b} = \boldsymbol{y}_{-b}$, we define the joint distribution of (y_b, y_b^Γ) to be the maximal coupling of $g_{\mathsf{y}_b|\mathbf{y}_{-b}=\boldsymbol{y}_{-b}}$ and $f_{\mathsf{x}_b|\mathbf{x}_{-b}=\boldsymbol{y}_{-b}}$ that achieves $\|g_{\mathsf{y}_b|\mathbf{y}_{-b}=\boldsymbol{y}_{-b}} - f_{\mathsf{x}_b|\mathbf{x}_{-b}=\boldsymbol{y}_{-b}}\|_{\mathsf{TV}}$. It is easy to see that the marginal distribution of $\mathbf{y}$ is $g_{\mathbf{y}}$ and the marginal distribution of $\mathbf{y}^\Gamma$ is $g_{\mathbf{y}}\Gamma$ (see Def. 7). Then, we have

$$W_2^2(g_{\mathbf{y}}, g_{\mathbf{y}}\Gamma) \overset{(a)}{\leq} \sum_{t\in[p]} \Big[\mathbb{P}(b = t)\mathbb{P}(y_t \neq y_t^\Gamma | b = t) + \mathbb{P}(b \neq t)\mathbb{P}(y_t \neq y_t^\Gamma | b \neq t)\Big]^2$$

$$\overset{(b)}{=} \sum_{t\in[p]} \Big[\frac{1}{p}\mathbb{P}(\mathsf{y}_t \neq \mathsf{y}_t^{\Gamma}|b=t)\Big]^2$$

$$\overset{(c)}{=} \frac{1}{p^2}\sum_{t\in[p]} \Big[\int_{\boldsymbol{y}_{-t}\in\mathcal{X}^{p-1}} \mathbb{P}(\mathsf{y}_t \neq \mathsf{y}_t^{\Gamma}|b=t, \mathbf{y}_{-t}=\boldsymbol{y}_{-t}) g_{\mathbf{y}_{-t}|b=t}(\boldsymbol{y}_{-t}|b=t) d\boldsymbol{y}_{-t}\Big]^2$$

$$\overset{(d)}{=} \frac{1}{p^2}\sum_{t\in[p]} \Big[\int_{\boldsymbol{y}_{-t}\in\mathcal{X}^{p-1}} \|g_{\mathsf{y}_t|\mathbf{y}_{-t}=\boldsymbol{y}_{-t}} - f_{\mathsf{x}_t|\mathbf{x}_{-t}=\boldsymbol{y}_{-t}}\|_{\mathsf{TV}}\, g_{\mathbf{y}_{-t}}(\boldsymbol{y}_{-t}) d\boldsymbol{y}_{-t}\Big]^2$$

$$= \frac{1}{p^2}\sum_{t\in[p]} \Big[\mathbb{E}_{\mathbf{y}_{-t}}\Big[\|g_{\mathsf{y}_t|\mathbf{y}_{-t}=\boldsymbol{y}_{-t}} - f_{\mathsf{x}_t|\mathbf{x}_{-t}=\boldsymbol{y}_{-t}}\|_{\mathsf{TV}}\Big]\Big]^2, \tag{105}$$

where (a) follows from Def. 6 and the Bayes rule, (b) follows because $\mathbb{P}(b=t) = \frac{1}{p}$ and $\mathbb{P}(\mathsf{y}_t \neq \mathsf{y}_t^{\Gamma}|b \neq t) = 0$, (c) follows by the law of total probability, and (d) follows because $g_{\mathbf{y}_{-t}|b=t}(\boldsymbol{y}_{-t}|b=t) = g_{\mathbf{y}_{-t}}(\boldsymbol{y}_{-t})$ and by the construction of the coupling between $\mathbf{y}$ and $\mathbf{y}^{\Gamma}$. Then, (102) follows by using Jensen's inequality in (105).

Proof of bound (103) on $W_2(g_{\mathbf{y}}\Gamma, f_{\mathbf{x}})$: We first show that $f_{\mathbf{x}}$ is an invariant measure for Γ, i.e., $f_{\mathbf{x}} = f_{\mathbf{x}}\Gamma$, implying $W_2(g_{\mathbf{y}}\Gamma, f_{\mathbf{x}}) = W_2(g_{\mathbf{y}}\Gamma, f_{\mathbf{x}}\Gamma)$, and then Γ is a contraction with respect to the W_2 distance with rate $1 - \frac{(1-\|\Theta\|_{\mathrm{op}})}{p}$, i.e., $W_2(g_{\mathbf{y}}\Gamma, f_{\mathbf{x}}\Gamma) \leq \Big(1 - \frac{(1-\|\Theta\|_{\mathrm{op}})}{p}\Big)W_2(g_{\mathbf{y}}, f_{\mathbf{x}})$, implying (103).

Proof of $f_{\mathbf{x}}$ being an invariant measure for Γ: We have

$$f_{\mathbf{x}}\Gamma(\boldsymbol{x}) \overset{(100)}{=} \int_{\boldsymbol{x}'\in\mathcal{X}^p} f_{\mathbf{x}}(\boldsymbol{x}')\Gamma(\boldsymbol{x}|\boldsymbol{x}')d\boldsymbol{x}' \overset{(99)}{=} \int_{\boldsymbol{x}'\in\mathcal{X}^p} f_{\mathbf{x}}(\boldsymbol{x}')\Big(\frac{1}{p}\sum_{t\in[p]}\Gamma_t(\boldsymbol{x}|\boldsymbol{x}')\Big)d\boldsymbol{x}'$$

$$\overset{(99)}{=} \frac{1}{p}\sum_{t\in[p]}\int_{\boldsymbol{x}'\in\mathcal{X}^p} f_{\mathbf{x}}(\boldsymbol{x}')\mathbb{1}(\boldsymbol{x}_{-t} = \boldsymbol{x}'_{-t}) f_{\mathsf{x}_t|\mathbf{x}_{-t}}(x_t|\boldsymbol{x}'_{-t})d\boldsymbol{x}'$$

$$= \frac{1}{p}\sum_{t\in[p]} f_{\mathsf{x}_t|\mathbf{x}_{-t}}(x_t|\boldsymbol{x}_{-t})\int_{x'_t\in\mathcal{X}} f_{\mathbf{x}}(\boldsymbol{x}_{-t}, x'_t)dx'_t$$

$$= \frac{1}{p}\sum_{t\in[p]} f_{\mathsf{x}_t|\mathbf{x}_{-t}}(x_t|\boldsymbol{x}_{-t}) f_{\mathbf{x}_{-t}}(\boldsymbol{x}_{-t}) = f_{\mathbf{x}}(\boldsymbol{x}).$$

Proof of Γ being a contraction w.r.t the W_2 distance: Let π^* be the coupling between $\mathbf{x}$ and $\mathbf{y}$ that achieves $W_2(g_{\mathbf{y}}, f_{\mathbf{x}})$ i.e.,

$$\pi^* = \operatorname*{arg\,min}_{\pi:\pi(\mathbf{x})=f(\mathbf{x}),\pi(\mathbf{y})=g(\mathbf{y})} \sqrt{\sum_{t\in[p]}\Big[\mathbb{P}_{\pi}(\mathsf{x}_t \neq \mathsf{y}_t)\Big]^2}. \tag{106}$$

We construct random variables $\mathbf{x}'$ and $\mathbf{y}'$ as well as a coupling π' between them such that the marginal distribution of $\mathbf{x}'$ is $f_{\mathbf{x}}\Gamma$ and the marginal distribution of $\mathbf{y}'$ is $g_{\mathbf{y}}\Gamma$. We start by selecting an index $b \in [p]$ at random, and defining

$$\mathsf{y}'_v \triangleq \mathsf{y}_v \quad \text{and} \quad \mathsf{x}'_v \triangleq \mathsf{x}_v \quad \text{for all} \quad v \neq b. \tag{107}$$

Then, given b, $\mathbf{y}'_{-b} = \boldsymbol{y}_{-b}$, and $\mathbf{x}'_{-b} = \boldsymbol{x}_{-b}$, we define the joint distribution of $(\mathsf{y}'_b, \mathsf{x}'_b)$ to be the maximal coupling of $f_{\mathsf{x}_b|\mathbf{x}_{-b}}(\cdot|\boldsymbol{y}_{-b})$ and $f_{\mathsf{x}_b|\mathbf{x}_{-b}}(\cdot|\boldsymbol{x}_{-b})$ that achieves $\|f_{\mathsf{x}_b|\mathbf{x}_{-b}=\boldsymbol{y}_{-b}} - f_{\mathsf{x}_b|\mathbf{x}_{-b}=\boldsymbol{x}_{-b}}\|_{\mathsf{TV}}$.

Now, for every $t \in [p]$, we bound $\mathbb{P}_{\pi'}(\mathsf{y}'_t \neq \mathsf{x}'_t)$ in terms of $\mathbb{P}_{\pi^*}(\mathsf{y}_t \neq \mathsf{x}_t)$. To that end, we have

$$\mathbb{P}_{\pi'}(\mathsf{y}'_t \neq \mathsf{x}'_t) \overset{(a)}{=} \mathbb{P}(b=t)\mathbb{P}_{\pi'}(\mathsf{y}'_t \neq \mathsf{x}'_t|b=t) + \mathbb{P}(b\neq t)\mathbb{P}_{\pi'}(\mathsf{y}'_t \neq \mathsf{x}'_t|b\neq t)$$

$$\overset{(b)}{=} \frac{1}{p}\mathbb{P}_{\pi'}(\mathsf{y}'_t \neq \mathsf{x}'_t|b=t) + \Big(1-\frac{1}{p}\Big)\mathbb{P}_{\pi^*}(\mathsf{y}_t \neq \mathsf{x}_t), \tag{108}$$

where (a) follows from the Bayes rule and (b) follows because $\mathbb{P}(b=t)=\frac{1}{p}$ and (107). Focusing on $\mathbb{P}_{\pi'}(y'_t \neq x'_t|b=t)$ and using the law of total probability, we have

$$\begin{aligned}
&\mathbb{P}_{\pi'}(y'_t \neq x'_t|b=t) \\
&= \int\limits_{\boldsymbol{y}_{-t},\boldsymbol{x}_{-t}\in\mathcal{X}^{p-1}} \mathbb{P}_{\pi'}(y'_t \neq x'_t|b=t, \mathbf{y}'_{-t}=\boldsymbol{y}_{-t}, \mathbf{x}'_{-t}=\boldsymbol{x}_{-t})\pi'_{\mathbf{y}'_{-t},\mathbf{x}'_{-t}|b=t}(\boldsymbol{y}_{-t},\boldsymbol{x}_{-t}|b=t)d\boldsymbol{y}_{-t}d\boldsymbol{x}_{-t} \\
&\stackrel{(a)}{=} \int\limits_{\boldsymbol{y}_{-t},\boldsymbol{x}_{-t}\in\mathcal{X}^{p-1}} \|f_{x_t|\mathbf{x}_{-t}=\boldsymbol{y}_{-t}} - f_{x_t|\mathbf{x}_{-t}=\boldsymbol{x}_{-t}}\|_{\mathsf{TV}}\pi^*_{\mathbf{y}_{-t},\mathbf{x}_{-t}}(\boldsymbol{y}_{-t},\boldsymbol{x}_{-t})d\boldsymbol{y}_{-t}d\boldsymbol{x}_{-t} \\
&= \mathbb{E}_{\pi^*_{\mathbf{y}_{-t},\mathbf{x}_{-t}}}\Big[\|f_{x_t|\mathbf{x}_{-t}=\boldsymbol{y}_{-t}} - f_{x_t|\mathbf{x}_{-t}=\boldsymbol{x}_{-t}}\|_{\mathsf{TV}}\Big] \\
&\stackrel{(b)}{\leq} \mathbb{E}_{\pi^*_{\mathbf{y}_{-t},\mathbf{x}_{-t}}}\Big[\sum_{u\in[p]\setminus\{t\}} \mathbb{1}(r_u = x_u \neq y_u = s_u)\mathbb{1}(r_v = s_v = x_v \forall v < u) \\
&\qquad\qquad\qquad \times \mathbb{1}(r_v = s_v = y_v \forall v > u)\|f_{x_t|\mathbf{x}_{-t}=\boldsymbol{r}_{-t}} - f_{x_t|\mathbf{x}_{-t}=\boldsymbol{s}_{-t}}\|_{\mathsf{TV}}\Big] \\
&\stackrel{(82)}{\leq} \mathbb{E}_{\pi^*_{\mathbf{y}_{-t},\mathbf{x}_{-t}}}\Big[\sum_{u\in[p]\setminus\{t\}} \Theta_{tu}\mathbb{1}(y_u \neq x_u)\Big] = \sum_{u\in[p]\setminus\{t\}} \Theta_{tu}\mathbb{P}_{\pi^*}(y_u \neq x_u),
\end{aligned} \tag{109}$$

where (a) follows by the construction of the coupling between $\mathbf{y}'$ and $\mathbf{x}'$, and (b) follows by triangle inequality. Putting together (108) and (109), we have

$$\mathbb{P}_{\pi'}(y'_t \neq x'_t) \leq \frac{1}{p}\sum_{u\in[p]\setminus\{t\}} \Theta_{tu}\mathbb{P}_{\pi^*}(y_u \neq x_u) + \Big(1-\frac{1}{p}\Big)\mathbb{P}_{\pi^*}(y_t \neq x_t). \tag{110}$$

Next, we use (110) to show contraction of Γ. We have

$$\begin{aligned}
W_2^2(g_{\mathbf{y}}\Gamma, f_{\mathbf{x}}\Gamma) &\stackrel{(a)}{\leq} \sum_{t\in[p]} \Big[\mathbb{P}_{\pi'}(y'_t \neq x'_t)\Big]^2 \\
&\stackrel{(110)}{\leq} \sum_{t\in[p]} \bigg[\frac{1}{p}\sum_{j\in[p]\setminus\{i\}} \Theta_{tu}\mathbb{P}_{\pi^*}(y_j \neq x_j) + \Big(1-\frac{1}{p}\Big)\mathbb{P}_{\pi^*}(y_t \neq x_t)\bigg]^2 \\
&\stackrel{(b)}{\leq} \bigg\|\Big(1-\frac{1}{p}\Big)I + \frac{1}{p}\Theta\bigg\|_{\mathrm{op}}^2 \sum_{t\in[p]} \Big[\mathbb{P}_{\pi^*}(y_t \neq x_t)\Big]^2 \\
&\stackrel{(c)}{=} \bigg\|\Big(1-\frac{1}{p}\Big)I + \frac{1}{p}\Theta\bigg\|_{\mathrm{op}}^2 W_2^2(g_{\mathbf{y}}, f_{\mathbf{x}}) \\
&\stackrel{(d)}{\leq} \bigg(\Big(1-\frac{1}{p}\Big) + \frac{1}{p}\|\Theta\|_{\mathrm{op}}\bigg)^2 W_2^2(g_{\mathbf{y}}, f_{\mathbf{x}}),
\end{aligned} \tag{111}$$

where (a) follows from Def. 6, (b) follows by some linear algebraic manipulations, (c) follows from Def. 6 and (106), and (d) follows from the triangle inequality. Then, contraction of Γ follows by taking square root on both sides of (111).

Case (ii) ($S \subset [p]$)**:** We can directly verify that the matrix $\Theta_S \triangleq \{\Theta_{tu}\}_{t,u\in S}$ is such that $\|\Theta_S\|_{\mathrm{op}} \leq \|\Theta\|_{\mathrm{op}}$. Further, we note that for any $\boldsymbol{y}_{S^C} \in \mathcal{X}^{p-|S|}$, the random vector $\mathbf{x}_S|\mathbf{x}_{S^C} = \boldsymbol{y}_{S^C}$ with distribution $f_{\mathbf{x}_S|\mathbf{x}_{S^C}=\boldsymbol{y}_{S^C}}$ satisfies the Dobrushin's uniqueness condition (Def. 4) with coupling matrix Θ_S. Then, by performing an analysis similar to the one above, we have

$$\begin{aligned}
W_2(g_{\mathbf{y}_S|\mathbf{y}_{S^C}}, f_{\mathbf{x}_S|\mathbf{x}_{S^C}}) &\leq \frac{1}{(1-\|\Theta_S\|_{\mathrm{op}})}\sqrt{\sum_{t\in S}\mathbb{E}\Big[\|g_{y_t|\mathbf{y}_{-t}=\boldsymbol{y}_{-t}} - f_{x_t|\mathbf{x}_{-t}=\boldsymbol{y}_{-t}}\|_{\mathsf{TV}}^2\Big|\mathbf{y}_{S^C} = \boldsymbol{y}_{S^C}\Big]} \\
&\stackrel{(a)}{\leq} \frac{1}{(1-\|\Theta\|_{\mathrm{op}})}\sqrt{\sum_{t\in S}\mathbb{E}\Big[\|g_{y_t|\mathbf{y}_{-t}=\boldsymbol{y}_{-t}} - f_{x_t|\mathbf{x}_{-t}=\boldsymbol{y}_{-t}}\|_{\mathsf{TV}}^2\Big|\mathbf{y}_{S^C} = \boldsymbol{y}_{S^C}\Big]}
\end{aligned}$$

where (a) follows because $\frac{1}{(1-\|\Theta_S\|_{\mathrm{op}})} \leq \frac{1}{(1-\|\Theta\|_{\mathrm{op}})}$. This completes the proof.

E.2 Proof of Prop. 3: Tail bounds for arbitrary functions under LSI

Fix a function $q : \mathcal{X}^p \to \mathbb{R}$. Fix any pseudo derivative $\widetilde{\nabla} q$ for q and any pseudo Hessian $\widetilde{\nabla}^2 q$ for q. To prove Prop. 3, we bound the p-th moment of $q(\mathbf{x}) - \mathbb{E}\big[q(\mathbf{x})\big]$ by certain norms of $\widetilde{\nabla}^2 q$ and $\mathbb{E}_{\mathbf{x}}\big[\widetilde{\nabla} q(\mathbf{x})\big]$. To that end, first, we claim that in order to control the p-th moment of $q(\mathbf{x}) - \mathbb{E}\big[q(\mathbf{x})\big]$, it is sufficient to control the p-th moment of $\|\nabla q(\mathbf{x})\|_2$. Then, using (84), we note that the p-th moment of $\|\nabla q(\mathbf{x})\|_2$ is bounded by the p-th moment of $\|\widetilde{\nabla} q(\mathbf{x})\|_2$. Next, we claim that the p-th moment of $\|\widetilde{\nabla} q(\mathbf{x})\|_2$ is bounded by a linear combination of appropriate norms of $\widetilde{\nabla}^2 q$ and $\mathbb{E}_{\mathbf{x}}\big[\widetilde{\nabla} q(\mathbf{x})\big]$. We formalize the claims below and divide the proof across App. E.2.1 and App. E.2.2.

Lemma 18 (Bounded p-th moments of $q(\mathbf{x}) - \mathbb{E}\big[q(\mathbf{x})\big]$ and $\|\widetilde{\nabla} q(\mathbf{x})\|_2$). *If a random vector $\mathbf{x}$ satisfies $\mathrm{LSI}_{\mathbf{x}}(\sigma^2)$, then for any arbitrary function $q : \mathcal{X}^p \to \mathbb{R}$,*

$$\big\|q(\mathbf{x}) - \mathbb{E}\big[q(\mathbf{x})\big]\big\|_{L_p} \le \sigma\sqrt{2p}\,\big\|\|\nabla q(\mathbf{x})\|_2\big\|_{L_p} \quad \textit{for any } p \ge 2 \quad . \tag{112}$$

Further, for any pseudo derivative $\widetilde{\nabla} q(\boldsymbol{x})$ and any pseudo Hessian $\widetilde{\nabla}^2 q(\boldsymbol{x})$ for q, and even $p \ge 2$,

$$\big\|\|\widetilde{\nabla} q(\mathbf{x})\|_2\big\|_{L_p} \le 2c\sigma\Big(\max_{\boldsymbol{x}\in\mathcal{X}^p} \|\widetilde{\nabla}^2 q(\boldsymbol{x})\|_{\mathrm{F}} + \sqrt{p}\max_{\boldsymbol{x}\in\mathcal{X}^p} \|\widetilde{\nabla}^2 q(\boldsymbol{x})\|_{\mathrm{op}}\Big) + 4\|\mathbb{E}_{\mathbf{x}}\big[\widetilde{\nabla} q(\mathbf{x})\big]\|_2 \tag{113}$$

where $c \ge 0$ is a universal constant.

Given these lemmas, we proceed to prove Prop. 3. Combining (112) and (113) for any even $p \ge 2$, there exists a universal constant c' such that

$$\big\|q(\mathbf{x}) - \mathbb{E}\big[q(\mathbf{x})\big]\big\|_{L_p} \le c'\sigma^2\Big(\sqrt{p}\max_{\boldsymbol{x}\in\mathcal{X}^p} \|\widetilde{\nabla}^2 q(\boldsymbol{x})\|_{\mathrm{F}} + p\max_{\boldsymbol{x}\in\mathcal{X}^p} \|\widetilde{\nabla}^2 q(\boldsymbol{x})\|_{\mathrm{op}} + \sqrt{p}\|\mathbb{E}_{\mathbf{x}}\big[\widetilde{\nabla} q(\mathbf{x})\big]\|_2\Big). \tag{114}$$

Now, we complete the proof by using (114) along with Markov's inequality for a specific choice of p. For any even $p \ge 2$, we have

$$\begin{aligned}
&\mathbb{P}\Big[\big|q(\mathbf{x}) - \mathbb{E}\big[q(\mathbf{x})\big]\big| > ec'\sigma^2\Big(\sqrt{p}\max_{\boldsymbol{x}\in\mathcal{X}^p} \|\widetilde{\nabla}^2 q(\boldsymbol{x})\|_{\mathrm{F}} + p\max_{\boldsymbol{x}\in\mathcal{X}^p} \|\widetilde{\nabla}^2 q(\boldsymbol{x})\|_{\mathrm{op}} + \sqrt{p}\|\mathbb{E}_{\mathbf{x}}\big[\widetilde{\nabla} q(\mathbf{x})\big]\|_2\Big)\Big] \\
&= \mathbb{P}\Big[\big|q(\mathbf{x}) - \mathbb{E}\big[q(\mathbf{x})\big]\big|^p > \big(ec'\sigma^2\big)^p\Big(\sqrt{p}\max_{\boldsymbol{x}\in\mathcal{X}^p} \|\widetilde{\nabla}^2 q(\boldsymbol{x})\|_{\mathrm{F}} + p\max_{\boldsymbol{x}\in\mathcal{X}^p} \|\widetilde{\nabla}^2 q(\boldsymbol{x})\|_{\mathrm{op}} + \sqrt{p}\|\mathbb{E}_{\mathbf{x}}\big[\widetilde{\nabla} q(\mathbf{x})\big]\|_2\Big)^p\Big] \\
&\overset{(a)}{\le} \frac{\mathbb{E}\Big[\big|q(\mathbf{x}) - \mathbb{E}\big[q(\mathbf{x})\big]\big|^p\Big]}{\big(ec'\sigma^2\big)^p\Big(\sqrt{p}\max_{\boldsymbol{x}\in\mathcal{X}^p} \|\widetilde{\nabla}^2 q(\boldsymbol{x})\|_{\mathrm{F}} + p\max_{\boldsymbol{x}\in\mathcal{X}^p} \|\widetilde{\nabla}^2 q(\boldsymbol{x})\|_{\mathrm{op}} + \sqrt{p}\|\mathbb{E}_{\mathbf{x}}\big[\widetilde{\nabla} q(\mathbf{x})\big]\|_2\Big)^p} \\
&\overset{(114)}{\le} e^{-p},
\end{aligned}$$

where (a) follows from Markov's inequality. The proof is complete by choosing an appropriate universal constant c'', and and performing basic algebraic manipulations after letting

$$p = \frac{1}{c''\sigma^2}\min\Big(\frac{\varepsilon^2}{\mathbb{E}\big[\|\widetilde{\nabla} q(\mathbf{x})\|_2\big]^2 + \max_{\boldsymbol{x}\in\mathcal{X}^p} \|\widetilde{\nabla}^2 q(\boldsymbol{x})\|_{\mathrm{F}}^2}, \frac{\varepsilon}{\max_{\boldsymbol{x}\in\mathcal{X}^p} \|\widetilde{\nabla}^2 q(\boldsymbol{x})\|_{\mathrm{op}}}\Big).$$

We note that a even $p \ge 2$ can be ensured by choosing appropriate c''.

E.2.1 Proof of Lem. 18 (112): Bounded p-th moment of $q(\mathbf{x}) - \mathbb{E}\big[q(\mathbf{x})\big]$

Fix any $p \ge 2$. We start by using the following result from [4, Theorem 3.4] since $\mathbf{x}$ satisfies $\mathrm{LSI}_{\mathbf{x}}(\sigma^2)$:

$$\big\|q(\mathbf{x}) - \mathbb{E}\big[q(\mathbf{x})\big]\big\|_{L_p}^2 \le \big\|q(\mathbf{x}) - \mathbb{E}\big[q(\mathbf{x})\big]\big\|_{L_2}^2 + 2\sigma^2(p-2)\,\big\|\|\nabla q(\mathbf{x})\|_2\big\|_{L_p}^2. \tag{115}$$

Then, we bound the first term in (115) by using the fact that logarithmic Sobolev inequality implies Poincare inequality with the same constant:

$$\big\|q(\mathbf{x}) - \mathbb{E}\big[q(\mathbf{x})\big]\big\|_{L_2}^2 = \mathbb{V}\mathrm{ar}(q(\mathbf{x})) \le \sigma^2\mathbb{E}_{\mathbf{x}}\Big[\|\nabla q(\mathbf{x})\|_2^2\Big]. \tag{116}$$

Putting together (115) and (116), we have

$$
\begin{aligned}
\left\|q(\mathbf{x})-\mathbb{E}\left[q(\mathbf{x})\right]\right\|_{L_p}^2 &\leq \sigma^2\mathbb{E}_{\mathbf{x}}\Big[\,\|\nabla q(\mathbf{x})\|_2^2\,\Big]+2\sigma^2(p-2)\,\|\|\nabla q(\mathbf{x})\|_2\|_{L_p}^2\\
&\overset{(a)}{\leq} \sigma^2\Big(\mathbb{E}_{\mathbf{x}}\Big[\,\|\nabla q(\mathbf{x})\|_2^p\,\Big]\Big)^{2/p}+2\sigma^2(p-2)\,\|\|\nabla q(\mathbf{x})\|_2\|_{L_p}^2\\
&\overset{(b)}{=} \sigma^2\,\|\|\nabla q(\mathbf{x})\|_2\|_{L_p}^2+2\sigma^2(p-2)\,\|\|\nabla q(\mathbf{x})\|_2\|_{L_p}^2 \leq 2\sigma^2 p\,\|\|\nabla q(\mathbf{x})\|_2\|_{L_p}^2\,,
\end{aligned}
\tag{117}
$$

where (a) follows by Jensen's inequality and (b) follows by the definition of p-th moment. Taking square root on both sides of (117) completes the proof.

E.2.2 Proof of Lem. 18 (113): Bounded p-th moment of $\|\widetilde{\nabla} q(\mathbf{x})\|_2$

Fix any even $p \geq 2$. Fix any pseudo derivative $\widetilde{\nabla} q$ and any pseudo Hessian $\widetilde{\nabla}^2 q$. We start by obtaining a convenient bound on $\|\widetilde{\nabla} q(\boldsymbol{x})\|_2$ for every $\boldsymbol{x}\in\mathcal{X}^p$ and then proceed to bound the p-th moment of $\|\widetilde{\nabla} q(\mathbf{x})\|_2$.

Consider a p-dimensional standard normal random vector $\mathbf{g}$ independent of $\mathbf{x}$. For a given $\mathbf{x}=\boldsymbol{x}\in\mathcal{X}^p$, the random variable $\dfrac{\widetilde{\nabla} q(\boldsymbol{x})^\top\mathbf{g}}{\|\widetilde{\nabla} q(\boldsymbol{x})\|_2}$ is a standard normal random variable. Then, for every $\boldsymbol{x}\in\mathcal{X}^p$, we have

$$
\left\|\frac{\widetilde{\nabla} q(\boldsymbol{x})^\top\mathbf{g}}{\|\widetilde{\nabla} q(\boldsymbol{x})\|_2}\right\|_{L_p} \overset{(a)}{=} \left(\mathbb{E}_{\mathbf{g}|\mathbf{x}=\boldsymbol{x}}\left[\left(\frac{\widetilde{\nabla} q(\boldsymbol{x})^\top\mathbf{g}}{\|\widetilde{\nabla} q(\boldsymbol{x})\|_2}\right)^p\right]\right)^{1/p} \overset{(b)}{\geq} \frac{\sqrt{p}}{2}. \tag{118}
$$

where (a) follows from the definition of p-th moment, and (b) follows since $\|g\|_{L_p}\geq\frac{\sqrt{p}}{2}$ for any standard normal random variable g and even $p\geq 2$. Rearranging (118), we have

$$
\|\widetilde{\nabla} q(\boldsymbol{x})\|_2 \leq \frac{2}{\sqrt{p}}\Big(\mathbb{E}_{\mathbf{g}|\mathbf{x}=\boldsymbol{x}}\Big[\big(\widetilde{\nabla} q(\boldsymbol{x})^\top\mathbf{g}\big)^p\Big]\Big)^{1/p}. \tag{119}
$$

Now, we proceed to bound the p-th moment of $\|\widetilde{\nabla} q(\mathbf{x})\|_2$ as follows:

$$
\begin{aligned}
\|\|\widetilde{\nabla} q(\mathbf{x})\|_2\|_{L_p} &\overset{(a)}{=} \Big(\mathbb{E}_{\mathbf{x}}\big[\|\widetilde{\nabla} q(\mathbf{x})\|_2^p\big]\Big)^{1/p}\\
&\overset{(119)}{\leq} \frac{2}{\sqrt{p}}\Big(\mathbb{E}_{\mathbf{x},\mathbf{g}}\Big[\big(\widetilde{\nabla} q(\mathbf{x})^\top\mathbf{g}\big)^p\Big]\Big)^{1/p}\\
&\overset{(b)}{=} \frac{2}{\sqrt{p}}\left\|\widetilde{\nabla} q(\mathbf{x})^\top\mathbf{g}\right\|_{L_p}\\
&\overset{(c)}{\leq} \frac{2}{\sqrt{p}}\Big(\left\|\widetilde{\nabla} q(\mathbf{x})^\top\mathbf{g}-\mathbb{E}_{\mathbf{x}}\big[\widetilde{\nabla} q(\mathbf{x})^\top\mathbf{g}\big]\right\|_{L_p}+\left\|\mathbb{E}_{\mathbf{x}}\big[\widetilde{\nabla} q(\mathbf{x})^\top\mathbf{g}\big]\right\|_{L_p}\Big)
\end{aligned}
\tag{120}
$$

where (a) and (b) follow from the definition of p-th moment and (c) follows by Minkowski's inequality. We claim that

$$
\left\|\widetilde{\nabla} q(\mathbf{x})^\top\mathbf{g}-\mathbb{E}_{\mathbf{x}}\big[\widetilde{\nabla} q(\mathbf{x})^\top\mathbf{g}\big]\right\|_{L_p} \leq c\sigma\Big(\sqrt{p}\max_{\boldsymbol{x}\in\mathcal{X}^p}\|\widetilde{\nabla}^2 q(\boldsymbol{x})\|_{\mathrm{F}}+p\max_{\boldsymbol{x}\in\mathcal{X}^p}\|\widetilde{\nabla}^2 q(\boldsymbol{x})\|_{\mathrm{op}}\Big),\ \text{and} \tag{121}
$$

$$
\left\|\mathbb{E}_{\mathbf{x}}\big[\widetilde{\nabla} q(\mathbf{x})^\top\mathbf{g}\big]\right\|_{L_p} \leq 2\sqrt{p}\left\|\mathbb{E}_{\mathbf{x}}\big[\widetilde{\nabla} q(\mathbf{x})\big]\right\|_2, \tag{122}
$$

where $c\geq 0$ is a universal constant. Putting together (120) to (122) completes the proof. It remains to prove our claims (121) and (122) which we now do one-by-one.

Proof of bound (121): We start by obtaining a bound on $\big(\mathbb{E}_{\mathbf{x}|\mathbf{g}=\boldsymbol{g}}\big[\big(\widetilde{\nabla} q(\mathbf{x})^\top\boldsymbol{g}-\mathbb{E}_{\mathbf{x}|\mathbf{g}=\boldsymbol{g}}\big[\widetilde{\nabla} q(\mathbf{x})^\top\boldsymbol{g}\big]\big)^p\big]\big)^{1/p}$ for every $\mathbf{g}=\boldsymbol{g}$. Then, we bound $\|\widetilde{\nabla} q(\mathbf{x})^\top\mathbf{g}-\mathbb{E}_{\mathbf{x}}\big[\widetilde{\nabla} q(\mathbf{x})^\top\mathbf{g}\big]\|_{L_p}$.

To that end, we define $h_{\boldsymbol{g}}(\mathbf{x}) \triangleq \widetilde{\nabla} q(\mathbf{x})^\top \boldsymbol{g} - \mathbb{E}_{\mathbf{x}|\mathbf{g}=\boldsymbol{g}}\big[\widetilde{\nabla} q(\mathbf{x})^\top \boldsymbol{g}\big]$ and observe that $\mathbb{E}_{\mathbf{x}|\mathbf{g}=\boldsymbol{g}}\big[h_{\boldsymbol{g}}(\mathbf{x})\big] = 0$. Now, applying Lem. 18 (112) to $h_{\boldsymbol{g}}(\cdot)$, we have

$$\|h_{\boldsymbol{g}}(\mathbf{x})\|_{L_p} \le \sigma\sqrt{2p}\Big(\mathbb{E}_{\mathbf{x}|\mathbf{g}=\boldsymbol{g}}\Big[\,\|\nabla h_{\boldsymbol{g}}(\mathbf{x})\|_2^p\Big]\Big)^{1/p} \overset{(a)}{\le} \sigma\sqrt{2p}\Big(\mathbb{E}_{\mathbf{x}|\mathbf{g}=\boldsymbol{g}}\Big[\Big\|\nabla\big[\boldsymbol{g}^\top\widetilde{\nabla} q(\mathbf{x})\big]\Big\|_2^p\Big]\Big)^{1/p}$$
$$\overset{(84)}{\le} \sigma\sqrt{2p}\Big(\mathbb{E}_{\mathbf{x}|\mathbf{g}=\boldsymbol{g}}\Big[\Big\|\boldsymbol{g}^\top\widetilde{\nabla}^2 q(\mathbf{x})\Big\|_2^p\Big]\Big)^{1/p}, \quad (123)$$

where (a) follows from the definition of $h_{\boldsymbol{g}}(\mathbf{x})$. Now, to obtain a bound on the RHS of (123), we further fix $\mathbf{x} = \boldsymbol{x}$. Then, we let $\mathbf{g}'$ be another p-dimensional standard normal vector and apply an inequality similar to (119) to $\boldsymbol{g}^\top\widetilde{\nabla}^2 q(\boldsymbol{x})$ obtaining

$$\Big\|\boldsymbol{g}^\top\widetilde{\nabla}^2 q(\boldsymbol{x})\Big\|_2 \le \frac{2}{\sqrt{p}}\Big(\mathbb{E}_{\mathbf{g}'|\mathbf{x}=\boldsymbol{x},\mathbf{g}=\boldsymbol{g}}\Big[\Big(\boldsymbol{g}^\top\widetilde{\nabla}^2 q(\boldsymbol{x})\mathbf{g}'\Big)^p\Big]\Big)^{1/p},$$

which implies

$$\Big(\mathbb{E}_{\mathbf{x}|\mathbf{g}=\boldsymbol{g}}\Big[\Big\|\boldsymbol{g}^\top\widetilde{\nabla}^2 q(\mathbf{x})\Big\|_2^p\Big]\Big)^{1/p} \le \frac{2}{\sqrt{p}}\Big(\mathbb{E}_{\mathbf{x},\mathbf{g}'|\mathbf{g}=\boldsymbol{g}}\Big[\Big(\nabla\boldsymbol{g}^\top\widetilde{\nabla}^2 q(\mathbf{x})\mathbf{g}'\Big)^p\Big]\Big)^{1/p}. \quad (124)$$

Putting together (123) and (124), and using the definition of $h_{\boldsymbol{g}}(\mathbf{x})$, we have

$$\Big(\mathbb{E}_{\mathbf{x}|\mathbf{g}=\boldsymbol{g}}\Big[\Big(\widetilde{\nabla} q(\mathbf{x})^\top\boldsymbol{g} - \mathbb{E}_{\mathbf{x}|\mathbf{g}=\boldsymbol{g}}\Big[\widetilde{\nabla} q(\mathbf{x})^\top\boldsymbol{g}\Big]\Big)^p\Big]\Big)^{1/p} \le 2\sqrt{2}\sigma\Big(\mathbb{E}_{\mathbf{x},\mathbf{g}'|\mathbf{g}=\boldsymbol{g}}\Big[\Big(\boldsymbol{g}^\top\widetilde{\nabla}^2 q(\mathbf{x})\mathbf{g}'\Big)^p\Big]\Big)^{1/p}. \quad (125)$$

Now, we proceed to bound $\|\widetilde{\nabla} q(\mathbf{x})^\top\mathbf{g} - \mathbb{E}_{\mathbf{x}}\big[\widetilde{\nabla} q(\mathbf{x})^\top\mathbf{g}\big]\|_{L_p}$ as follows:

$$\Big\|\widetilde{\nabla} q(\mathbf{x})^\top\mathbf{g} - \mathbb{E}_{\mathbf{x}}\Big[\widetilde{\nabla} q(\mathbf{x})^\top\mathbf{g}\Big]\Big\|_{L_p} \overset{(a)}{=} \Big(\mathbb{E}_{\mathbf{x},\mathbf{g}}\Big[\Big(\widetilde{\nabla} q(\mathbf{x})^\top\mathbf{g} - \mathbb{E}_{\mathbf{x}}\Big[\widetilde{\nabla} q(\mathbf{x})^\top\mathbf{g}\Big]\Big)^p\Big]\Big)^{1/p}$$
$$\overset{(125)}{\le} 2\sqrt{2}\sigma\Big(\mathbb{E}_{\mathbf{g},\mathbf{x},\mathbf{g}'}\Big[\Big(\mathbf{g}^\top\widetilde{\nabla}^2 q(\mathbf{x})\mathbf{g}'\Big)^p\Big]\Big)^{1/p}, \quad (126)$$

where (a) follows from the definition of p-th moment. Finally, to bound the RHS of (126), we fix $\mathbf{x} = \boldsymbol{x}$ and bound the p-th norm of the quadratic form $\mathbf{g}^\top\widetilde{\nabla}^2 q(\boldsymbol{x})\mathbf{g}'$ by the Hanson-Wright inequality resulting in

$$\Big(\mathbb{E}_{\boldsymbol{g},\mathbf{g}'|\mathbf{x}=\boldsymbol{x}}\Big[\Big(\mathbf{g}^\top\widetilde{\nabla}^2 q(\boldsymbol{x})\mathbf{g}'\Big)^p\Big]\Big)^{1/p} \le c\Big(\sqrt{p}\|\widetilde{\nabla}^2 q(\boldsymbol{x})\|_{\mathrm{F}} + p\|\widetilde{\nabla}^2 q(\boldsymbol{x})\|_{\mathrm{op}}\Big)$$
$$\le c\Big(\sqrt{p}\max_{\boldsymbol{x}\in\mathcal{X}^p}\|\widetilde{\nabla}^2 q(\boldsymbol{x})\|_{\mathrm{F}} + p\max_{\boldsymbol{x}\in\mathcal{X}^p}\|\widetilde{\nabla}^2 q(\boldsymbol{x})\|_{\mathrm{op}}\Big), \quad (127)$$

where $c \ge 0$ is a universal constant. Then, (121) follows by putting together (126) and (127).

Proof of bound (122): By linearity of expectation, we have

$$\|\mathbb{E}_{\mathbf{x}}\big[\widetilde{\nabla} q(\mathbf{x})^\top\mathbf{g}\big]\|_{L_p} = \|\big(\mathbb{E}_{\mathbf{x}}\big[\widetilde{\nabla} q(\mathbf{x})\big]\big)^\top\mathbf{g}\|_{L_p}. \quad (128)$$

We note that the random variable $\dfrac{(\mathbb{E}_{\mathbf{x}}[\widetilde{\nabla} q(\mathbf{x})])^\top\mathbf{g}}{\|\mathbb{E}_{\mathbf{x}}[\widetilde{\nabla} q(\mathbf{x})]\|_2}$ is a standard normal random variable. Therefore,

$$\Bigg\|\frac{\big(\mathbb{E}_{\mathbf{x}}\big[\widetilde{\nabla} q(\mathbf{x})\big]\big)^\top\mathbf{g}}{\|\mathbb{E}_{\mathbf{x}}\big[\widetilde{\nabla} q(\mathbf{x})\big]\|_2}\Bigg\|_{L_p} \overset{(a)}{=} \Bigg(\mathbb{E}_{\mathbf{g}}\Bigg[\Bigg(\frac{\big(\mathbb{E}_{\mathbf{x}}\big[\widetilde{\nabla} q(\mathbf{x})\big]\big)^\top\mathbf{g}}{\|\mathbb{E}_{\mathbf{x}}\big[\widetilde{\nabla} q(\mathbf{x})\big]\|_2}\Bigg)^p\Bigg]\Bigg)^{1/p} \overset{(b)}{\le} 2\sqrt{p}, \quad (129)$$

where (a) follows from the definition of p-th moment, and (b) follows since $\|g\|_{L_p} \le 2\sqrt{p}$ for any standard normal variable g. Then, (122) follows by using (129) in (128).

F Identifying weakly dependent random variables

In App. E, we derived (in Prop. 2) that a random vector (supported on a compact set) satisfies the logarithmic Sobolev inequality if it satisfies the Dobrushin's uniqueness condition (in Def. 4). Further, we also derived (Prop. 3) tail bounds for a random vector satisfying the logarithmic Sobolev inequality. Combining the two, we see that in order to use the tail bound, the random vector needs to satisfy the Dobrushin's uniqueness condition, i.e, the elements of the random vector should be weakly dependent. In this section, we show that any random vector (outside Dobrushin's regime) that is a τ-Sparse Graphical Model (to be defined) can be reduced to satisfy the Dobrushin's uniqueness condition. In particular, we show that by conditioning on a subset of the random vector, the unconditioned subset of the random vector (in the conditional distribution) are only weakly dependent. We exploit this trick in Lem. 12 and Lem. 14 to enable application of the tail bound in App. E. The result below is a generalization of the result in [12] for discrete random vectors to continuous random vectors.

We start by defining the notion of τ-Sparse Graphical Model.

Definition 8 (**τ-Sparse Graphical Model**). *A pair of random vectors $\{\mathbf{x}, \mathbf{z}\}$ supported on $\mathcal{X}^p \times \mathcal{Z}^{p_z}$ is a τ-Sparse Graphical Model for model-parameters $\tau \triangleq (\alpha, \zeta, x_{\max}, \Theta)$ and denoted by τ-SGM if $\mathcal{X} = \{-x_{\max}, x_{\max}\}$, and*

1. *for any realization $\boldsymbol{z} \in \mathcal{Z}^{p_z}$, the conditional probability distribution of $\mathbf{x}$ given $\mathbf{z} = \boldsymbol{z}$ is given by $f_{\mathbf{x}|\mathbf{z}}\big(\cdot|\boldsymbol{z}; \theta(\boldsymbol{z}), \Theta\big)$ in (4) for a vector $\theta(\boldsymbol{z}) \in \mathbb{R}^p$ depending on $\boldsymbol{z}$ and a symmetric matrix $\Theta \in \mathbb{R}^{p\times p}$ (independent of $\boldsymbol{z}$) with $\Theta_{tt} = 0$ for all $t \in [p]$,*
2. $\max\{\max_{\boldsymbol{z}\in\mathcal{Z}^{p_z}} \|\theta(\boldsymbol{z})\|_\infty, \|\Theta\|_{\max}\} \le \alpha$, *and*
3. $\|\Theta\|_\infty \le \zeta$.

Now, we provide the main result of this section.

Proposition 4 (**Identifying weakly dependent random variables**). *Given a pair of random vectors $\{\mathbf{x}, \mathbf{z}\}$ supported on $\mathcal{X}^p \times \mathcal{Z}^{p_z}$ that is a τ-SGM (Def. 8) with $\tau \triangleq (\alpha, \zeta, x_{\max}, \Theta)$, and a scalar $\lambda \in (0, \zeta]$, there exists $L \triangleq 32\zeta^2 \log 4p/\lambda^2$ subsets $S_1, \cdots, S_L \subseteq [p]$ that satisfy the following properties:*

(a) For any $t \in [p]$, we have $\sum_{u=1}^{L} \mathbb{1}(t \in S_u) = \lceil \lambda L/(8\zeta) \rceil$.

(b) For any $u \in [L]$,

(i) the pair of random vectors $\{\mathbf{x}_{S_u}, (\mathbf{x}_{-S_u}, \mathbf{z})\}$ correspond to a τ_1-SGM with $\tau_1 \triangleq (\alpha + 2x_{\max}\zeta, \lambda, x_{\max}, \Theta_{S_u})$ where $\Theta_{S_u} \triangleq \{\Theta_{tv}\}_{t,v\in S_u}$, and

(ii) the random vector $\mathbf{x}_{S_u}$ conditioned on $(\mathbf{x}_{-S_u}, \mathbf{z})$ satisfies the Dobrushin's uniqueness condition (Def. 4) with coupling matrix $2\sqrt{2}x_{\max}^2\Theta_{S_u}$ whenever $\lambda \in \Big(0, \frac{1}{2\sqrt{2}x_{\max}^2}\Big]$ with $\|\Theta_{S_u}\|_{\mathrm{op}} \le \lambda$.

Proof of Prop. 4: Identifying weakly dependent random variables. We prove each part one-by-one.

Proof of part (a): From Def. 8, for any realization $\boldsymbol{z} \in \mathcal{Z}^{p_z}$, the conditional probability distribution of $\mathbf{x}$ given $\mathbf{z} = \boldsymbol{z}$ is given by $f_{\mathbf{x}|\mathbf{z}}\big(\cdot|\boldsymbol{z}; \theta(\boldsymbol{z}), \Theta\big)$ in (4) where $\theta(\boldsymbol{z}) \in \mathbb{R}^p$ is a vector and $\Theta \in \mathbb{R}^{p\times p}$ is a symmetric matrix with $\Theta_{tt} = 0$ for all $t \in [p]$ and $\|\Theta\|_\infty \le \zeta$. Consider the matrix $A \triangleq \frac{1}{\zeta}\Theta$. Since A has zeros on the diagonal and $\|A\|_\infty \le 1$, we can apply [12, Lem. 12] on A with $\eta = \frac{\lambda}{\zeta}$. Then part (a) follows directly from [12, Lem. 12.1].

Proof of part (b) (i): To prove this part, consider the conditional distribution of $\mathbf{x}_{S_u}$ conditioned on $\mathbf{x}_{-S_u} = \boldsymbol{x}_{-S_u}$ and $\mathbf{z} = \boldsymbol{z}$ for any $u \in [L]$. We have

$$f_{\mathbf{x}_{S_u}|\mathbf{x}_{-S_u},\mathbf{z}}(\boldsymbol{x}_{S_u}|\boldsymbol{x}_{-S_u}, \boldsymbol{z}; \theta(\boldsymbol{z}), \Theta) \propto \exp\bigg(\sum_{t\in S_u}\Big(\theta_t(\boldsymbol{z}) + 2\sum_{v\notin S_u}\Theta_{tv}x_v\Big)x_t + \sum_{t\in S_u}\sum_{v\in S_u}\Theta_{tv}x_tx_v\bigg). \tag{130}$$

We can re-parameterize $f_{\mathbf{x}_{S_u}|\mathbf{x}_{-S_u},\mathbf{z}}(\boldsymbol{x}_{S_u}|\boldsymbol{x}_{-S_u},\boldsymbol{z};\theta(\boldsymbol{z}),\Theta)$ in (130) as follows

$$f_{\mathbf{x}_{S_u}|\mathbf{x}_{-S_u},\mathbf{z}}(\boldsymbol{x}_{S_u}|\boldsymbol{x}_{-S_u},\boldsymbol{z};\upsilon(\boldsymbol{z},\boldsymbol{x}_{-S_u}),\Upsilon) \propto \exp\Big([\upsilon(\boldsymbol{z},\boldsymbol{x}_{-S_u})]^\top \boldsymbol{x}_{S_u} + \boldsymbol{x}_{S_u}^\top \Upsilon \boldsymbol{x}_{S_u}\Big), \quad \text{where}$$

$$\upsilon(\boldsymbol{z},\boldsymbol{x}_{-S_u}) \in \mathbb{R}^{|S_u|\times 1}, \quad \text{with} \quad \upsilon_t(\boldsymbol{z},\boldsymbol{x}_{-S_u}) \triangleq \theta_t(\boldsymbol{z}) + 2\sum_{k\notin S_u}\Theta_{tv}x_k \quad \text{for} \quad t\in S_u, \text{ and} \quad (131)$$

$$\Upsilon = \Upsilon^\top \in \mathbb{R}^{|S_u|\times|S_u|} \quad \text{with} \quad \Upsilon_{tv} \triangleq \Theta_{tv}, \text{ and } \Upsilon_{tt}=0 \text{ for all } t,v\in S_u. \quad (132)$$

Thus, to show that the random vector $\mathbf{x}_{S_u}$ conditioned on $\mathbf{x}_{-S_u}$ and $\mathbf{z}$ corresponds to an τ_1-SGM with $\tau_1 \triangleq (\alpha + 2x_{\max}\zeta, \lambda, x_{\max}, \Theta_{S_u})$, it suffices to establish that

$$\max\Big\{\max_{\boldsymbol{z}\in\mathcal{Z}^{p_z}}\|\upsilon(\boldsymbol{z},\boldsymbol{x}_{-S_u})\|_\infty, \|\Upsilon\|_{\max}\Big\} \overset{(i)}{\leq} \alpha + 2x_{\max}\zeta \quad \text{and} \quad \|\Upsilon\|_\infty \overset{(ii)}{\leq} \lambda. \quad (133)$$

To establish (i) in (133), we note that

$$\|\Upsilon\|_{\max} \overset{(132)}{\leq} \|\Theta\|_{\max} \overset{(a)}{\leq} \alpha \quad \text{and} \quad (134)$$

$$\|\upsilon(\boldsymbol{z},\boldsymbol{x}_{-S_u})\|_\infty \overset{(b)}{\leq} \|\theta(\boldsymbol{z})\|_\infty + 2\max_{t\in S_u}\|\Theta_t\|_1\|\boldsymbol{x}\|_\infty \overset{(c)}{\leq} \|\theta(\boldsymbol{z})\|_\infty + 2x_{\max}\|\Theta\|_\infty \overset{(d)}{\leq} \alpha + 2x_{\max}\zeta, \quad (135)$$

where (a) and (d) follow from Def. 8, (b) follows from (131) and the triangle inequality, and (c) follows from the definition of $\|\cdot\|_\infty$ and Def. 8. Then, from (134) and (135), we have

$$\max\Big\{\max_{\boldsymbol{z}\in\mathcal{Z}^{p_z}}\|\upsilon(\boldsymbol{z},\boldsymbol{x}_{-S_u})\|_\infty, \|\Upsilon\|_{\max}\Big\} \leq \alpha + 2x_{\max}\zeta$$

as claimed. Next, to establish (ii) in (133), we again apply [12, Lem. 12] on the matrix $A = \frac{1}{\zeta}\Theta$ with $\eta = \frac{\lambda}{\zeta}$. Then, [12, Lem. 12.2] implies that

$$\sum_{v\in S_u}\left|\frac{\Theta_{tv}}{\zeta}\right| \leq \frac{\lambda}{\zeta} \quad \text{for all } t\in S_u, u\in[L]. \quad (136)$$

Therefore, we have

$$\|\Upsilon\|_\infty = \max_{t\in S_u}\Big(\sum_{v\in S_u}|\Upsilon_{tv}|\Big) \overset{(132)}{=} \max_{t\in S_u}\Big(\sum_{v\in S_u}|\Theta_{tv}|\Big) \overset{(136)}{\leq} \lambda, \quad (137)$$

as desired. The proof for this part is now complete.

Proof of part (b) (ii): We start by noting that the operator norm of a symmetric matrix is bounded by the infinity norm of the matrix. Then, from the analysis in part (b) (i), for any $u \in S_u$, we have

$$\|\Theta_{S_u}\|_{\text{op}} \leq \|\Theta_{S_u}\|_\infty \overset{(132)}{=} \|\Upsilon\|_\infty \overset{(137)}{\leq} \lambda.$$

Therefore, $\|2\sqrt{2}x_{\max}^2\Theta_{S_u}\|_\infty \leq 1$ whenever $\lambda \leq 1/2\sqrt{2}x_{\max}^2$. It remains to show that for every $u\in[L], t\in S_u, v\in S_u\backslash\{t\}$, $\mathbf{z}=\boldsymbol{z}$, and $\boldsymbol{x}_{-t},\widetilde{\boldsymbol{x}}_{-t}\in\mathcal{X}^{p-1}$ differing only in the v^{th} coordinate,

$$\|f_{\mathbf{x}_t|\mathbf{x}_{-t}=\boldsymbol{x}_{-t},\mathbf{z}=\boldsymbol{z}} - f_{\mathbf{x}_t|\mathbf{x}_{-t}=\widetilde{\boldsymbol{x}}_{-t},\mathbf{z}=\boldsymbol{z}}\|_{\mathsf{TV}} \leq 2\sqrt{2}x_{\max}^2\Theta_{tv}.$$

To that end, fix any $u\in[L]$, any $t\in S_u$, any $v\in S_u\backslash\{t\}$, any $\mathbf{z}=\boldsymbol{z}$, and any $\boldsymbol{x}_{-t},\widetilde{\boldsymbol{x}}_{-t}\in\mathcal{X}^{p-1}$ differing only in the v^{th} coordinate. We have

$$\|f_{\mathbf{x}_t|\mathbf{x}_{-t}=\boldsymbol{x}_{-t},\mathbf{z}=\boldsymbol{z}} - f_{\mathbf{x}_t|\mathbf{x}_{-t}=\widetilde{\boldsymbol{x}}_{-t},\mathbf{z}=\boldsymbol{z}}\|_{\mathsf{TV}}^2 \overset{(a)}{\leq} \frac{1}{2}\mathsf{KL}\left(f_{\mathbf{x}_t|\mathbf{x}_{-t}=\boldsymbol{x}_{-t},\mathbf{z}=\boldsymbol{z}}\,\Big\|\,f_{\mathbf{x}_t|\mathbf{x}_{-t}=\widetilde{\boldsymbol{x}}_{-t},\mathbf{z}=\boldsymbol{z}}\right)$$

$$\overset{(b)}{=} \frac{1}{2}(2\Theta_{tv}x_v - 2\Theta_{tv}\widetilde{x}_v)^2 x_{\max}^2 \overset{(c)}{\leq} 8x_{\max}^4\Theta_{tv}^2,$$

where (a) follows from Pinsker's inequality, (b) follows by (i) applying [11, Theorem 1] to the exponential family parameterized as per $f_{\mathbf{x}_t|\mathbf{x}_{-t},\mathbf{z}}$ in (8), (ii) noting that $f_{\mathbf{x}_t|\mathbf{x}_{-t}=\boldsymbol{x}_{-t},\mathbf{z}=\boldsymbol{z}} \propto \exp\big([\theta_t(\boldsymbol{z}) + 2\Theta_t^\top\boldsymbol{x}]x_t\big)$ and $f_{\mathbf{x}_t|\mathbf{x}_{-t}=\widetilde{\boldsymbol{x}}_{-t},\mathbf{z}=\boldsymbol{z}} \propto \exp\big([\theta_t(\boldsymbol{z}) + 2\Theta_t^\top\widetilde{\boldsymbol{x}}]x_t\big)$, and (iii) noting that the Hessian of the log partition function for any regular exponential family is the covariance matrix of the associated sufficient statistic which is bounded by $x_{\max}^2$ when $\mathcal{X}=\{-x_{\max},x_{\max}\}$, and (c) follows because $x_v,\widetilde{x}_v\in\{-x_{\max},x_{\max}\}$. This completes the proof. □

G Supporting concentration results

In this section, we provide a corollary of Prop. 3 that is used to prove the concentration results in Lem. 12 and Lem. 14. To show any concentration result for the random vector $\mathbf{x}$ conditioned on $\mathbf{z}$ via Prop. 3, we need $\mathbf{x}|\mathbf{z}$ to satisfy the logarithmic Sobolev inequality (defined in (81)). From Prop. 2, for this to be true, we need the random vector x_t conditioned on $(\mathbf{x}_{-t}, \mathbf{z})$ to satisfy the logarithmic Sobolev inequality for all $t \in [p]$. In the result below, we show this holds with a proof in App. G.1. We define a $\tau \triangleq (\alpha, \zeta, x_{\max}, \Theta)$-dependent constant:

$$C_{3,\tau} \triangleq \exp\left(x_{\max}(\alpha + 2\zeta x_{\max})\right). \tag{138}$$

Lemma 19 (Logarithmic Sobolev inequality for $x_t|\mathbf{x}_{-t}, \mathbf{z}$). *Given a pair of random vectors $\{\mathbf{x}, \mathbf{z}\}$ supported on $\mathcal{X}^p \times \mathcal{Z}^{p_z}$ that is a τ-SGM (Def. 8) with $\tau \triangleq (\alpha, \zeta, x_{\max}, \Theta)$, $x_t|\mathbf{x}_{-t}, \mathbf{z}$ satisfies $\mathrm{LSI}_{x_t|\mathbf{x}_{-t}=\boldsymbol{x}_{-t}, \mathbf{z}=\boldsymbol{z}}\Big(\frac{8x_{\max}^2}{\pi^2}C_{3,\tau}^2\Big)$ for all $t \in [p]$, $\boldsymbol{x}_{-t} \in \mathcal{X}^{p-1}$, and $\boldsymbol{z} \in \mathcal{Z}^{p_z}$.*

Now, we state the desired corollary of Prop. 3 with a proof in App. G.2. The corollary makes use of some $\tau \triangleq (\alpha, \zeta, x_{\max}, \Theta)$-dependent constants:

$$C_{4,\tau} \triangleq 1 + \alpha x_{\max} + 4x_{\max}^2\zeta \quad \text{and} \quad C_{5,\tau} \triangleq \frac{32x_{\max}^3 C_{3,\tau}^4}{\pi^2}. \tag{139}$$

Corollary 2. *Suppose a pair of random vectors $\{\mathbf{x}, \mathbf{z}\}$ supported on $\mathcal{X}^p \times \mathcal{Z}^{p_z}$ corresponds to a τ-SGM (Def. 8) with $\tau \triangleq (\alpha, \zeta, x_{\max}, \Theta)$, and $\mathbf{x}$ conditioned on $\mathbf{z}$ satisfies the Dobrushin's uniqueness condition (Def. 4) with coupling matrix $\overline{\Theta}$. For any $\theta, \overline{\theta} \in \Lambda_\theta$ and $\Theta \in \Lambda_\Theta$, define the functions q_1 and q_2 as*

$$q_1(\mathbf{x}) \triangleq \sum_{t\in[p]} (\omega_t x_t)^2 \quad \textit{and} \quad q_2(\mathbf{x}) \triangleq \sum_{t\in[p]} \omega_t x_t \exp\Big(-[\theta_t + 2\Theta_t^\top \mathbf{x}]x_t\Big) \quad \textit{where} \quad \omega = \overline{\theta} - \theta.$$

Then, for any $\varepsilon > 0$

$$\mathbb{P}\Big[\big|q_i(\mathbf{x}) - \mathbb{E}\big[q_i(\mathbf{x})\big|\mathbf{z}\big]\big| \geq \varepsilon\Big|\mathbf{z}\Big] \leq \exp\left(\frac{-c\big(1 - \|\overline{\Theta}\|_{\mathrm{op}}\big)^4\varepsilon^2}{c_i\|\omega\|_2^2}\right) \quad \textit{for} \quad i = 1, 2, \tag{140}$$

where c is a universal constant, $c_1 \triangleq 16\alpha^2 x_{\max}^2 C_{5,\tau}^2$, and $c_2 \triangleq C_{3,\tau}^2 C_{7,\tau}^2 C_{5,\tau}^2$ with $C_{3,\tau}$ defined in (138) and $C_{4,\tau}$ and $C_{5,\tau}$ defined in (139).

G.1 Proof of Lem. 19: Logarithmic Sobolev inequality for $x_t|\mathbf{x}_{-t}, \mathbf{z}$

Let u be the uniform distribution on $\mathcal{X}$. Then, u satisfies $\mathrm{LSI}_u\Big(\frac{8x_{\max}^2}{\pi^2}\Big)$ (see [14, Corollary 2.4]). Then, using the Holley-Stroock perturbation principle (see [16, Page 31], [19, Lemma 1.2]), for every $t \in [p]$, $\boldsymbol{x}_{-t} \in \mathcal{X}^{p-1}$, and $\boldsymbol{z} \in \mathcal{Z}^{p_z}$, $x_t|\mathbf{x}_{-t} = \boldsymbol{x}_{-t}, \mathbf{z} = \boldsymbol{z}$ satisfies the logarithmic Sobolev inequality with a constant bounded by

$$\frac{8x_{\max}^2 \exp(\sup_{x_t\in\mathcal{X}} \psi(x_t; \boldsymbol{x}_{-t}, \boldsymbol{z}) - \inf_{x_t\in\mathcal{X}} \psi(x_t; \boldsymbol{x}_{-t}, \boldsymbol{z}))}{\pi^2}$$

where $\psi(x_t; \boldsymbol{x}_{-t}, \boldsymbol{z}) \triangleq -[\theta_t(\boldsymbol{z}) + 2\Theta_t^\top \boldsymbol{x}]x_t$. We have

$$\begin{aligned}\exp(\sup_{x_t\in\mathcal{X}} \psi(x_t; \boldsymbol{x}_{-t}, \boldsymbol{z}) - \inf_{x_t\in\mathcal{X}} \psi(x_t; \boldsymbol{x}_{-t}, \boldsymbol{z})) &\overset{(a)}{=} \exp\left(2\big|\theta_t(\boldsymbol{z}) + 2\Theta_t^\top \boldsymbol{x}\big|x_{\max}\right)\\ &\overset{(b)}{\leq} \exp\left((2\alpha + 4\zeta x_{\max})x_{\max}\right) \overset{(138)}{=} C_{3,\tau}^2,\end{aligned}$$

where (a) follows from Def. 8 and (b) follows by using Def. 8 along with triangle inequality and Cauchy–Schwarz inequality.

G.2 Proof of Cor. 2

To apply Prop. 3 to the random vector $\mathbf{x}$ conditioned on $\mathbf{z}$, we need $\mathbf{x}|\mathbf{z}$ to satisfy the logarithmic Sobolev inequality. From Prop. 2, this is true if (i) $f_{\min} = \min_{t\in[p],\boldsymbol{x}\in\mathcal{X}^p,\boldsymbol{z}\in\mathcal{X}^{p_z}} f_{\mathsf{x}_t|\mathbf{x}_{-t},\mathbf{z}}(x_t|\boldsymbol{x}_{-t},\boldsymbol{z}) > 0$ (see (83)), (ii) $\mathbf{x}|\mathbf{z}$ satisfies the Dobrushin's uniqueness condition, and (iii) $\mathsf{x}_t|\mathbf{x}_{-t},\mathbf{z}$ satisfies the logarithmic Sobolev inequality for all $t \in [p]$. By assumption, $\mathbf{x}|\mathbf{z}$ satisfies the Dobrushin's uniqueness condition with coupling matrix $\overline{\Theta}$. From Lem. 19, $\mathsf{x}_t|\mathbf{x}_{-t},\mathbf{z}$ satisfies $\mathrm{LSI}_{\mathsf{x}_t|\mathbf{x}_{-t}=\boldsymbol{x}_{-t},\mathbf{z}=\boldsymbol{z}}\Big(\frac{8x_{\max}^2C_{3,\tau}^2}{\pi^2}\Big)$. It remains to show that $f_{\min} > 0$. Consider any $t \in [p]$, any $\boldsymbol{x} \in \mathcal{X}^p$, and any $\boldsymbol{z} \in \mathcal{X}^{p_z}$. We have

$$
\begin{aligned}
f_{\mathsf{x}_t|\mathbf{x}_{-t},\mathbf{z}}(x_t|\boldsymbol{x}_{-t},\boldsymbol{z}) &\stackrel{(a)}{=} \frac{\exp\Big([\theta_t(\boldsymbol{z}) + 2\Theta_t^\top \boldsymbol{x}]x_t\Big)}{\int_{\mathcal{X}}\exp\Big([\theta_t(\boldsymbol{z}) + 2\Theta_t^\top \boldsymbol{x}]x_t\Big)dx_t} \\
&\stackrel{(b)}{\geq} \frac{\exp\Big(-|\theta_t(\boldsymbol{z}) + 2\Theta_t^\top \boldsymbol{x}|x_{\max}\Big)}{\int_{\mathcal{X}}\exp\Big(|\theta_t(\boldsymbol{z}) + 2\Theta_t^\top \boldsymbol{x}|x_{\max}\Big)dx_t} \\
&\stackrel{(c)}{\geq} \frac{\exp\Big(-(|\theta(\boldsymbol{z})| + 2\|\Theta_t\|_1\|\boldsymbol{x}\|_\infty)x_{\max}\Big)}{\int_{\mathcal{X}}\exp\Big((|\theta(\boldsymbol{z})| + 2\|\Theta_t\|_1\|\boldsymbol{x}\|_\infty)x_{\max}\Big)dx_t} \\
&\stackrel{(d)}{\geq} \frac{\exp\Big(-(\alpha + 2\zeta x_{\max})x_{\max}\Big)}{\int_{\mathcal{X}}\exp\Big((\alpha + 2\zeta x_{\max})x_{\max}\Big)dx_t} \stackrel{(e)}{=} \frac{1}{2x_{\max}C_{3,\tau}^2},
\end{aligned}
$$

where (a) follows from (8), (b) and (d) follow from Def. 8, (c) follows by triangle inequality and Cauchy–Schwarz inequality, and (e) follows because $\int_{\mathcal{X}} dx_t = 2x_{\max}$. Therefore, $f_{\min} = \frac{1}{2x_{\max}C_{3,\tau}^2}$. Putting (i), (ii), and (iii) together, and using Prop. 2, we see that $\mathbf{x}|\mathbf{z}$ satisfies $\mathrm{LSI}_{\mathbf{x}}\Big(\frac{C_{5,\tau}}{(1-\|\overline{\Theta}\|_{\mathrm{op}})^2}\Big)$ where $C_{5,\tau}$ was defined in (139).

Now, we apply Prop. 3 to q_1 and q_2 one-by-one. The general strategy is to choose appropriate pseudo derivatives and pseudo Hessians for both q_1 and q_2, and evaluate the corresponding terms appearing in Prop. 3.

Concentration for q_1: Fix any $\boldsymbol{x} \in \mathcal{X}^p$. We start by decomposing $q_1(\boldsymbol{x})$ as follows:

$$q_1(\boldsymbol{x}) = \varpi^\top r(\boldsymbol{x}) \tag{141}$$

where $\varpi \triangleq (\omega_1^2, \cdots, \omega_p^2)$ and $r(\boldsymbol{x}) \triangleq (r_1(\boldsymbol{x}), \cdots, r_p(\boldsymbol{x}))$ with $r_t(\boldsymbol{x}) = x_t^2$ for every $t \in [p]$. Next, we define $H : \mathcal{X}^p \to \mathbb{R}^{p\times p}$ such that

$$H_{tu}(\boldsymbol{x}) = \frac{dr_u(\boldsymbol{x})}{dx_t} \quad \text{for every } t, u \in [p]. \tag{142}$$

Pseudo derivative: We bound the ℓ_2 norm of the gradient of $q_1(\boldsymbol{x})$ as follows:

$$
\begin{aligned}
\|\nabla q_1(\boldsymbol{x})\|_2^2 = \sum_{t\in[p]}\Big(\frac{dq_1(\boldsymbol{x})}{dx_t}\Big)^2 \stackrel{(141)}{=} \sum_{t\in[p]}\Big(\frac{\varpi^\top dr(\boldsymbol{x})}{dx_t}\Big)^2 &\stackrel{(142)}{=} \|H(\boldsymbol{x})\varpi\|_2^2 \\
&\stackrel{(a)}{\leq} \|H(\boldsymbol{x})\|_{\mathrm{op}}^2 \|\varpi\|_2^2 \\
&\stackrel{(b)}{\leq} \|H(\boldsymbol{x})\|_1 \|H(\boldsymbol{x})\|_\infty \|\varpi\|_2^2
\end{aligned} \tag{143}
$$

where (a) follows because induced matrix norms are submultiplicative and (b) follows because the matrix operator norm is bounded by square root of the product of matrix one norm and matrix infinity norm. Now, we claim that the one norm and the infinity norm of $H(\boldsymbol{x})$ are bounded as follows:

$$\max\Big\{\max_{\boldsymbol{x}\in\mathcal{X}^p}\|H(\boldsymbol{x})\|_1, \max_{\boldsymbol{x}\in\mathcal{X}^p}\|H(\boldsymbol{x})\|_\infty\Big\} \leq 2x_{\max}. \tag{144}$$

Taking this claim as given at the moment, we continue with our proof. Combining (143) and (144), we have

$$\max_{\boldsymbol{x}\in\mathcal{X}^p} \|\nabla q_1(\boldsymbol{x})\|_2^2 \leq 4x_{\max}^2 \|\overline{\omega}\|_2^2 = 4x_{\max}^2 \sum_{t\in[p]} \omega_t^4 \leq 4x_{\max}^2 \max_{u\in[p]} \omega_u^2 \sum_{t\in[p]} \omega_t^2 \stackrel{(a)}{\leq} 16x_{\max}^2 \alpha^2 \|\omega\|_2^2,$$

where (a) follows because $\omega \in 2\Lambda_\theta$. Therefore, we choose the pseudo derivative (see Def. 5) as follows:

$$\widetilde{\nabla} q_1(\boldsymbol{x}) = 4x_{\max}\alpha \|\omega\|_2. \tag{145}$$

Pseudo Hessian: Fix any $\rho \in \mathbb{R}$. We bound $\|\nabla(\rho^\top \widetilde{\nabla} q_1(\boldsymbol{x}))\|_2^2$ (see Def. 5) as follows:

$$\|\nabla(\rho^\top \widetilde{\nabla} q_1(\boldsymbol{x}))\|_2^2 = \sum_{u\in[p]} \Big(\frac{d\rho^\top \widetilde{\nabla} q_1(\boldsymbol{x})}{dx_u}\Big)^2 \stackrel{(145)}{=} 0.$$

Therefore, we choose the pseudo Hessian (see Def. 5) as follows:

$$\widetilde{\nabla}^2 q_1(\boldsymbol{x}) = 0. \tag{146}$$

The concentration result in (140) for q_1 follows by applying Prop. 3 with the pseudo discrete derivative defined in (145) and the pseudo discrete Hessian defined in (146).

It remains to show that the one-norm and the infinity-norm of $H(\boldsymbol{x})$ are bounded as in (144).

Bounds on the one-norm and the infinity-norm of $H(\boldsymbol{x})$: We have

$$H_{tu}(\boldsymbol{x}) = \begin{cases} 2x_t & \text{if} \quad t = u, \\ 0 & \text{otherwise.} \end{cases} \tag{147}$$

Therefore,

$$\|H(\boldsymbol{x})\|_1 = \max_{u\in[p]} \sum_{t\in[p]} |H_{tu}(\boldsymbol{x})| \stackrel{(147)}{\leq} \max_{u\in[p]} 2|x_u| \stackrel{(a)}{\leq} 2x_{\max} \quad \text{and}$$
$$\|H(\boldsymbol{x})\|_\infty = \max_{t\in[p]} \sum_{u\in[p]} |H_{tu}(\boldsymbol{x})| \stackrel{(147)}{\leq} \max_{t\in[p]} 2|x_t| \stackrel{(a)}{\leq} 2x_{\max},$$

where (a) follows from Def. 8.

Concentration for q_2: Fix any $\boldsymbol{x} \in \mathcal{X}^p$. We start by decomposing $q_2(\boldsymbol{x})$ as follows:

$$q_2(\boldsymbol{x}) = \omega^\top r(\boldsymbol{x}) \tag{148}$$

where $r(\boldsymbol{x}) \triangleq (r_1(\boldsymbol{x}), \cdots, r_p(\boldsymbol{x}))$ with $r_t(\boldsymbol{x}) = x_t \exp\big(-[\theta_t + 2\Theta_t^\top \boldsymbol{x}]x_t\big)$ for every $t \in [p]$. Next, we define $H : \mathcal{X}^p \to \mathbb{R}^{p\times p}$ such that

$$H_{tu}(\boldsymbol{x}) = \frac{dr_u(\boldsymbol{x})}{dx_t} \quad \text{for every } t, u \in [p]. \tag{149}$$

Pseudo derivative: We bound the ℓ_2 norm of the gradient of $q_2(\boldsymbol{x})$ as follows:

$$\begin{aligned} \|\nabla q_2(\boldsymbol{x})\|_2^2 = \sum_{t\in[p]} \Big(\frac{dq_2(\boldsymbol{x})}{dx_t}\Big)^2 \stackrel{(148)}{=} \sum_{t\in[p]} \Big(\frac{\omega^\top dr(\boldsymbol{x})}{dx_t}\Big)^2 &\stackrel{(149)}{=} \|H(\boldsymbol{x})\omega\|_2^2 \\ &\stackrel{(a)}{\leq} \|H(\boldsymbol{x})\|_{\text{op}}^2 \|\omega\|_2^2 \\ &\stackrel{(b)}{\leq} \|H(\boldsymbol{x})\|_1 \|H(\boldsymbol{x})\|_\infty \|\omega\|_2^2 \end{aligned} \tag{150}$$

where (a) follows because induced matrix norms are submultiplicative and (b) follows because the matrix operator norm is bounded by square root of the product of matrix one norm and matrix infinity norm. Now, we claim that the one norm and the infinity norm of $H(\boldsymbol{x})$ are bounded as follows:

$$\max\left\{\max_{\boldsymbol{x}\in\mathcal{X}^p}\|H(\boldsymbol{x})\|_1, \max_{\boldsymbol{x}\in\mathcal{X}^p}\|H(\boldsymbol{x})\|_\infty\right\} \leq C_{3,\tau}C_{4,\tau}. \tag{151}$$

where $C_{3,\tau}$ and $C_{4,\tau}$ were defined in (138) and (139) respectively. Taking this claim as given at the moment, we continue with our proof. Combining (150) and (151), we have

$$\max_{\boldsymbol{x}\in\mathcal{X}^p}\|\nabla q_2(\boldsymbol{x})\|_2^2 \leq C_{3,\tau}^2 C_{4,\tau}^2 \|\omega\|_2^2.$$

Therefore, we choose the pseudo derivative (see Def. 5) as follows:

$$\widetilde{\nabla} q_2(\boldsymbol{x}) = C_{3,\tau}C_{4,\tau}\|\omega\|_2. \tag{152}$$

Pseudo Hessian: Fix any $\rho \in \mathbb{R}$. We bound $\|\nabla(\rho^\top \widetilde{\nabla} q_2(\boldsymbol{x}))\|_2^2$ (see Def. 5) as follows:

$$\|\nabla(\rho^\top \widetilde{\nabla} q_2(\boldsymbol{x}))\|_2^2 = \sum_{u\in[p]}\Big(\frac{d\rho^\top \widetilde{\nabla} q_2(\boldsymbol{x})}{dx_u}\Big)^2 \overset{(152)}{=} 0.$$

Therefore, we choose the pseudo Hessian (see Def. 5) as follows:

$$\widetilde{\nabla}^2 q_2(\boldsymbol{x}) = 0. \tag{153}$$

The concentration result in (140) for q_1 follows by applying Prop. 3 with the pseudo discrete derivative defined in (152) and the pseudo discrete Hessian defined in (153).

It remains to show that the one-norm and the infinity-norm of $H(\boldsymbol{x})$ are bounded as in (151).

Bounds on the one-norm and the infinity-norm of H: We have

$$H_{tu}(\boldsymbol{x}) = \begin{cases}\big[1-[\theta_u+2\Theta_u^\top\boldsymbol{x}]x_u\big]\exp\big(-[\theta_u+2\Theta_u^\top\boldsymbol{x}]x_u\big) & \text{if} \quad t=u,\\ -2\Theta_{tu}x_u^2\exp\big(-[\theta_u+2\Theta_u^\top\boldsymbol{x}]x_u\big) & \text{otherwise.}\end{cases} \tag{154}$$

Therefore,

$$\begin{aligned}\|H(\boldsymbol{x})\|_1 &= \max_{u\in[p]}\sum_{t\in[p]}|H_{tu}(\boldsymbol{x})|\\ &\overset{(154)}{=} \max_{u\in[p]}\big|1-[\theta_u+2\Theta_u^\top\boldsymbol{x}]x_u\big|\exp\big(-[\theta_u+2\Theta_u^\top\boldsymbol{x}]x_u\big)\\ &\qquad\qquad +2\max_{u\in[p]}x_u^2\exp\big(-[\theta_u+2\Theta_u^\top\boldsymbol{x}]x_u\big)\sum_{t\neq u}|\Theta_{tu}|\\ &\overset{(a)}{\leq} (1+\alpha x_{\max}+4x_{\max}^2\zeta)\exp\big(x_{\max}(\alpha+2\zeta x_{\max})\big) \overset{(b)}{=} C_{3,\tau}C_{4,\tau}\end{aligned}$$

where (a) follows from Def. 8 along with triangle inequality and Cauchy–Schwarz inequality and (b) follows from (138) and (139). Similarly, we have

$$\begin{aligned}\|H(\boldsymbol{x})\|_\infty &= \max_{t\in[p]}\sum_{u\in[p]}|H_{tu}(\boldsymbol{x})|\\ &\overset{(154)}{=} \max_{t\in[p]}\big|1-[\theta_t+2\Theta_t^\top\boldsymbol{x}]x_t\big|\exp\big(-[\theta_t+2\Theta_t^\top\boldsymbol{x}]x_t\big)\\ &\qquad\qquad +2\max_{t\in[p]}\sum_{u\neq t}|\Theta_{tu}|x_u^2\exp\big(-[\theta_u+2\Theta_u^\top\boldsymbol{x}]x_u\big)\\ &\overset{(a)}{\leq} (1+\alpha x_{\max}+4x_{\max}^2\zeta)\exp\big(x_{\max}(\alpha+2\zeta x_{\max})\big) \overset{(b)}{=} C_{3,\tau}C_{4,\tau}\end{aligned}$$

where (a) follows from Def. 8 along with triangle inequality and Cauchy–Schwarz inequality and (b) follows from (138) and (139).