Journal of Machine Learning Research 26 (2025) 1-66 Submitted 4/24; Revised 10/25; Published 11/25

Collaborative likelihood-ratio estimation over graphs

Alejandro de la Concha ALEJANDRO.DE_LA_CONCHA_DUARTE@ENS-PARIS-SACLAY.FR
Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, 91190 Gif-sur-Yvette, France
Department of Mathematics, University of Luxembourg, 4364 Esch-sur-Alzette, Luxembourg

Nicolas Vayatis NICOLAS.VAYATIS@ENS-PARIS-SACLAY.FR
Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, 91190 Gif-sur-Yvette, France

Argyris Kalogeratos ARGYRIS.KALOGERATOS@ENS-PARIS-SACLAY.FR
Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, 91190 Gif-sur-Yvette, France

Editor: Krishnakumar Balasubramanian

Abstract

This paper introduces the *Collaborative Likelihood-ratio Estimation problem*, which is relevant for applications involving multiple statistical estimation tasks that can be mapped to the nodes of a fixed graph expressing pairwise task similarity. Each graph node v observes i.i.d. data from two unknown node-specific pdfs, p_v and q_v, and the goal is to estimate the likelihood-ratios (or density-ratios), $r_v(x) = \frac{q_v(x)}{p_v(x)}$, for all v. Our contribution is multifold: we present a non-parametric collaborative framework that leverages the graph structure of the problem to solve the tasks more efficiently; we present a concrete method that we call Graph-based Relative Unconstrained Least-Squares Importance Fitting (GRULSIF) along with an efficient implementation; we derive convergence rates that highlight the role of the main variables of the problem. Our theoretical results explicit the conditions under which the collaborative estimation leads to performance gains compared to solving each estimation task independently. Finally, in a series of experiments, we demonstrate that the joint likelihood-ratio estimation of GRULSIF at all graph nodes is more accurate compared to state-of-the-art methods that operate independently at each node, and we verify that the behavior of GRULSIF is in agreement with our theoretical analysis.

Keywords: Unsupervised learning, f-divergence, likelihood-ratio estimation, kernel methods, graph regularization, multitask learning.

1. Introduction

The quantification and statistical validation of the difference between two probability measures, P and Q, is a fundamental problem in Machine Learning and Statistics. The likelihood-ratio between the associated pdfs, $r(x) = \frac{q(x)}{p(x)}$, can serve this purpose and has been used in designing statistical tests and detectors (Tartakovsky et al., 2014; Lehmann and Romano, 2022). In many practical cases where p and q are unknown and only sampled data are available from each of them, $r(\cdot)$ needs to be estimated from the available data via non-parametric methods that require minimal assumptions about p and q. A critical component in likelihood-ratio estimation (LRE) is the variational representation of f-divergences (ϕ-divergence), which shows how measuring the dissimilarity between P and Q, that is $\mathcal{D}_\phi(P\|Q)$, is equivalent to solving a functional optimization problem (Nguyen et al., 2008, 2010; Sugiyama et al., 2012; Agrawal and Horel, 2021; Birrell et al., 2022a). Interestingly,

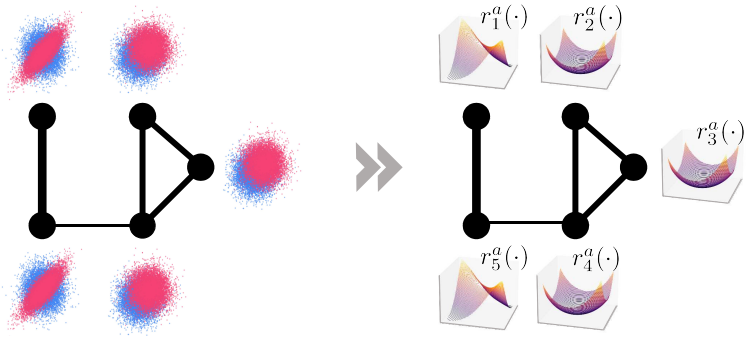

Figure 1: Collaborative LRE over a graph. Given observations from two probabilistic models p_v (blue) and q_v (pink) at each node v of a graph (left), the proposed Collaborative LRE framework aims at estimating jointly all the N associated likelihood-ratios r_v (right) in a collaborative and distributed manner. This example shows how using only the input data $\mathcal{X} \subset \mathbb{R}^2$, essentially a vector-valued function $\mathbf{r}(\cdot) = (r_1(\cdot), ..., r_N(\cdot))$ is derived holding the likelihood-ratios.

under some conditions on P, Q, and $\mathcal{D}_\phi(P\|Q)$, the solution to this optimization problem leads to an approximation of the likelihood-ratio.

The existing research has focused on the single LRE problem, however, modern challenges are posed when multiple entities need to solve similar problems using local data samples. For instance, each entity may correspond to an agency or sensor collecting local data for different geographic areas, e.g. meteorological data, air pollution sensor networks, medical surveys across geographic regions. Then, one's interest may be to estimate a likelihood-ratio at each location, which can be subsequently used in application tasks. In such application settings, Collaborative LRE could play a crucial role by ensuring that the heterogeneity among what each entity observes locally will not get diluted by any sort of global data aggregation. Under certain conditions, this approach can enhance the estimation of the likelihood-ratios associated to each entity, hence the performance in the application task of one's interest, as compared to ignoring the graph structure of the problem. Therefore, the intriguing question we put forward in this work is how can these entities collaborate to solve their LRE tasks better than operating independently on their own.

Contribution. This work introduces *the Collaborative LRE* problem for multiple data sources: each data source v, represented as a node in a fixed graph, intends to compare two node-specific pdfs, p_v and q_v, using local i.i.d. observations. The novelty lies in the collaborative estimation of the likelihood-ratios, $r_v(x) = \frac{q_v(x)}{p_v(x)}$, instead of independently at each node. Our fundamental hypothesis is that the graph structure conveys valuable information about how similar are expected to be the estimation tasks at any two nodes. A visual summary of the problem is provided in Fig. 1. We also present an algorithm for the *Collaborative LRE* problem, called Graph-based Relative Unconstrained Least Squares Importance Fitting (GRULSIF). As in previous works in LRE, GRULSIF is defined in terms of the variational representation of ϕ-divergences (Nguyen et al., 2008, 2010). In this graph-based extension where multiple likelihood-ratios need to be estimated, we consider a functional space of vector-valued functions whose geometry encodes the node similarity hypothesis. Our framework assumes that the likelihood-ratios we aim to estimate $\{r_v\}_{v \in V}$ are elements of a Reproducing Kernel Hilbert Space (RKHS) $\mathbb{H}$ shared among all nodes,

and that at two adjacent nodes u and v, the likelihood-ratios r_u and r_v will be close to each other in $\mathbb{H}$. This graph-based setting has the following characteristics:

1. The Collaborative LRE is an instance of Multitask Learning, which has been shown to improve the generalization in supervised problems (Maurer, 2006a,b; Yousefi et al., 2018; Zhang and Yang, 2021) and has motivated special and often distributed optimization schemes (Nassif et al., 2020a,b,c; Zhang and Yang, 2021).
2. The GRULSIF method is characterized by convergence rates that are derived thanks also to the available theoretical results in the literature. Moreover, we identify under which conditions the Collaborative LRE outperforms the independent LRE, as well as the sensitivity of the former to important problem variables, such as the amount of available data per node, the number of nodes, and whether prior information is provided in the form of a graph structure.
3. Our numerical implementation exploits the graph structure of the problem and a shared RKHS for all nodes, and scales well in the number of nodes. GRULSIF provides also guidelines for hyperparameter selection and its sensitivity to different choices.

LRE applications and motivation for graph-based extensions. The interplay between ϕ-divergence and likelihood-ratio has led to several applications, such as Transfer Learning, Hypothesis Testing, and Change-point Detection. Transfer Learning relaxes the classical hypothesis that the training and the test datasets are samples of the same distribution, and relies instead on importance weighting that trains a predictive model by focusing on training losses that are weighted according to the test-over-training likelihood-ratio, $r(x) = \frac{q_{\text{test}}(x)}{p_{\text{train}}(x)}$ (Huang et al., 2006; Sugiyama et al., 2007; Yamada et al., 2013; Lu et al., 2023). In Hypothesis Testing, statistical tests based on ϕ-divergences are suitable when the forms of q and p are unknown, and only data samples are available. The test statistic takes the form of an approximated ϕ-divergence via empirical averages over estimated likelihood-ratios (Sugiyama et al., 2011b, 2012; Yamada et al., 2013). Similarly, in non-parametric Change-point Detection the goal is to detect the moment at which a time-series changes behavior from p to q, which are both unknown (Liu et al., 2013; Ferrari et al., 2023). The workflow in such applications comprises two stages: the likelihood-ratio is first estimated, and then it is used to compute application-specific scores, e.g. a weighting function or a test statistic. This explains the broad interest of the research community in generic LRE approaches that can implement the first of the above stages.

The graph-based counterpart of this problem introduced in this work is motivated by the recent technological advances that have increased the capacity to monitor many aspects of daily life or natural phenomena by collecting and analyzing data coming from multiple sources. Imagine, for example, a network of sensors monitoring air quality of a city, where sensors that are close to each other are likely to record similar signals. Using the data from each sensor and the similarities between them, one could design a two-sample test based on likelihood-ratio estimation to detect pollution peaks. In another example, the similarity between hospitals in terms of the patient profiles they receive, can be exploited by graph-based LRE to enable Transfer Learning and update their diagnostic algorithms in response to changes in the behavior of a disease.

The emergence of such systems that are made of multiple sensors or agents, each of them generating their own data, poses the intriguing question how to adapt LRE approaches so

they integrate this heterogeneity while enabling collaboration. This challenge has already been mentioned in Sugiyama et al. (2012) (chapter: "Conclusions and Future Directions"), where the need for such methods was acknowledged so that likelihood-ratio-based techniques to be applied to real-world problems such as sensory data, finance and neuroscience. Besides, Multitask Learning and Federated Learning literature acknowledges graph-based modeling as a promising research direction to address within their scope. The underlying hypothesis is that agents share similarities that can be modeled via graph-based techniques, and then those similarities can be leveraged to design more efficient machine learning methods. This work is on this direction by proposing the GRULSIF framework that can enable applications such as the aforementioned. Worth noting, this framework has already been employed in works developed in parallel as foundation for Change-Point Detection and Two-Sample Testing that account for graph-structured data (de la Concha et al., 2023, 2025).

Organization of the paper. Sec. 2 introduces the Collaborative LRE problem and provides the building elements of our framework. Sec. 3 presents the main technical contribution of the paper, the GRULSIF method. Sec. 4 provides theoretical guarantees on GRULSIF's excess risk. Sec. 5 discusses the elements allowing an efficient GRULSIF implementation. Sec. 6 illustrates the performance of GRULSIF in experiments on synthetic and real-life data. Our concluding remarks follow in Sec. 7. Technical details are in the Appendix.

2. Problem definition and background

In this section, we start by giving general notations and a problem statement. Then, we provide a brief background for each of the diverse technical elements we combine in this work, while at the same time justifying a series of choices that have been made.

General notations. Let a_i be the i-th entry of a vector a; when the vector is itself indexed by j, we refer to its i-th entry by $a_{j,i}$. A_{ij} denotes the entry at the i-th row and j-th column of a matrix A, and $A_{i,:}$ is its i-th row. $\mathbf{1}_M$ and $\mathbf{0}_M$ represent vectors with M ones and zeros, respectively, I_M is the $M \times M$ identity matrix. The Euclidean norm and dot product appear as $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$, and when referring to a functional space $\mathbb{H}$ they are indexed as $\|\cdot\|_{\mathbb{H}}$ and $\langle \cdot, \cdot \rangle_{\mathbb{H}}$. We consider a fixed, undirected and positive-weighted graph $G = (V, E, W)$, without self-loops, defined by the set V containing N nodes, the set of edges E, and a non-negative weight matrix $W \in \mathbb{R}^{N \times N}$ such that $W_{uu} = 0$, $\forall u \in V$, and $W_{uv} = W_{vu} \geq 0$. The set containing the neighbors of node v is $\mathrm{ng}(v) = \{u : W_{uv} \neq 0\}$, and the node degree is $d_v = |\mathrm{ng}(v)| = \sum_u \mathbb{1}\{W_{uv} \neq 0\}$, where $|\mathrm{set}| \in \mathbb{N}^*$ is the cardinality of a set, and $\mathbb{1}\{\mathrm{condition}\} \in \{0, 1\}$ is the indicator function. The combinatorial graph Laplacian operator is defined by $\mathcal{L} = \mathrm{diag}((d_v)_{v \in V}) - W$, where $\mathrm{diag}(\cdot)$ is a diagonal matrix with the input elements in its diagonal. In the rest, composite objects (vectors, matrices, sets, etc.) referring to all the nodes of a graph appear in bold font.

2.1 Collaborative likelihood-ratio estimation (LRE): Problem statement

In a given fixed, undirected, and positive-weighted graph $G = (V, E, W)$, suppose each node v has independent and identically distributed (i.i.d.) observations from two unknown probability measures P_v and Q_v (see the formal Assumption 2): n_v observations from P_v, and respectively n'_v others from Q_v. Let the measures P_v and Q_v admit the pdfs p_v and

q_v with respect to (w.r.t.) a reference measure ρ over a common input space $\mathcal{X}$; then, the available observations are defined as:

$$\begin{cases} \mathbf{X} &= \{\mathbf{X}_v\}_{v\in V} = \big\{\{x_{v,1},...,x_{v,n_v}\}\big\}_{v\in V}, \quad \forall v,i: \;\; x_{v,i} \overset{\text{i.i.d.}}{\sim} p_v, \;\; x_{v,i}\in\mathcal{X}; \\ \mathbf{X}' &= \{\mathbf{X}'_v\}_{v\in V} = \big\{\{x'_{v,1},...,x'_{v,n'_v}\}\big\}_{v\in V}, \quad \forall v,i: \;\; x'_{v,i} \overset{\text{i.i.d.}}{\sim} q_v, \;\; x'_{v,i}\in\mathcal{X}. \end{cases} \tag{1}$$

Consider a LRE task at each node v, quantifying how different p_v and q_v are by learning a node-specific model f_v that approximates the likelihood-ratio function. The Collaborative LRE problem aims to learn jointly all the N models by leveraging the similarity between tasks of adjacent nodes (see Fig. 1). Note that for technical reasons, explained in Sec. 2.3 and later in the paper, instead of the typical likelihood-ratio function $r_v(\cdot)=\frac{q_v(\cdot)}{p_v(\cdot)}$, we propose to approximate the α-relative likelihood-ratio function expressed in Eq. 2.

2.2 Likelihood-ratio functions

Let $(\mathcal{X},\Xi)$ be a measurable space. We denote by $\mathcal{M}^+_\sigma(\mathcal{X})$ the set of positive σ-finite measures over $(\mathcal{X},\Xi)$, $\mathcal{M}(\mathcal{X})$ refers to the space of all finite signed measures, and $\mathcal{P}(\mathcal{X})$ is the set of probability measures. Any $\rho\in\mathcal{M}(\mathcal{X})$ admits a unique decomposition as the difference of two positive measures, $\rho=\rho^+-\rho^-$, where at least one of them is finite (Hahn decomposition theorem). Then, the total variation measure of $\rho\in\mathcal{M}(\mathcal{X})$ is defined as $|\rho|=\rho^++\rho^-$.

For two probability measures $P,Q\in\mathcal{P}(\mathcal{X})$, the Radon-Nikodym derivative $r=\frac{dQ}{dP}(x)$ exists iff $Q\ll P$ (consequence of the Radon-Nikodym theorem). When both P,Q admit a pdf, p and q, w.r.t. a reference measure ρ, the Radon-Nikodym derivative is rewritten as $r(x)=\frac{dQ}{dP}(x)=\frac{q(x)}{p(x)}$. The quotient $\frac{q(x)}{p(x)}$ is known as the *likelihood-ratio* or *density-ratio*.

In practice, Q may not be absolutely continuous w.r.t. P (i.e. $Q\not\ll P$), which implies that the Radon-Nikodym derivative does not exist and hence the LRE problem is ill-defined. A flexible workaround is the α-regularization that chooses a value $0\leq\alpha\leq 1$ and introduces the convex combination of P and Q as a composite probability measure: $P^\alpha=(1-\alpha)P+\alpha Q$. An advantage is that setting $\alpha>0$ ensures $Q\ll P^\alpha$; in addition, when P and Q admit a pdf p and q, the likelihood-ratio is replaced by the *α-relative likelihood-ratio function* (Yamada et al., 2011), $r^\alpha:\mathcal{X}\to\mathbb{R}$, indexed by α:

$$r^\alpha(x)=\frac{q(x)}{(1-\alpha)p(x)+\alpha q(x)}<\frac{1}{\alpha}, \quad \text{for any } 0<\alpha<1. \tag{2}$$

Notably, when $\alpha>0$, the ratio r^α is always bounded above by $1/\alpha$ and does not suffer from numerical problems typically faced when trying to approximate an unbounded function.

2.3 Likelihood-ratio estimation with ϕ-divergences

ϕ-divergence. ϕ-divergences offer a unified view to various statistical problems, as they can be used both for the quantification of the dissimilarity between two probability measures P and Q and for training probabilistic models. In the LRE context, variational representations of ϕ-divergences have been used to define valid surrogate cost functions such that estimating the likelihood-ratio to amount to solving an optimization problem defined in a functional space $\mathcal{F}$. This has the advantage of minimizing the required assumptions on P and Q, and allows the direct approximation of the likelihood-ratio. In most applications, for efficient

estimation it is more intuitive to impose regularity conditions on the likelihood-ratio via the geometry of the functional space $\mathcal{F}$, rather than specifying explicitly P and Q.

Formally, a ϕ-divergence is a positive measure that quantifies the dissimilarity between two probability measured P and Q defined over an input space $\mathcal{X}$:

$$\mathcal{D}_\phi(P\|Q) = \begin{cases} \int \phi\Big(\frac{dQ}{dP}\Big)(x)dP(x), & \text{if } Q \ll P; \\ +\infty, & \text{if } Q \not\ll P, \end{cases} \tag{3}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a proper closed convex function from $(-\infty, \infty)$ to $[0, \infty]$ with $\phi(1) = 0$, and such that its domain $\mathrm{Dom}(\phi) = \{x \in \mathbb{R} \,|\, \phi(x) < \infty\}$ is an interval with endpoints $a_\phi < 1 < b_\phi$ (Csiszár, 1967; Broniatowski and Keziou, 2006; Birrell et al., 2022a). If $\mathcal{D}_\phi(P\|Q) \geq 0$ holds, and if ϕ is strictly convex at 1, then $\mathcal{D}_\phi(P\|Q) = 0$ iff $P = Q$. Notably, for $z \in \mathbb{R}$, setting $\phi(z) = -\log(z)$ recovers the KL-divergence (Kullback, 1959), while $\phi(z) = \frac{1}{2}(z-1)^2$ recovers Pearson's χ^2-divergence (Pearson, 1900).

Variational representation of ϕ-divergences. Deriving valid variational representations is an active research topic and there are several formulations built using different assumptions (Broniatowski and Keziou, 2006; Nguyen et al., 2008, 2010; Agrawal and Horel, 2021; Birrell et al., 2022a,b; Bach, 2024). In this work, we focus on the following result.

Theorem 1 *(Theorem 4.3 in Broniatowski and Keziou (2006)). Let $\mathcal{F}$ be some class of measurable real-valued functions defined on $\mathcal{X}$, $\mathcal{B}$ the set of all bounded measurable real-valued functions defined on $\mathcal{X}$, and* $\mathrm{span}(\mathcal{F} \cup \mathcal{B})$ *the set of all linear combinations of the set $\mathcal{F} \cup \mathcal{B}$; let also the set:*

$$\mathcal{M}_\mathcal{F} = \left\{ Q \in \mathcal{M} \;\middle|\; \int |f| \, d|Q| < \infty, \; \forall f \in \mathcal{F} \right\}, \tag{4}$$

where $|Q|$ denotes the total variation of the signed finite measure Q.

Assume that ϕ is differentiable. Then, for all $Q \in \mathcal{M}_\mathcal{F}$, such that $\mathcal{D}_\phi(P\|Q)$ is finite and $\phi'(\frac{dQ}{dP})$ belongs to $\mathrm{span}(\mathcal{F} \cup \mathcal{B})$*, $\mathcal{D}_\phi(P\|Q)$ admits the dual representation:*

$$\mathcal{D}_\phi(P\|Q) = \sup_{g \in \mathrm{span}(\mathcal{F} \cup \mathcal{B})} \int g(x')dQ(x') - \int \phi^\star(g)(x) \, dP(x) \tag{5}$$

where $\phi^\star$ denotes the convex conjugate of ϕ. And the function $g^* = \phi'(\frac{dQ}{dP})$ is a dual optimal solution. Furthermore, if ϕ is essentially smooth, then g^* is the unique dual solution P-almost everywhere.

Solving Problem 5 to estimate the ϕ-divergence, depends on the functional space $\mathcal{F}$ and the set defined by the subdifferential of ϕ, which is evaluated at the likelihood-ratio $r(x)$ for $x \in \mathcal{X}$. This link has been exploited to define convex functional optimization problems that first estimate the likelihood-ratio and then the associated ϕ-divergence (Nguyen et al., 2008, 2010; Sugiyama et al., 2012). This approach makes no hypothesis about the form of q and p, hence leads to non-parametric algorithms that only need data observations coming from p and q. Note that we write $\mathcal{D}_\phi^\alpha(P\|Q) := \mathcal{D}_\phi(P^\alpha\|Q)$ to define a ϕ-divergence in the form of Eq. 3, but in terms of the relative likelihood-ratio r^α. Moreover, $\mathcal{D}_\phi^\alpha(P\|Q)$ is a valid

dissimilarity function, since $\mathcal{D}_\phi^\alpha(P\|Q) \geq 0$, and when ϕ is strictly convex around 1 it can be verified that $\mathcal{D}_\phi^\alpha(P\|Q) = 0$ iff $P = Q$. Then, the result of Theorem 1 links the variational formulation of $\mathcal{D}_\phi^\alpha(P\|Q)$ and the relative likelihood-ratio r^α.

To address the Collaborative LRE problem stated in Sec. 2.1, we define a surrogate cost function based on the variational representation of $\mathcal{D}_\phi^\alpha(P_v\|Q_v)$ to jointly estimate all $\{r_v^\alpha\}_{v\in V}$, i.e. one (relative) likelihood-ratio function $r_v^\alpha(x) = \frac{q_v(x)}{(1-\alpha)p_v(x)+\alpha q_v(x)} \in \mathbb{R}$ for each node over a vector-valued functional space $\mathbb{G}$. We choose a normed vector space $\mathbb{G}$ whose norm encodes the given information regarding the similarity between the graph nodes, which in this case amounts to the similarity between their likelihood-ratio functions $\{r_v^\alpha\}_{v\in V}$.

2.4 Non-parametric estimation

LRE can be addressed using different functional spaces, e.g. Neural Networks (Nowozin et al., 2016; Rhodes et al., 2020; Kato and Teshima, 2021) or Reproducing Kernel Hilbert Spaces (RKHSs) (Sugiyama et al., 2007; Nguyen et al., 2010; Yamada et al., 2013; Kanamori et al., 2011; Zellinger et al., 2023; de la Concha et al., 2024; Nguyen et al., 2024). The choice depends on the availability of data, the computational resources, or other prior information about the likelihood-ratio. In this work, we focus on RKHSs as they offer numerous comparative advantages: they provide geometrical operations defined in Hilbert spaces that facilitate the estimation and theoretical analysis; they allow us to learn in rich infinite-dimensional spaces, and the complexity of the approximated function can be elegantly expressed by the norm $\|\cdot\|_{\mathbb{H}}$.

Traditional Kernel Methods model scalar functions in a RKHS space associated with a positive definite kernel. In our context, we would like to approximate the vector-valued function $\mathbf{r}^\alpha(\cdot) = (r_1^\alpha(\cdot), ..., r_N^\alpha(\cdot)) \in \mathbb{H}^N$, where each dimension is associated with a node of the graph. Moreover, we would like the functional space to be rich enough to approximate all those N likelihood-ratios. This is possible via a Vector-Valued Reproducing Kernel Hilbert Space (VV-RKHS), which is a multivariate generalization of the scalar RKHS (Micchelli and Pontil, 2005; Carmeli et al., 2006; Álvarez et al., 2012). Formal definitions follow.

Scalar RKHS. Let $\mathcal{X}$ be a set, and $\mathbb{H} \subset \mathbb{R}^{\mathcal{X}}$ a class of functions forming a real Hilbert space with inner-product $\langle\cdot,\cdot\rangle_{\mathbb{H}}$. $\mathrm{K} : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is a reproducing kernel function of $\mathbb{H}$ if:

1. $\mathbb{H}$ contains all functions of the form: $\forall x \in \mathcal{X}, \mathrm{K}(x,\cdot) : t \to \mathrm{K}(x,t)$.
2. For every $x \in \mathcal{X}$ and $f \in \mathbb{H}$ the reproducing property holds: $f(x) = \langle f, \mathrm{K}(x,\cdot)\rangle_{\mathbb{H}}$.

If a reproducing kernel K exists, then $\mathbb{H}$ is called a RKHS. An RKHS has a unique reproducing kernel, and conversely, a function K describes at most one RKHS.

Vector-valued Kernels and associated RKHS. A positive vector-valued kernel in $\mathbb{R}^N$ on $\mathcal{X}\times\mathcal{X}$ is a map $\mathbf{\Gamma} : \mathcal{X}\times\mathcal{X} \to \mathbb{R}^{N\times N}$ such that, for all $n \in \mathbb{N}$, $x_1, ..., x_n \in \mathcal{X}$ and $a_1, ..., a_n \in \mathbb{C}$:

$$\sum_{i,j=1}^{n} a_i \bar{a}_j \langle \mathbf{\Gamma}(x_i, x_j)c, c\rangle \geq 0, \quad \forall c \in \mathbb{R}^N. \tag{6}$$

As in the scalar case, the positive vector-valued kernel is associated with a unique vector-valued RKHS (VV-RKHS) denoted by $\mathbb{G}$ (see Theorem 1 in Micchelli and Pontil (2005)).

However, the conditions to characterize the associated functional spaces are different in this case (Micchelli and Pontil, 2005; Carmeli et al., 2006), more precisely it is required:

1. For every $c \in \mathbb{R}^N$ and $x \in \mathcal{X}$: $\mathbf{\Gamma}(x,\cdot)c \in \mathbb{G}$.
2. For every $\mathbf{f} \in \mathbb{G}$: $\langle \mathbf{f}, \mathbf{\Gamma}(x,\cdot)c \rangle_{\mathbb{G}} = \langle \mathbf{f}(x), c \rangle$.

In this case, the reproducing kernel function returns a matrix in $\mathbb{R}^{N\times N}$ (instead of a scalar), and the elements of the associated Hilbert space are vector-valued functions $\mathbf{f} : \mathcal{X} \to \mathbb{R}^N$.

2.5 Graph functions and graph smoothness

A *graph function* $\vartheta : V \to \mathcal{Y}$ assigns to each node of a graph an element of a given metric space $\mathcal{Y}$. When $\mathcal{Y} = \mathbb{R}$, $\vartheta = (\vartheta_1, ..., \vartheta_N) \in \mathbb{R}^N$ is also called *graph signal* (Shuman et al., 2013). The smoothness of ϑ w.r.t. a graph is defined as:

$$S(\vartheta) = \sum_{v \in V} \sum_{u \in \mathrm{ng}(v)} W_{uv} (\vartheta(u) - \vartheta(v))^2. \tag{7}$$

For the needs of our discussion, we introduce a generalization of this notion for $\mathcal{Y} = \mathbb{H}$, formally for graph functions $\vartheta(\cdot) = (\vartheta_1(\cdot), ..., \vartheta_N(\cdot)) \in \mathbb{H}^N$, which implies $\vartheta_v(\cdot) \in \mathbb{H}$, $\forall v \in V$:

$$S(\vartheta(\cdot)) = \sum_{v \in V} \sum_{u \in \mathrm{ng}(v)} W_{uv} \left\| \vartheta_u(\cdot) - \vartheta_v(\cdot) \right\|_{\mathbb{H}}^2. \tag{8}$$

The lower $S(\vartheta(\cdot))$ is, the smoother we say the function $\vartheta(\cdot)$ is w.r.t. the graph. Smoothness formalizes the idea that two adjacent nodes u and v have similar functions $\vartheta_u(\cdot)$, $\vartheta_v(\cdot)$.

Notice, for any $x \in \mathcal{X}$, $\vartheta(x) = (\vartheta_1(x), ..., \vartheta_N(x)) \in \mathbb{R}^N$ gives a graph signal. Moreover, when $\mathbb{H}$ is a scalar RKHS with a bounded kernel, e.g. $\exists C > 0$ s.t. $\sup_{x \in \mathcal{X}} \sqrt{\mathrm{K}(x,x)} \leq C < \infty$, the smoothness of the graph function $\vartheta(\cdot)$ w.r.t. the norm $\|\cdot\|_{\mathbb{H}}$ (Expr. 8) implies also the smoothness of the graph signal $\vartheta(x)$ in the classical sense (Expr. 7). More clearly, for $x \in \mathcal{X}$:

$$\begin{aligned}
(\vartheta_u(x) - \vartheta_v(x))^2 &= \left| \langle [\vartheta_u(\cdot) - \vartheta_v(\cdot)], \mathrm{K}(x,\cdot) \rangle_{\mathbb{H}} \right|^2 && \text{(Reproducing property of } \mathbb{H}\text{)} \\
&\leq \left\| \mathrm{K}(x,\cdot) \right\|_{\mathbb{H}}^2 \left\| \vartheta_u(\cdot) - \vartheta_v(\cdot) \right\|_{\mathbb{H}}^2 && \text{(Cauchy-Schwarz inequality)} \\
&\leq C^2 \left\| \vartheta_u(\cdot) - \vartheta_v(\cdot) \right\|_{\mathbb{H}}^2. && \text{(Kernel boundedness)}
\end{aligned} \tag{9}$$

$$\begin{aligned}
\Rightarrow\ S(\vartheta(x)) &= \sum_{v \in V} \sum_{u \in \mathrm{ng}(v)} W_{uv} (\vartheta_u(x) - \vartheta_v(x))^2 \\
&\leq C^2 \sum_{v \in V} \sum_{u \in \mathrm{ng}(v)} W_{uv} \left\| \vartheta_u(\cdot) - \vartheta_v(\cdot) \right\|_{\mathbb{H}}^2 = C^2 S(\vartheta(\cdot)).
\end{aligned} \tag{10}$$

Further comments on what graph smoothness quantifies in a Multitask Learning context are provided in Sec. 3B.

3. Graph-based Relative Unconstrained Least-Squares Importance Fitting (GRULSIF)

The proposed Collaborative LRE framework estimates jointly the N likelihood-ratios at the nodes of a graph (see Fig. 1), in a collaborative and distributed manner. We approximate

each node's r_v^α with a function $f_v \in \mathbb{H}$. Note that $\mathbf{r}^\alpha = (r_1, ..., r_N)$ defines a graph function, which we assume to be smooth over the graph (see Sec. 2.5). This essentially suggests that two adjacent nodes, u and v, generally exhibit similar likelihood-ratios, r_u^α and r_v^α. This in turn means that the learned models, f_u and f_v, will be close w.r.t. the norm $\|\cdot\|_{\mathbb{H}}$ and, as a consequence, will have similar outputs $f_u(x)$ and $f_v(x)$ for an input point $x \in \mathcal{X}$. Notice that, by the likelihood-ratio definition, this hypothesis is true when $p_v = q_v$, $\forall v$, even if there is heterogeneity among the nodes (generally, $p_v \neq p_u$). The latter is the basis of our approach, taking inspiration from the RULSIF method of Yamada et al. (2011), hence named *Graph-based Relative Unconstrained Least-Squares Importance Fitting* (GRULSIF). Our estimation strategy capitalizes over the elements described in Sec. 2 as follows:

1. The variational representation of Theorem 1 will allow us to define a functional optimization problem at the node level, aiming to approximate the relative likelihood-ratio r_v^α, while requiring minimal hypotheses for $\{p_v\}_{v \in V}$ and $\{q_v\}_{v \in V}$.
2. The concept of graph smoothness and VV-RKHS will encode the geometry of the problem, and will formalize a *collaborative estimation procedure*.
3. The properties of VV-RKHS, more precisely the Representer theorem, will provide the required elements to translate the optimization problem from a potentially infinite-dimensional space into a simple optimization problem in $\mathbb{R}^L$, where L is the total number of available observations from all the nodes. Moreover, this approach will lead to efficient likelihood-ratio estimators, $\hat{f}_v$, that can be evaluated at any point $x \in \mathcal{X}$ by just computing a dot product in $\mathbb{R}^L$.

This line of reasoning is general enough to be applicable to any ϕ-divergences to produce likelihood-ratio estimates that account for a graph structure. However, for the rest of the paper, we will focus on the Pearson's χ^2-divergence. The main reason for using this specific ϕ-divergence is that collaborative LRE takes the form of an unconstrained penalized least-squares problem. Moreover, the likelihood-ratio estimates are the solution to a linear system. Leveraging these features, we can seamlessly adapt existing and efficient optimization techniques tailored for penalized least-squares, and hence integrate a mature theoretical framework to gain insight into the properties of the estimators. Such advantages may not be offered or be readily available for other ϕ-divergences.

A) Node-level relative likelihood-ratio estimation

At each node v, we approximate the relative likelihood-ratio r_v^α via the variational representation of the χ^2-divergence. To explain the properties of the resulting surrogate cost function, we focus on estimating a single relative likelihood-ratio r^α (i.e. $N = 1$) by the elements of a scalar RKHS $\mathbb{H}$, equipped with a scalar reproducing kernel K and a feature map $\varphi(\cdot)$ (see Sec. 2.4). To formally define the problem, we require $\mathbb{H}$ to satisfy the following standard hypothesis of the Kernel Methods literature.

Assumption 1 ***(Kernel boundedness).*** *The reproducing kernel map* $\mathrm{K}(\cdot, \cdot)$ *is measurable and can be upper-bounded by a constant* $C > 0$*:*

$$\sup_{x \in \mathcal{X}} \sqrt{\mathrm{K}(x, x)} \leq C < \infty. \tag{11}$$

This is satisfied by popular kernels, such as the Gaussian and the Laplacian kernels, and in general, for continuous kernel maps $\mathrm{K}(\cdot,\cdot)$ defined in a compact input space $\mathcal{X}$.

Theorem 2 states the variational representation of the χ^2-divergence when $\mathcal{F}=\mathbb{H}$. The result is a direct consequence of Theorem 1, and its proof is provided in Appendix A.

Theorem 2 *Under Assumption 1, the variational formulation of the χ^2-divergence between P^α and Q takes the form:*

$$PE(P^\alpha\|Q)=\int\frac{(r^\alpha(y)-1)^2}{2}dP^\alpha(y)\geq\sup_{f\in\mathbb{H}}\int f(x')\,dQ(x')-\int\frac{f^2(y)}{2}\,dP^\alpha(y)-\frac{1}{2},\qquad(12)$$

where f is an approximation of the likelihood-ratio r^α.

The optimization problem appearing in Expr. 12 can be interpreted as a least-squares approximation of r^α. This becomes more evident in Problem 13 where we state the LRE as a least-square optimization problem in RKHS with respect to the norm of the space of square-integrable functions w.r.t. P^α, denoted by $\mathcal{L}^2(P^\alpha)$:

$$\begin{aligned}\inf_{f\in\mathbb{H}}\hat{L}&=\inf_{f\in\mathbb{H}}\int\frac{f^2(y)}{2}\,dP^\alpha(y)-\int f(x')\,dQ(x')\\&\approx\inf_{f\in\mathbb{H}}\int\frac{f^2(y)}{2}\,dP^\alpha(y)-\int f(y)r^\alpha(y)\,dP^\alpha(y)+\int\frac{(r^\alpha(y))^2}{2}\,dP^\alpha(y)\qquad(13)\\&=\inf_{f\in\mathbb{H}}\int\frac{[f-r^\alpha]^2(y)}{2}dP^\alpha(y),\end{aligned}$$

where '$\approx$' indicates that both optimization problems are equivalent up to some constant terms. The second equality is brought by $\mathbb{E}_{p^\alpha(y)}[r^\alpha(y)^2]=C_1$, where $C_1<\infty$ is a constant, and the change of measure identity $\mathbb{E}_{p^\alpha(y)}[g(y)r^\alpha(y)]=\mathbb{E}_{q(x')}[g(x')]$.

For Problem 13 to be well-defined, the optimal solution f^* should exist and $f^*\in\mathbb{H}$, as well as f^* should coincide with the relative likelihood-ratio r^α we want to estimate. To ensure this, notice that Assumption 1 implies $\mathbb{H}\subset\mathcal{L}^2$:

$$\begin{aligned}\mathbb{E}_{p^\alpha(y)}[f^2(y)]&=\mathbb{E}_{p^\alpha(y)}[(\langle f,\mathrm{K}(y,\cdot)\rangle_{\mathbb{H}})^2]&&\text{(Representer property)}\\&\leq\|f\|^2_{\mathbb{H}}\,\mathbb{E}_{p^\alpha(y)}[\|\mathrm{K}(y,\cdot)\|^2_{\mathbb{H}}]&&\text{(Cauchy–Schwarz inequality)}\\&\leq C^2\,\|f\|^2_{\mathbb{H}}<\infty.&&\text{(Assumption 1)}\end{aligned}$$

For a well-defined model, i.e. $r^\alpha\in\mathbb{H}$, it is clear that $f^*=r^\alpha$ is P^α-almost everywhere.

Moreover, it is worth to comment on whether this model specification constitutes a restrictive hypothesis in the usual LRE context. Notice that the fact that r^α is an upper-bounded function implies $r^\alpha\in\mathcal{L}^2(P^\alpha)$. Then, the elements of a RKHS $\mathbb{H}$ that is dense w.r.t. the $\mathcal{L}^2(P^\alpha)$ norm can approximate as much as we want r^α. Most LRE methods deal with the case $\mathcal{X}\subset\mathbb{R}^d$, where it can be shown that translation invariant kernels (e.g. Gaussian, Laplacian, and the Matérn class of kernels) are universal kernels that can approximate any function in $\mathcal{L}^2(P^\alpha)$. Similar results can be obtained when $\mathcal{X}$ is a locally compact Hausdorff space (Sriperumbudur et al., 2011).

To conclude, the approximation properties of RKHSs along with the α-regularization imply that the proposed Problem 12 is a powerful functional optimization technique for approximating relative likelihood-ratios. Moreover, under the hypothesis $r^\alpha\in\mathbb{H}$, the approximation error would be zero.

B) Multitasking formulation of the LRE over graphs

Our goal is to estimate the vector-valued function $\mathbf{r}^\alpha = (r_1^\alpha, ..., r_N^\alpha)$. As in the previous section, we need a surrogate cost function leading to an optimization problem with nice numerical properties and a functional space with good approximation properties. Moreover, we would like to leverage pior knowledge about the similarities between the relative likelihood-ratios to obtain better generalization. Thus, the LRE problem based on χ^2-divergence, motivates a multitasking formulation of the relative likelihood-ratios estimation over a graph, through the following objective function:

$$\begin{aligned}
&\underset{\{f_v\}_{v\in V}\in\mathbb{H}^N}{\operatorname{argmin}} \frac{1}{N}\sum_{v\in V}\left(\frac{1}{2}\mathbb{E}_{p_v^\alpha(y)}[[r_v^\alpha - f_v]^2(y)]\right) + \frac{\lambda}{4}\sum_{u,v\in V} W_{uv}\,\|f_u - f_v\|_{\mathbb{H}}^2 + \frac{\lambda\gamma}{2}\sum_{v\in V}\|f_v\|_{\mathbb{H}}^2 \\
= &\underset{\{f_v\}_{v\in V}\in\mathbb{H}^N}{\operatorname{argmin}} \frac{1}{N}\sum_{v\in V}\left(\frac{1}{2}\mathbb{E}_{p_v^\alpha(y)}[f_v^2(x)] - \mathbb{E}_{q_v(x')}[f_v(x)]\right) + \frac{\lambda}{4}\sum_{u,v\in V} W_{uv}\,\|f_u - f_v\|_{\mathbb{H}}^2 + \frac{\lambda\gamma}{2}\sum_{v\in V}\|f_v\|_{\mathbb{H}}^2 .
\end{aligned} \tag{14}$$

The first term of the objective function is a loss asking for a good approximation at each node; the second term introduces our hypothesis that adjacent nodes are expected to have similar likelihood-ratios; the third term is a penalization term aiming to reduce the risk of overfitting (Sheldon, 2008), where $\lambda, \gamma > 0$ are penalization coefficients.

Let us now define the vector-valued kernel:

$$\mathbf{\Gamma}(x, x') = \mathrm{K}(x, x')(\mathcal{L} + \gamma I_N)^{-1} \in \mathbb{R}^{N\times N}. \tag{15}$$

Given the properties of the graph Laplacian $\mathcal{L}$, it can be shown that $\mathbf{\Gamma}(\cdot,\cdot)$ is a positive vector-valued kernel inducing a VV-RKHS $\mathbb{G}$, in which the norm of any $\mathbf{f} \in \mathbb{G}$ is defined as:

$$\|\mathbf{f}\|_{\mathbb{G}}^2 = \frac{1}{2}\sum_{u,v\in V} W_{uv}\,\|f_u - f_v\|_{\mathbb{H}}^2 + \gamma\sum_{v\in V}\|f_v\|_{\mathbb{H}}^2 . \tag{16}$$

Notice that the norm in $\mathbb{G}$ incorporates *both* the geometry induced by the structure of the graph Laplacian $\mathcal{L}$ and the geometry of the scalar RKHS $\mathbb{H}$.

As explained in Sec. 2.1, we assume that there is access to two samples, $\mathbf{X}$ and $\mathbf{X}'$. Then, the optimization Problem 14 can be written as a penalized empirical risk minimization problem in terms of the elements of $\mathbb{G}$, the vector-valued functions $\mathbf{f} = (f_1, ..., f_N)$:

$$\min_{\mathbf{f}\in\mathbb{G}} \frac{1}{N}\sum_{v\in V}\left(\frac{1-\alpha}{2n_v}\sum_{i=1}^{n_v} f_v^2(x_{v,i}) + \frac{\alpha}{2n'_v}\sum_{i=1}^{n'_v} f_v^2(x'_{v,i}) - \frac{1}{n'_v}\sum_{i=1}^{n'_v} f_v(x'_{v,i})\right) + \frac{\lambda}{2}\|\mathbf{f}\|_{\mathbb{G}}^2 . \tag{17}$$

The VV-RKHS formulation allows us to apply the Representer theorem (Theorem 5 in Micchelli and Pontil (2005)), meaning that the solution to Problem 17 takes the form:

$$\hat{\mathbf{f}}(\cdot) = \sum_{i=1}^{L}\mathbf{\Gamma}(x_i, \cdot)c_i = \sum_{i=1}^{L}\mathrm{K}(x_i, \cdot)(\mathcal{L} + \gamma I_N)^{-1}c_i, \tag{18}$$

where $L = \sum_{v\in V}(n_v + n'_v)$ is the total number of observations in all nodes, and $c_i \in \mathbb{R}^N$, $i = 1, ..., L$. The second equality comes from Eq. 15. More specifically, the node-level ap-

proximation takes now the form:

$$\hat{f}_v(\cdot) = \sum_{i=1}^{L} \mathrm{K}(\cdot, x_i)[(\mathcal{L}+I_N)^{-1}]_{v,:}c_i = \sum_{i=1}^{L} \mathrm{K}(\cdot, x_i)\theta_{v,i} = \varphi(\cdot)^{\mathsf{T}}\theta_v, \tag{19}$$

where, for the second equality we have defined $\theta_{v,i} = [(\mathcal{L}+I_N)^{-1}]_{v,:}c_i$, we define the feature map w.r.t. all observations as the function $\varphi : \mathcal{X} \to \mathbb{R}^L$, $\varphi(x) = (\mathrm{K}(x,x_1),...,\mathrm{K}(x,x_L))^{\mathsf{T}} \in \mathbb{R}^L$. The last equality uses $\theta_v = (\theta_{v,1},...,\theta_{v,L})^{\mathsf{T}} \in \mathbb{R}^L$, which is an abuse of notation that is helpful for the presentation. By the definition of $\hat{f}_v(\cdot)$, its norm in $\mathbb{H}$ can be elegantly written as:

$$\|\hat{f}_v\|^2_{\mathbb{H}} = \theta_v^{\mathsf{T}}\mathcal{K}\theta_v, \tag{20}$$

where $\mathcal{K} \in \mathbb{R}^{L\times L}$ is the Gram matrix associated with the scalar kernel function $\mathrm{K}(\cdot,\cdot)$. We conclude from Expr. 19 that approximating $(r_1^\alpha,...,r_N^\alpha)$ amounts to estimating the node parameters θ_v, for all $v \in V$.

Further remarks on graph smoothness. A clearer interpretation of graph smoothness in multitasking, can be given by developing the penalization term introduced in Expr. 16:

$$\begin{aligned}
\|\mathbf{f}\|^2_{\mathbb{G}} &= \frac{1}{2}\sum_{u,v\in V} W_{uv}\left(\|f_u\|^2_{\mathbb{H}} + \|f_v\|^2_{\mathbb{H}} - 2\langle f_v, f_u\rangle_{\mathbb{H}}\right) + \gamma\sum_{v\in V}\|f_v\|^2_{\mathbb{H}} \\
&= \sum_{v\in V}\sum_{u\le v} W_{uv}\left(\|f_u\|^2_{\mathbb{H}} + \|f_v\|^2_{\mathbb{H}} - 2\langle f_v, f_u\rangle_{\mathbb{H}}\right) + \gamma\sum_{v\in V}\|f_v\|^2_{\mathbb{H}} \\
&= \sum_{v\in V}(d_v+\gamma)\|f_v\|^2_{\mathbb{H}} - 2\sum_{u\le v} W_{uv}\langle f_v, f_u\rangle_{\mathbb{H}}.
\end{aligned} \tag{21}$$

By hypothesis, we have $W_{uv} \ge 0$, which implies that $\|\mathbf{f}\|^2_{\mathbb{G}}$ decreases as the dot product $\langle f_v, f_u\rangle_{\mathbb{H}}$ of the functions associated with connected nodes increases. This dot product can be interpreted as a similarity measure between the functions $f_u, f_v \in \mathbb{H}$ w.r.t. the geometry of the underlying scalar RKHS. These observations imply that a vector-valued function $\mathbf{f} = (f_1,...,f_N)$ becomes smoother w.r.t. the graph as the functions associated to connected nodes become more similar w.r.t. the geometry of the RKHS.

The exact meaning of the similarity measure $\langle f_v, f_u\rangle_{\mathbb{H}}$ depends on the chosen kernel. For example, consider the case of the input space $\mathcal{X} = \mathbb{R}^d$ and a translation-invariant positive definite kernel $k(x,t) = \psi(x-t)$, where ψ and its Fourier transform $\hat{\phi}$ are integrable functions in $\mathbb{R}^d$. In this setting, each element of the scalar RKHS is an integrable and continuous function f whose norm is given by:

$$\|f\|^2_{\mathbb{H}} = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \frac{|\hat{f}(w)|^2}{\hat{\phi}(w)} dw < +\infty. \tag{22}$$

Then, the similarity $\langle f_v, f_u\rangle_{\mathbb{H}}$ can be computed in terms of the Fourier transforms $\hat{f}_u$, $\hat{f}_v$:

$$\langle f_u, f_v\rangle_{\mathbb{H}} = \frac{1}{2(\pi)^d}\int_{\mathbb{R}^d} \frac{\hat{f}_u(w)\overline{\hat{f}_v(w)}}{\hat{\phi}(w)} dw. \tag{23}$$

For translation-invariant kernels, $\langle f_v, f_u\rangle_{\mathbb{H}}$ measures the similarity of both functions within the Fourier domain, weighted by $\frac{1}{\hat{\phi}(w)}$.

C) LRE as a quadratic problem in $\mathbb{R}^{NL}$

Let $\mathbf{\Theta} = \text{vec}(\theta_1^\mathsf{T}, ..., \theta_N^\mathsf{T})^\mathsf{T} \in \mathbb{R}^{NL}$ be the concatenation of all node parameter vectors in a vector. Let us also introduce the following terms associated with a specific feature map (here, this is $\varphi(\cdot)$), which need to be computed only once at the beginning:

$$
\begin{aligned}
&H_v = \frac{1}{n_v} \sum_{x \in \mathbf{X}_v} \varphi(x)\varphi(x)^\mathsf{T} \in \mathbb{R}^{L \times L}, \quad H'_v = \frac{1}{n'_v} \sum_{x' \in \mathbf{X}'_v} \varphi(x')\varphi(x')^\mathsf{T} \in \mathbb{R}^{L \times L}, \quad h'_v = \frac{1}{n'_v} \sum_{x' \in \mathbf{X}'_v} \varphi(x') \in \mathbb{R}^{L}, \\
&\mathbf{H} = \text{block}(H_1, ..., H_N) \in \mathbb{R}^{LN \times LN}, \quad \mathbf{H}' = \text{block}(H'_1, ..., H'_N) \in \mathbb{R}^{LN \times LN}, \\
&\mathbf{h}' = \text{vec}(h'_1, h'_2, ..., h'_N)^\mathsf{T} \in \mathbb{R}^{LN}.
\end{aligned} \tag{24}
$$

Above, $\text{block}(H_1, ..., H_n)$ denotes a block diagonal matrix with each block corresponding to one of the square matrices $H_1, ..., H_n$, and $\text{vec}(h'_1, .., h'_n)$ denotes the concatenation of the input vectors $h'_1, ..., h'_n$ in a single vector. By putting everything together and by using Eq. 20, 24, we conclude that the functional optimization Problem 14 can be restated as a quadratic problem w.r.t. the vector $\mathbf{\Theta}$:

$$
\begin{aligned}
\min_{\mathbf{\Theta} \in \mathbb{R}^{NL}} \Phi(\mathbf{\Theta}) &= \min_{\mathbf{\Theta} \in \mathbb{R}^{NL}} \frac{1}{N} \sum_{v \in V} \left(\frac{1-\alpha}{2} \theta_v^\mathsf{T} H_v \theta_v + \frac{\alpha}{2} \theta_v^\mathsf{T} H'_v \theta_v - {h'}_v^\mathsf{T} \theta_v \right) \\
&\qquad + \frac{\lambda}{4} \sum_{u,v \in V} W_{uv} (\theta_v - \theta_u)^\mathsf{T} \mathcal{K} (\theta_v - \theta_u) + \frac{\lambda\gamma}{2} \sum_{v \in V} \theta_v^\mathsf{T} \mathcal{K} \theta_v \\
&= \min_{\mathbf{\Theta} \in \mathbb{R}^{NL}} \frac{1}{N} \left(\frac{1-\alpha}{2} \mathbf{\Theta}^\mathsf{T} \mathbf{H} \mathbf{\Theta} + \frac{\alpha}{2} \mathbf{\Theta}^\mathsf{T} \mathbf{H}' \mathbf{\Theta} - {\mathbf{h}'}^\mathsf{T} \mathbf{\Theta} \right) \\
&\qquad + \frac{\lambda}{2} \mathbf{\Theta}^\mathsf{T} (I_N \otimes \mathcal{K}^{\frac{1}{2}})^\mathsf{T} [\mathcal{L} \otimes I_L] (I_N \otimes \mathcal{K}^{\frac{1}{2}}) \mathbf{\Theta} \\
&\qquad + \frac{\lambda\gamma}{2} \mathbf{\Theta}^\mathsf{T} (I_N \otimes \mathcal{K}) \mathbf{\Theta} \\
\iff \min_{\mathbf{\Theta} \in \mathbb{R}^{NL}} \Phi(\mathbf{\Theta}) &= \min_{\mathbf{\Theta} \in \mathbb{R}^{NL}} \mathbf{\Theta}^\mathsf{T} \mathbf{A} \mathbf{\Theta} - \frac{1}{N} {\mathbf{h}'}^\mathsf{T} \mathbf{\Theta},
\end{aligned} \tag{25}
$$

where $\otimes$ is the Kronecker product of two matrices and:

$$
\mathbf{A} = \frac{1}{N} \left(\frac{1-\alpha}{2} \mathbf{H} + \frac{\alpha}{2} \mathbf{H}' \right) + \frac{\lambda}{2} (I_N \otimes \mathcal{K}^{\frac{1}{2}})^\mathsf{T} [\mathcal{L} \otimes I_L] (I_N \otimes \mathcal{K}^{\frac{1}{2}}) + \frac{\lambda\gamma}{2} (I_N \otimes \mathcal{K}). \tag{26}
$$

Notice that $\mathbf{A}$ is a semi-positive definite matrix given that $\mathcal{L}$ and $\mathcal{K}$ are semi-positive definite as well, which implies that Problem 25 is a quadratic optimization problem in $\mathbf{\Theta}$. We will exploit this fact in Sec. 5 to propose an efficient optimization procedure.

D) Pearson's χ^2-divergence estimation

We can use the estimated likelihood-ratio $\hat{f}_v$ from Eq. 19, and Eq. 24 to approximate the following expectation that corresponds to the loss $\ell_v(\theta_v)$ at node v:

$$\frac{1}{2}\mathbb{E}_{p_v^\alpha(y)}[f_v^2(y)] - \mathbb{E}_{q_v(x')}[f_v(x)] = \frac{1-\alpha}{2}\mathbb{E}_{p_v(x)}[f_v^2(x)] + \frac{\alpha}{2}\mathbb{E}_{q_v(x')}[f_v^2(x')] - \mathbb{E}_{q_v(x')}[f_v(x')] \quad (27)$$

$$\approx \frac{1-\alpha}{2}\left(\sum_{x\in X_v}\frac{\hat{f}_v(x)^2}{n_v}\right) + \frac{\alpha}{2}\left(\sum_{x'\in X'_v}\frac{\hat{f}_v(x')^2}{n'_v}\right) - \sum_{x'\in X'_v}\frac{\hat{f}_v(x')}{n'_v}$$

$$= \frac{1-\alpha}{2}\hat{\theta}_v^\mathsf{T} H_v \hat{\theta}_v + \frac{\alpha}{2}\hat{\theta}_v^\mathsf{T} H'_v \hat{\theta}_v - h_v'^\mathsf{T}\hat{\theta}_v \;=:\; \hat{L}_v(\hat{\theta}_v). \quad (28)$$

This expression leads to the more compact and convenient formulation of Problem 25:

$$\min_{\mathbf{\Theta}\in\mathbb{R}^{NL}} \frac{1}{N}\left(\sum_{v\in V}\hat{L}_v(\theta_v)\right) + \frac{\lambda}{2}\mathbf{\Theta}^\mathsf{T}(I_N\otimes\mathcal{K}^{\frac{1}{2}})^\mathsf{T}[\mathcal{L}\otimes I_L](I_N\otimes\mathcal{K}^{\frac{1}{2}})\mathbf{\Theta} + \frac{\lambda\gamma}{2}\mathbf{\Theta}^\mathsf{T}(I_N\otimes\mathcal{K})\mathbf{\Theta}. \quad (29)$$

Moreover, we use Eq. 12 to propose an approximation of $PE(p_v^\alpha\|q_v)$ based on the estimated parameters $\hat{\mathbf{\Theta}}$ and the available samples X_v and X'_v:

$$\hat{PE}_v^\alpha(X_v\|X'_v) = -\hat{L}_v(\hat{\theta}_v) - \frac{1}{2}. \quad (30)$$

Problem 29 and Eq. 30 highlight how minimizing the former amounts to maximizing the estimated χ^2-divergence, while at the same time accounting for the structure of the graph and the geometry of the RKHS $\mathbb{H}$. Finally, let us define the following expression for $\mathbf{f}\in\mathbb{G}$:

$$PE_v^\alpha(f_v) = \mathbb{E}_{q_v(x')}[f_v(x')] - \frac{1}{2}\mathbb{E}_{p_v^\alpha(y)}[f_v^2(y)] - \frac{1}{2}, \quad \forall v\in V. \quad (31)$$

Notice that, as a consequence of Theorem 1, we have:

$$PE_v^\alpha(r_v^\alpha) = PE(p_v^\alpha\|q_v) \quad \text{and} \quad PE_v^\alpha(r_v^\alpha) \geq PE_v^\alpha(f_v). \quad (32)$$

4. Convergence guarantees

In this section, we discuss the generalization properties of GRULSIF, more precisely the gains brought by the Collaborative LRE when Pearson's χ^2-divergence is used in the surrogate cost function. The main result of this section is summarized in Theorem 3.

For the rest of the section, we assume $n_v = n'_v = n$, i.e. that we have the same number of observations from p_v and q_v at each node v, and all the nodes have the same sample size. Moreover, we assume that observations come in pairs $z_v = (x_v, x'_v)$ as realizations of a probabilistic model described by the joint pdf $p_{z,v}$ with marginal pdfs p_v and q_v.

Let us start by defining the functional space:

$$\mathcal{F}_G = \{\mathbf{f} = (f_1, ..., f_N)\in\mathbb{G} : \frac{1}{2}\|\mathbf{f}\|_\mathbb{G}^2 \leq \Lambda^2\}, \quad (33)$$

where $\Lambda \geq 0$ is a positive constant controlling the smoothness of the vector-valued function to be learned w.r.t. the graph G and the Hilbert space $\mathbb{H}$. The first thing to notice is that Problem 17 can alternatively be written in terms of the functional space $\mathcal{F}_G$:

$$\min_{\mathbf{f}\in\mathcal{F}_G} \frac{1}{N}\sum_{v\in V}\left(\frac{(1-\alpha)}{2n}\sum_{i=1}^{n} f_v^2(x_{v,i}) + \frac{\alpha}{2n}\sum_{i=1}^{n} f_v^2(x'_{v,i}) - \frac{1}{n}\sum_{i=1}^{n} f_v(x'_{v,i})\right). \tag{34}$$

Assumption 2 ***(Independent observations).*** *$\{z_{v,i}\}_{v\in V, i=1,...n} = \{(x_{v,i}, x'_{v,i})\}_{v\in V, i=1,...n}$ represent nN pairs of independent observations: for $u, v \in V$ and $i, j \in \{1, ..., n\}$, $z_{v,i} = (x_{v,i}, x'_{v,i})$ is independent of $z_{u,j} = (x_{u,j}, x'_{u,j})$ if $v \neq u$ or $i \neq j$. Moreover, for each node $v \in V$, the pairs $\{(x_{v,i}, x'_{v,i})\}_{i=1,...,n}$ are also identically distributed under the joint law $p_{z,v}$ with marginals p_v and q_v, where $x_{v,i} \sim p_v$ and $x'_{v,i} \sim q_v$.*

The data independence assumption appears in previous LRE works that study a single data source (Nguyen et al., 2008, 2010; Sugiyama et al., 2012). The above graph-based variant is standard in theoretical analyses based in multitasking involving Vector-Valued Kernels (Maurer, 2006a; Yousefi et al., 2018), yet it can be considered as being a strong hypothesis for many applications.

Assumption 3 ***(Well-specified model).*** *There exists $\Lambda > 0$ such that $\mathbf{r}^\alpha = (r_1^\alpha, ..., r_N^\alpha) \in \mathcal{F}_G$, where $\mathcal{F}_G$ is defined by Eq. 33.*

Assumption 3 states that the proposed statistical model is well-defined in the functional space. In particular, it implies: i) $r_v \in \mathbb{H}$, for all $v \in V$, a common hypothesis in the LRE literature (Nguyen et al., 2008, 2010; Sugiyama et al., 2012; Nguyen et al., 2024); ii) it introduces the parameter Λ, which relates to the regularization constant λ (Problem 25), and formalizes the a priori information encoded in the graph that is required to estimate the vector $\mathbf{r}^\alpha$.

Let $\varphi(y)$ the feature map associated with the RKHS $\mathbb{H}$, and let us consider $g, h \in \mathbb{H}$, and define the operator $g \otimes h : \mathbb{H} \to \mathbb{H}$ as $g \otimes h(f) = \langle f, h\rangle_{\mathbb{H}} g$. Then, we can define the covariance operator associated to the node $v \in V$ as:

$$\Sigma_v = \mathbb{E}_{p_v^\alpha(y)}[\varphi(y) \otimes \varphi(y)]. \tag{35}$$

Assuming the feature space $\mathcal{X}$ is compact and the K is continuous, the Mercer's theorem implies (Aronszajn, 1950; Dieuleveut, 2017):

$$\Sigma_v = \sum_{i=1}^{\infty} \mu_{v,i} \tilde{\varphi}_{v,i} \otimes \tilde{\varphi}_{v,i}, \tag{36}$$

where $\{\tilde{\varphi}_{v,i}\}_{i\in\mathbb{N}}$ forms a Hilbert basis of $\mathbb{H}$, with associated eigenvalues $\{\mu_{v,i}\}_{i\in\mathbb{N}}$. Nevertheless, there exists more general settings where Expr. 36 is satisfied (see Dieuleveut (2017)).

Assumption 4 ***(Capacity condition.)*** *For each $v \in V$, assume Σ_v satisfies Eq. 36. Denote by I the set of indexes of non-zero eigenvalues $\{\mu_{v,i}\}_{i\in I}$ of the operator Σ_v arranged in decreasing order. We assume that $\mu_{v,i} \leq s_v^2 i^{-\zeta_v}$, $i \in I$, for some $\zeta_v > 1$ and some $s_v > 0$.*

The capacity condition quantifies the size of the RKHS $\mathbb{H}$ w.r.t. the eigenbasis $\{\tilde{\varphi}_{v,i}\}_{i\in I}$. Larger ζ_v values lead to faster eigenvalue decay, which means less basis functions are required to approximate $\mathbb{H}$, hence r_v^α can be approximated by a smaller space. When ζ_v approaches 1, a bigger space will be needed to approximate the elements of $\mathbb{H}$, including r_v^α. This assumption has been discussed and analyzed in previous works for obtaining optimal convergence rates in the context of Kernel Ridge regression (Caponnetto and Vito, 2006; Ying and Pontil, 2007; Steinwart et al., 2009).

Theorem 3 *If Assumptions 2-4 are satisfied, and $\mathcal{F}_G$ is a class of functions with ranges in $[-b,b]$, $b\in\mathbb{R}^+$. Then, for any $C\geq 1$ and $\delta\in(0,1)$, with probability at least $1-\delta$, the solution to Problem 34, $\hat{\mathbf{f}}=(\hat{f}_1,...,\hat{f}_N)$, satisfies:*

$$\begin{aligned}\frac{1}{N}\sum_{v\in V}\Big[PE(p_v^\alpha\|q_v)-PE_v^\alpha(\hat{f}_v)\Big]&\leq 2C(20^2)B_1\rho^*+\left(\frac{16B_0^2C+24B_0B_1}{nN}\right)\log\frac{1}{\delta}\\ \frac{1}{N}\sum_{v\in V}\mathbb{E}_{p_v^\alpha(y)}\Big[[\hat{f}_v-r_v^\alpha]^2\Big]&\leq 4C(20^2)B_1\rho^*+\left(\frac{32B_0^2C+48B_0B_1}{nN}\right)\log\frac{1}{\delta},\end{aligned}\tag{37}$$

where:

$$C_{\alpha,v}=\min\{\tfrac{1}{\alpha},\|r_v^\alpha\|_\infty\},\quad C_\alpha=\max_{v\in V}C_{\alpha,v},$$

$$B_0=\tfrac{1}{2}\Big[(b+C_\alpha)^2+4C_\alpha\Big],\quad B_1=\tfrac{1}{2}(b+C_\alpha)^2+(b+C_\alpha),$$

$$\zeta^*=\min_{v\in V}\zeta_v\ \ (\textit{recall}\ \ \zeta_v>1,\forall v\in V),\quad s_{\max}=\max_{v\in V}s_v,$$

$$\rho^*\leq 8B_0\sqrt{\frac{\zeta^*+1}{\zeta^*-1}}\left(\mathcal{T}_G\Lambda^2(b+C_a)^{2\zeta^*}\right)^{\frac{1}{1+\zeta^*}}n^{-\frac{\zeta^*}{1+\zeta^*}}N^{-\frac{1}{1+\zeta^*}}s_{\max}^{\frac{1}{1+\zeta^*}},$$

$$\mathcal{T}_G=\beta\gamma^{-1}+(1-\beta)\lambda_{\min^+}^{-1},\tag{38}$$

where $\mathcal{T}_G$ encodes the graph topology, $\lambda_{\min^+}$ denotes the smallest nonzero eigenvalue of $\mathcal{L}$, and $\beta=\frac{\#C(G)}{N}\in[0,1]$ is the ratio between the number of connected components of G and the total number of nodes.

The proof is given in Appendix C and it relies mainly on the framework of Local Rademacher Complexities for Multitask Learning introduced in Yousefi et al. (2018). Notice that the convergence rates are given in terms of the excess risk and the average $\mathcal{L}^2(P^\alpha)$ distance between $\hat{f}_v$ and r_v^α w.r.t. the measure p_v^α. The excess risk takes the form of the difference between the expected divergence $PE_v^\alpha(\hat{f}_v)$ and the true χ^2-divergence $PE(p_v^\alpha\|q_v)$ that former aims to approximate.

The convergence rates depend mainly on the number of observations per node ($n^{\frac{-\zeta^*}{1+\zeta^*}}$), the number of nodes in the graph ($N^{\frac{-1}{1+\zeta^*}}$), the smoothness of the function to be approximated ($\Lambda^{\frac{2}{1+\zeta^*}}$), the topology of G ($\mathcal{T}_G^{\frac{1}{1+\zeta^*}}$), and the effective dimension of the space to approximate each r_v^α, which is encoded by ζ^* and $s_{\max}$. When ζ^* is small and close to 1, the convergence rate can be as slow as $\mathcal{O}\left(\frac{\sqrt{s_{\max}\mathcal{T}_G}\Lambda}{\sqrt{nN}}\right)$, and as fast as $\mathcal{O}\left(\frac{1}{n}\right)$ when $\zeta_v\to\infty$ for all $v\in V$. This means that the gains of the collaborative estimation in terms of the

excess risk will be more relevant as ζ^* is smaller, since the number of nodes, smoothness, and graph topology play a role in the convergence rate. This situation occurs when a larger number of basis functions $\{\tilde{\varphi}_{v,i}\}_{i\in\mathbb{N}}$ are required to approximate the space $\mathbb{H}$, which could mean that r_v^α is harder to estimate with respect to w.r.t. the kernel function K and the data distribution. In that case, a larger number of nodes and a smoother $\mathbf{r}^\alpha$ over the graph would improve performance. However, the collaborative estimation will offer little advantage when the $\mathbb{H}$ is low-dimensional (large values of ζ^*), since convergence is governed by the number of observations per node. This suggests that GRULSIF can be used in the regime in which multitasking is also recommended: when there are many interrelated tasks with little data per task, and each of the tasks is complex to be solved using only its available local data (Yousefi et al., 2018; Zhang and Yang, 2021).

Example of collaborative vs independent LRE. To illustrate the gains collaboration may bring, consider the special case where all the relative likelihood-ratios are the same, with norms that are upper-bounded by a constant $M > 0$. Let G be a graph with weight matrix W^c parametrized by a constant, such as $c \geq 0$ as $W^c_{uv} = \frac{c}{N}$, for all $u \neq v \in V$, and $W^c_{uu} = 0$ for all $u \in V$. Notice that $W^{c=0} = \mathbf{0}_{N\times N}$ implies no collaboration at all (N connected components), and $W^{c>0}$ induces a fully connected graph without self-loops (a single connected component). For $c > 0$, the smallest non-zero eigenvalue of the graph Laplacian is $\lambda_{\min^+} = c$. Importantly, in both cases the norm $\|r^\alpha\|^2_{\mathbb{G}}$ remains the same. In fact, if we choose $\gamma = \frac{1}{N}$, then:

$$\|r^\alpha\|^2_{\mathbb{G}} = \gamma \sum_{v\in V} \|r^\alpha_v\|^2_{\mathbb{H}} \leq \frac{1}{N}\sum_{v\in V} M = M.$$

Thus, to satisfy Assumption 3, it suffices to take $\Lambda^2 = M$ in both cases.

Let us now compare the two approaches in terms of Theorem 3. In the case of independent LRE ($c = 0$), the leading term of Theorem 3 is upper-bounded as follows:

$$(\rho^*)^{c=0} \leq 8B_0\sqrt{\frac{\zeta^*+1}{\zeta^*-1}}\left(M(b+C_a)^{2\zeta^*}\right)^{\frac{1}{1+\zeta^*}} n^{-\frac{\zeta^*}{1+\zeta^*}} s_{\max}^{\frac{1}{1+\zeta^*}}.$$

On the other hand, in the case of full collaboration ($c > 0$), the bound becomes respectively:

$$\begin{aligned}(\rho^*)^{c>0} &\leq 8B_0\sqrt{\frac{\zeta^*+1}{\zeta^*-1}}\left(\left[1+\frac{1}{Nc}\right]M(b+C_a)^{2\zeta^*}\right)^{\frac{1}{1+\zeta^*}} n^{-\frac{\zeta^*}{1+\zeta^*}} N^{-\frac{1}{1+\zeta^*}} s_{\max}^{\frac{1}{1+\zeta^*}} \\ &\leq 16B_0\sqrt{\frac{\zeta^*+1}{\zeta^*-1}}\left(M(b+C_a)^{2\zeta^*}\right)^{\frac{1}{1+\zeta^*}} n^{-\frac{\zeta^*}{1+\zeta^*}} N^{-\frac{1}{1+\zeta^*}} s_{\max}^{\frac{1}{1+\zeta^*}}. \qquad (\text{If } c \geq \tfrac{1}{N})\end{aligned}$$

The main difference between those upper-bounds is that the term $N^{-\frac{1}{1+\zeta^*}}$ disappears in the independent setting. In other words, the convergence rates depend only on the number of observations per node n, and no effect from having multiple nodes. Contrary, the collaborative approach benefits from enforcing the prior information regarding the similarity between relative likelihood-ratios (larger values of c), and also from the number of nodes. The benefit of collaboration is mediated by the complexity of the problem encoded in the parameter ζ^*.

The case $\alpha = 0$**.** The convergence rates given in Theorem 3 still apply when the α-regularization is ignored, provided that $Q \ll P$ and for each $v \in V$ there exists $C_v > 0$ such that $\|r_v\|_\infty \leq C_v < \infty$. Although this hypothesis is quite standard in the literature (Nguyen et al., 2010; Sugiyama et al., 2011a; Kanamori et al., 2011; Nguyen et al., 2024), it is restrictive and hard to verify in the general LRE setting in the absence of prior knowledge about P and Q. In practice, if we set $\alpha = 0$ and this hypothesis is not satisfied, we may face situations where LRE approximates a function that either does not exist or is unbounded, which may hinder the numerical performance.

The impact on the convergence rates is visible in Theorem 3. If $\alpha = 0$, with the convention $\frac{1}{0} = \infty$, then $C_\alpha = \max_{v \in V} \|r_v\|_\infty$, meaning that the generalization bounds scale with the norms $\{\|r_v\|_\infty\}_{v \in V}$, which are unknown and potentially very large. α-regularization prevents the likelihood-ratios from exploding and provides some control over the convergence rates via the hyperparameter α. Of course, there may be situations where $Q_v \ll P_v$, $\forall v \in V$, and $\max_{v \in V} \|r\|_\infty \ll \frac{1}{\alpha}$; in this case Theorem 3 suggests that regularization is unnecessary. However, since such a condition cannot be verified in the general problem setting, we prefer to fix the same α for all the nodes to avoid additional hypotheses on $\{Q_v\}_{v \in V}$ and $\{P_v\}_{v \in V}$.

The case $Q_v = P_v, \forall v \in V$**.** Here, the likelihood-ratios are well-defined: $r_v(x) = r_v^\alpha(x) = 1$, $\forall v \in V$, for any $0 < \alpha < 1$. Furthermore, it easy to verify that $1 = C_{\alpha=0} = C_{\alpha \neq 0} = 1$. Therefore, under the assumptions stated at the beginning of this section and Theorem 3, the convergence behavior would be independent of the regularization in this case.

5. Practical implementation

A straightforward optimization of Problem 25 would set the derivative of the objective function to zero, i.e. $\nabla_{\boldsymbol{\Theta}} \Phi(\boldsymbol{\Theta}) = 0$, and solve to get the estimated parameters $\hat{\boldsymbol{\Theta}}$:

$$\hat{\boldsymbol{\Theta}} = \frac{1}{N} \mathbf{A}^\dagger \mathbf{h}', \tag{39}$$

where $A^\dagger$ denotes the pseudoinverse of $A^{NL \times NL}$. Nevertheless, the size of matrix A scales with the number of nodes in the graph (N) and the number of available observations (L). The total complexity of this optimization approach would be of scale $\mathcal{O}((LN)^3)$, which makes it prohibitive to compute in most practical situations. For deploying GRULSIF in practice, we propose in this section an optimization procedure that can handle efficiently large graphs and a substantial number of observations. Additionally, we detail the strategy to identify the regularization constants λ, $\gamma > 0$, and the hyperparameters related to the kernel that we will denote by σ, when $\mathrm{K}(\cdot, \cdot)$ is the Gaussian kernel σ is the width parameter.

5.1 Computing the node parameter updates via CBCGD

Instead of computing the peudoinverse $A^\dagger$ to solve Problem 25, we propose to use the Cyclic Block Coordinate Gradient Descent (CBCGD) method (Beck and Tetruashvili, 2013; Li et al., 2018). Our optimization schema operates in cycles, and each cycle involves multiple iterations of a block coordinate gradient descent (GD) one for each node v; therefore, the high-level complexity is $\mathcal{O}(\#\text{Cycles} \cdot \#\text{Nodes} \cdot \text{Cost\_of\_GD\_at\_one\_node})$. Starting from the

last term, CBCGD's i-th cycle has to estimate the node parameter $\hat{\theta}_v^{(i)}$ at each node v:

$$\hat{\theta}_v^{(i)} = (\lambda\gamma\mathcal{K} + \eta_v I_L)^{-1}\Bigg[\overbrace{\eta_v\hat{\theta}_v^{(i-1)} - \left[\left(\frac{1-\alpha}{N}H_v + \frac{\alpha}{N}H'_v\right)\hat{\theta}_v^{(i-1)} - \frac{h'_v}{N}\right]}^{\text{component depending on node } v} \\ \overbrace{-\lambda\mathcal{K}\left(d_v\hat{\theta}_v^{(i-1)} - \sum_{u\in \text{ng}(v)} W_{uv}\big(\mathbb{1}\{u<v\}\hat{\theta}_u^{(i)} + \mathbb{1}\{u\geq v\}\hat{\theta}_u^{(i-1)}\big)\right)}^{\text{component depending on the graph}}\Bigg], \tag{40}$$

where η_v is the node learning rate, recall that d_v is the node degree and $\mathbb{1}\{\cdot\}$ is the indicator matrix. Notice the elegance of the decomposition of the update into two components: one depending on the node v itself, and the other depending on the graph, i.e. only on v's neighbors. Important to note that, the node parameters are estimated asynchronously in each cycle in an arbitrary but fixed cyclic order; this is clear in the summation inside the graph-related component. Tweaking this order to adapt it to specific communication restrictions between nodes is possible, but it is left to the reader to specify the most convenient setting for her needs (see Wright (2015) for a review of the topic). CBCGD is easy to implement, and when applied to quadratic problems leads to a manageable complexity in terms of the number of cycles required to achieve convergence (Li et al., 2018). This kind of result is made explicit for the optimization schema described in Expr. 40 in the following theorem.

Theorem 4 *Suppose that for a dictionary D of size $L \geq 2$ we desire to solve the optimization Problem 25 via the CBCGD strategy, where the update w.r.t. the node parameter θ_v at the i-th cycle is computed as detailed in Eq. 40. Then, if we fix the learning rate for node v to be equal to the maximum eigenvalue $\eta_v = e_{\max}\left(\frac{(1-\alpha)}{N}H_v + \frac{\alpha}{N}H'_v + \lambda d_v\mathcal{K}\right)$, we will need at most the following number of cycles for achieving a pre-specified accuracy level $\epsilon > 0$:*

$$i_{\max} = \left\lceil \frac{\lambda\gamma c(C_{\min} + \lambda\gamma c) + 16C^2\log^2(3NL)}{\lambda\gamma c(C_{\min} + \lambda\gamma c)} \cdot \log\left(\frac{1}{\epsilon}\left(\Phi(\boldsymbol{\Theta}^{(0)}) - \Phi(\boldsymbol{\Theta}^*)\right)\right)\right\rceil, \tag{41}$$

where $\Phi(\boldsymbol{\Theta})$ is the cost function of Expr. 25, $c > 0$ is a positive constant, $C_{\min} = \min_{v\in V} C_v$:

$$\begin{aligned} C &= e_{\max}\left(\frac{1-\alpha}{N}\mathbf{H} + \frac{\alpha}{N}\mathbf{H}' + \lambda(I_N \otimes \mathcal{K}^{\frac{1}{2}})\left[\mathcal{L}\otimes I_L\right](I_N \otimes \mathcal{K}^{\frac{1}{2}})\right), \\ C_v &= e_{\max}\left(\frac{1-\alpha}{N}H_v + \frac{\alpha}{N}H'_v + \lambda d_v\mathcal{K}\right). \end{aligned} \tag{42}$$

The proof of Theorem 4 is provided in Appendix A.3. The computational complexity of the full optimization schema depends on two components: 1) estimating the optimal learning rates η_v and the inversion of the matrix $\lambda\gamma\mathcal{K} + \eta_v I_L$, operations to be done just once for each of the nodes, this step amounts to a computational cost of $\mathcal{O}(NL^3)$. 2) The cost of Eq. 40 across all nodes and cycles. The cost for a node v at a given cycle i is dominated by matrix-vector multiplications of dimension L, leading to a cycle cost of $\mathcal{O}(NL^2)$. As indicated by Eq. 25, the required number of CBCGD cycles for achieving a given accuracy level ϵ scales in $\mathcal{O}(\log^2(NL))$. The total cost of the second step is then $\mathcal{O}(NL^2\log^2(NL))$. The total cost of the whole optimization schema is then $\mathcal{O}(NL^3 + NL^2\log^2(NL))$.

5.2 Nyström dimensionality reduction strategy

The main computation burden of the CBCGD method is related to the dataset size, that is the number of observations L. This is a common problem in Kernel Methods and has motivated extensive research. For instance, the *random features* approach (Rahimi and Recht, 2007) uses a randomized feature map to approximate the input space by a low dimensional Euclidean space. Low-rank approximations, such as Nyström approximations (Williams and Seeger, 2000; Smola and Schökopf, 2000), use a subsample of observations as a dictionary, to define a finite dimensional space that preserves the approximation properties of the original space. In time-series analysis, Richard et al. (2009) proposed to grow the dictionary by adding new elements one-by-one, according to a *coherence threshold* that keeps the linear dependency of the dictionary elements as low as possible (i.e. as diverse as possible basis functions), while still being able to approximate any of the functions in $\mathbb{H}$. The usual approach followed in non-parametric LRE is simply to create a dictionary out of a subsample of the observations chosen uniformly at random (Sugiyama et al., 2012).

In general, the choice of the dictionary learning method depends on the task, the time complexity requirements, and the nature of the chosen kernel. Nyström approximations replace the feature map $\varphi(x)$ by its orthogonal projection into a finite-dimensional space $\mathbb{F} = \text{span}(\{\varphi(x) : x \in D_{\hat{L}}\})$, where $D_{\hat{L}} = \{x_i \in \mathcal{X}\}_{i=1}^{\hat{L}}$ is a set of carefully chosen points in the original input space (not restricted to data observations), and span$(\cdot)$ refers to the set of lineal combinations of the input elements, for some chosen $\hat{L} \ll L$. The points $\varphi(x_1), ..., \varphi(x_{\hat{L}})$ are known as *anchor points* in $\mathbb{H}$, and, via the associated kernel matrix $\mathcal{K}_{\hat{L}} \in \mathbb{R}^{\hat{L}\times\hat{L}}$, $[\mathcal{K}_{\hat{L}}]_{ij} = \mathrm{K}(x_i, x_j)$, they allow the definition of a new feature map:

$$\psi(\cdot) = \mathcal{K}_{\hat{L}}^{-\frac{1}{2}} \left(\mathrm{K}(\cdot, x_1), .., \mathrm{K}(\cdot, x_{\hat{L}})\right)^{\mathsf{T}}, \tag{43}$$

The idea is to choose the anchor points such that preserve the geometry of $\mathbb{H}$, i.e. the dot product in the infinite dimensional space $\mathbb{H}$ gets translated into a dot product in $\mathbb{R}^{\hat{L}}$:

$$\mathrm{K}(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathbb{H}} \approx \langle \psi(x), \psi(y) \rangle, \quad \forall x, y \in \mathcal{X}.$$

According to the empirical risk minimization and the Representer Theorem (Expr. 19), the node-level approximation in this new space takes the form:

$$f_v(x) = \sum_{i=1}^{L} \mathrm{K}(x, x_i)\theta_{v,i} = \Big\langle \sum_{i=1}^{L} \varphi(x_i)\theta_{v,i}, \varphi(x) \Big\rangle_{\mathbb{H}} \approx \langle w_v, \psi(x) \rangle, \tag{44}$$

where $w_v \in \mathbb{R}^{\hat{L}}$. This approximation can rephrase Problem 25 in terms of vectors in $\mathbb{R}^{\hat{L}}$ and the new feature map $\psi(\cdot)$:

$$\min_{\boldsymbol{\Theta}\in\mathbb{R}^{N\hat{L}}} \frac{1}{N} \sum_{v\in V} \left(\frac{1-\alpha}{2} \theta_v^{\mathsf{T}} H_{\psi,v} \theta_v + \frac{\alpha}{2} \theta_v^{\mathsf{T}} H'_{\psi,v} \theta_v - (h'_{\psi,v})^{\mathsf{T}} \theta_v \right) + \frac{\lambda}{4} \sum_{u,v\in V} W_{uv} \|\theta_v - \theta_u\|^2 + \frac{\lambda\gamma}{2} \sum_{v\in V} \|\theta_v\|^2, \tag{45}$$

where $\|\cdot\|$ refers to the Euclidean norm, and the terms $H_{\psi,v}, H'_{\psi,v} \in \mathbb{R}^{\hat{L}\times\hat{L}}$, and $h'_{\psi,v} \in \mathbb{R}^{\hat{L}}$ are those of Eq. 24, but now computed using their new associated feature map $\psi(\cdot)$. Recall that these terms need to be computed only once at the beginning. Moreover, Nyström

approximation does not affect the structure of the problem, which remains quadratic and can be solved via CBCGD, with each iteration taking the form:

$$\hat{\theta}_v^{(i)} = \frac{1}{\lambda\gamma+\eta_v}\Bigg[\overbrace{\eta_v\hat{\theta}_v^{(i-1)} - \left[\left(\frac{1-\alpha}{N}H_{\psi,v} + \frac{\alpha}{N}H'_{\psi,v}\right)\hat{\theta}_v^{(i-1)} - \frac{1}{N}h'_{\psi,v}\right]}^{\text{component depending on node } v} \\ \overbrace{-\lambda\left(d_v\hat{\theta}_v^{(i-1)} - \sum_{u\in V} W_{uv}\left(\mathbb{1}\{u<v\}\hat{\theta}_u^{(i)} + \mathbb{1}\{u\geq v\}\hat{\theta}_u^{(i-1)}\right)\right)}^{\text{component depending on the graph}}\Bigg]. \tag{46}$$

The final computational cost is the sum of the cost of encoding the data points via Expr. 43 and the optimization procedure. The encoding requires a matrix inversion $\mathcal{O}(\hat{L}^3)$ and L matrix-vector multiplications of dimension $\hat{L}$ (this is overall $\mathcal{O}(L\hat{L}^2)$, while CBCGD requires the estimation of the optimal learning rates η_v and has total cost $\mathcal{O}(N\hat{L}^3)$, and the cost of all node iterates across cycles that amounts to $\mathcal{O}(N\hat{L}^2\log^2(N\hat{L}))$. In conclusion, Nyström approximation enables the reduction of the computation complexity from $\mathcal{O}(NL^3 + NL^2\log^2(NL))$ to $\mathcal{O}(N\hat{L}^3 + L\hat{L}^2 + N\hat{L}^2\log^2(N\hat{L}))$, where $\hat{L} \ll L$.

Nyström approximation not only offers computational gains, but it also brings interesting features from a data accessibility perspective: notice that computing the node-level quantities H_v, H'_v, h'_v requires access to the full dataset (Expr. 24), while $H_{\psi,v}$, $H'_{\psi,v}$, $h'_{\psi,v}$ requires only the anchor points and the available samples at that node (X_v, X'_v) (Expr. 43). In this problem formulation, the update of the vector parameter θ_v of node v only requires the computation of $\psi(\cdot)$ using node's own local observations, and the use of the parameters of its neighbors $\{\theta_u\}_{u\in\text{ng}(v)}$. In conclusion, Nyström approximation combined with our optimization schema enables a distributed LRE at each node and, hence , limits to only indirect node access to foreign data of other nodes through the parameters θ_u for $u \in \text{ng}(v)$. This is appealing for applications with data access restrictions.

The remaining important question is how to select the set of anchor points. There are many strategies to address this problem; for example, Kernel PCA (Schölkopf et al., 1998), random sampling (Williams and Seeger, 2000; Talwalkar et al., 2008), greedy approaches (Bach and Jordan, 2002), or k-means clustering (Zhang et al., 2008). In this work we use the approach proposed by Richard et al. (2009) that is based on the coherence measure. That algorithm builds a dictionary of manageable size and low redundancy, has a low computational cost, and, under mild conditions, it produces good approximations of the whole space. The adaptation of this strategy in our context can be found in Appendix A.3.

5.3 POOL: a *no graph* variant

One important by-product of the GRULSIF framework and our supporting analysis, is that we can derive a reduced LRE variant, which we call POOL, that we call that disregards the graph, while enjoying all the other advantages of our non-parametric optimization formulation. By setting $W = \mathbf{0}_{N\times N}$, which neutralizes the graph component and hence the associated terms disappear from Eq. 46, we get POOL's optimization problem:

$$\min_{\boldsymbol{\Theta}\in\mathbb{R}^{N\hat{L}}} \frac{1}{N}\sum_{v\in V}\left(\frac{1-\alpha}{2}\theta_v^\intercal H_{\psi,v}\theta_v + \frac{\alpha}{2}\theta_v^\intercal H'_{\psi,v}\theta_v - (h'_{\psi,v})^\intercal\theta_v\right) + \frac{\lambda\gamma}{2}\sum_{v\in V}\|\theta_v\|^2. \tag{47}$$

This yields N independent quadratic problems admitting a closed form solution:

$$\hat{\theta}_v = \frac{1}{N}\left[\frac{1}{N}\left((1-\alpha)H_{\psi,v} + \alpha H'_{\psi,v}\right) + \lambda\gamma I_{\hat{L}}\right]^{-1} h'_{\psi,v}. \tag{48}$$

POOL leads to a total computational complexity of $\mathcal{O}(N\hat{L}^3 + L\hat{L}^2)$ (the term related with the cost of the CBCGD schema disappears). POOL can be relevant when it is believed that there is no graph behind the observed phenomena at the different locations, or in situations like those detailed in Sec. 4 where the collaborative estimation may offer little advantage. Moreover, POOL can be seen as a RULSIF variant (Yamada et al., 2011), where POOL's main differences are: i) its hyperparameters are selected jointly w.r.t. to the mean score $\frac{1}{N}\sum_{v\in V}\ell_v(\theta_v)$, while RULSIF selects independently the hyperparameters for each task; ii) POOL uses the Nyström dimensionality reduction technique over the full set of observations, while RULSIF uses a simple uniform random sampling at each node (Sugiyama et al., 2012).

5.4 Hyperparameter selection

The performance of GRULSIF depends on the penalization constants γ, λ, and the hyperparameters of the kernel K (e.g. for a Gaussian kernel, that would be only the width σ). As in previous works in non-parametric ϕ-divergence estimation, we use a cross-validation strategy (Sugiyama et al., 2007, 2011a; Yamada et al., 2011). The main difference is that in GRULSIF we aim to minimize the average of the cost function over all the nodes of the graph. Thus, the score used to identify the optimal hyperparameters is:

$$\hat{L}(\Theta,\sigma) = \frac{1}{N}\sum_{v\in V}\hat{L}_v(\theta_v,\sigma) = \frac{1}{N}\sum_{v\in V}\left(\frac{1-\alpha}{2}\theta_v^\mathsf{T} H_{\psi,v}(\sigma)\theta_v + \frac{\alpha}{2}\theta_v^\mathsf{T} H'_{\psi,v}(\sigma)\theta_v - h'_{\psi,v}(\sigma)^\mathsf{T}\theta_v\right), \tag{49}$$

where $H_{\psi,v}(\sigma)$, $H'_{\psi,v}(\sigma)(\sigma)$, $h'_{\psi,v}(\sigma)$ explicit the relationship between these operators and the hyperparameters of the kernel function K.

At each iteration of the cross-validation, we define two training sets, $\mathbf{X}_{\text{train}}$, $\mathbf{X}'_{\text{train}}$, to update $H_{\psi,v}(\sigma)$, $H'_{\psi,v}(\sigma)(\sigma)$, $h'_{\psi,v}(\sigma)$. We fix the two hyperparameters λ and γ to estimate the parameter $\hat{\Theta}(\sigma,\lambda,\gamma)$, which is the solution to the optimization problem:

$$\hat{\Theta}(\sigma,\gamma,\lambda) = \underset{\Theta}{\operatorname{argmin}}\,\frac{1}{N}\sum_{v\in V}\hat{L}_v(\theta,\sigma) + \frac{\lambda}{4}\sum_{u,v\in V}W_{uv}\left\|\theta_v-\theta_u\right\|^2 + \frac{\gamma}{2}\sum_{v\in V}\left\|\theta_v\right\|^2. \tag{50}$$

The solution of Problem 50 is found via Alg. 2. Finally, the parameter $\hat{\Theta}(\sigma,\gamma,\lambda)$ is used to identify which combination of parameter values $\sigma^*,\lambda^*,\gamma^*$ are optimal such that they minimize the expected value of the chosen score $\hat{L}(\hat{\Theta}(\sigma,\gamma,\lambda),\sigma)$ of Eq. 49. The implementation details of the model selection are provided in Alg. 1.

We can apply a similar approach to find the hyperparameters of POOL. As POOL ignores the graph structure ($W=\mathbf{0}_{N\times N}$), we fix $\lambda=1$, hence the penalization term related to the norm of each functional f_v will only depend on γ (Eq. 47). Then, we use cross-validation to identify the optimal values of the hyperparameters σ and γ. The score to decide for this choice is the same as one described in Eq. 49.

Once the anchor points of Nyström approximation and the hyperparameters have been fixed, we can learn the relative likelihood-ratios. For GRULSIF this amounts to estimating

Algorithm 1 – Model selection for GRULSIF hyperparameters tuning

Input: $\mathbf{X}, \mathbf{X}'$: the two sets of observations to be used for estimating the likelihood-ratios;
$G = (V, E, W)$: a given graph;
$\#_\sigma, \#_\lambda, \#_\gamma$: parameter grid to explore for values of σ, λ, γ;
R: the number of random splits.
Output: σ^* the optimal scale parameter for the Gaussian kernel, the associated dictionary D_{σ^*}, and the two penalization constants λ^* and γ^*.

Randomly split $\mathbf{X}$ and $\mathbf{X}'$ into R disjoint subsets $\{\mathrm{X}_r\}_{r=1}^R$ and $\{\mathrm{X}'_r\}_{r=1}^R$
for each $\sigma \in \#_\sigma$ **do**
  Compute a dictionary D_σ using the chosen kernel and hyperparameter σ
  for each $(\lambda, \gamma) \in \#_\lambda \times \#_\gamma$ **do**
    for each data subset $r = 1, ..., R$ **do**
      Let $\mathbf{X}'_{\text{train}} = \mathbf{X}' \backslash \mathrm{X}'_r$, $\mathbf{X}'_{\text{test}} = \mathrm{X}'_r$, and $\mathbf{X}_{\text{train}} = \mathbf{X} \backslash \mathrm{X}_r$, $\mathbf{X}_{\text{test}} = \mathrm{X}_r$
      Compute $h'_{\text{train}}(\sigma)$ and $H'_{\text{train}}(\sigma)$ using the observations in $\mathbf{X}'_{\text{train}}$ (see Eq. 24)
      Compute $H_{\text{train}}(\sigma)$ using the observations in $\mathbf{X}_{\text{train}}$ (see Eq. 24)
      Find $\hat{\mathbf{\Theta}}(\sigma, \gamma, \lambda)$ by solving
      Compute $h'_{\text{test}}(\sigma)$ and $H'_{\text{test}}(\sigma)$ using the observations in $\mathbf{X}'_{\text{test}}$
      Compute $H_{\text{test}}(\sigma)$ using the observations in $\mathbf{X}_{\text{test}}$
      Compute $\hat{L}^{(r)}(\hat{\mathbf{\Theta}}(\sigma, \gamma, \lambda)) = \frac{1}{N} \sum_{v \in V} \hat{L}_v(\hat{\theta}_v(\sigma, \gamma, \lambda))$
        using $h'_{\text{test}}(\sigma)$, $H'_{\text{test}}(\sigma)$, and $H_{\text{test}}(\sigma)$
    end for
    Compute $\hat{L}(\sigma, \lambda, \gamma) = \frac{1}{R} \sum_{r=1}^R \hat{L}^{(r)}(\mathbf{\Theta}(\sigma, \gamma, \lambda))$
  end for
end for
$\gamma^* = \mathrm{argmin}_{\sigma, \lambda, \gamma} \hat{\ell}(\sigma, \lambda, \gamma)$
return σ^*, D_{σ^*}, λ^*, γ^*

the parameter $\hat{\Theta}(\sigma^*, \lambda^*, \gamma^*)$ via Alg. 2, while POOL estimates $\hat{\Theta}(\sigma^*, \lambda^* = 1, \gamma^*)$ by solving the N independent quadratic problems of Eq. 48.

The hyperparameter α requires a more complex discussion. On one hand, it depends on the LRE application: it is clear that when $\alpha = 1$, the relative likelihood-ratio $r^\alpha(x) = \frac{q(x)}{(1-\alpha)p(x) + \alpha q(x)} = 1$ and the χ^2-divergence $P^\alpha(p \| q) = 0$, independently to p and q. That would make them meaningless as quantities for quantifying the difference between p and q. On the other extremity, if $\alpha = 0$, we recover the classical likelihood-ratio $r^{\alpha=0}(x) = r(x) = \frac{q(x)}{p(x)}$. As we mentioned in Sec. 2.3, in this case, $r(\cdot)$ may be an unbounded function and cause problems of convergence or numerical instability (Yamada et al., 2011). Such phenomena are visible in the rates provided in Theorem 3, where the constants depending on α become undefined. The role of α is to prevent this from happening, since it upper-bounds r^α:

$$r^\alpha(x) = \left[\frac{(1-\alpha)p(x) + \alpha q(x)}{q(x)} \right]^{-1} = \frac{1}{(1-\alpha) r^{-1}(x) + \alpha} \leq \frac{1}{\alpha}.$$

Therefore, at the level of estimating a single likelihood-ratio, the best way to see α is that it smooths the denominator of r^α, which is the reference measure to compare p with q. In that sense, values of $\alpha > 0$ that are far from 1 help in defining meaningful and sensitive estimators. In addition, α has an effect at the graph level: higher values make all estimates getting closer to 1 and consequently become more similar to each other (i.e.

Algorithm 2 – GRULSIF: Collaborative and distributed LRE over a graph

Input: $\mathbf{X}, \mathbf{X}'$: two samples with observations over the nodes of a graph graph $G = (V, E, W)$;
$\alpha \in [0, 1)$: parameter of the relative likelihood-ratio (Eq. 2);
σ, D: kernel hyperparameter, and a dictionary containing precomputed set of $\hat{L}$ anchor points associated with a kernel $\mathrm{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$;
λ, γ: constants multiplying the penalization terms;
$\hat{\mathbf{\Theta}}^{(0)}$, tol : initialization of node parameters, and tolerated relative error before termination.
Output: estimated parameters $\hat{\mathbf{\Theta}} = \mathrm{vec}(\hat{\theta}_1, ..., \hat{\theta}_N)$.

for each node $v \in \{1, ..., N\}$ **do**
  Compute $H_{\psi,v}$, $H'_{\psi,v}$, and $h'_{\psi,v}$ (see Eq. 24, using ψ instead of ϕ)
  Compute the learning rate by $\eta_v = e_{\max}\left(\frac{1-\alpha}{N} H_{\psi,v} + \frac{\alpha}{N} H'_{\psi,v} + \lambda d_v I_{\hat{L}}\right)$ (see Theorem 4)
end for
$i = 0$
repeat
  $i = i + 1$
  for each node $v \in \{1, ..., N\}$ **do**
    Update the node parameter $\hat{\theta}_v^{(i)}$ (see Eq. 46 and Sec. 5.2)
  end for
until $\frac{\|\hat{\mathbf{\Theta}}^{(i)} - \hat{\mathbf{\Theta}}^{(i-1)}\|}{\|\hat{\mathbf{\Theta}}^{(i-1)}\|} > tol$
return $\hat{\mathbf{\Theta}}^{(i)}$

stronger higher graph smoothness), which impacts the convergence rates of Collaborative LRE (see Theorem 3). To summarize, the optimal α value depends on the interplay between the convergence rates of the used estimation method and the performance achieved in the task we intend to address with the estimated likelihood-ratios. We investigate empirically the first point in dedicated experiments in the next section.

6. Experiments

The empirical evaluation of the GRULSIF framework is conducted for the objective of estimating the likelihood-ratio r_v^α for each node of a given fixed graph. In Sec. 6.1, we present synthetic experiments where the true likelihood-ratios are known by their design. The evaluation in real problems is challenging since the true likelihood-ratios are generally not known, however, in Sec. 6.2 we design a particular setting for seismic data where those true quantities can be safely assumed. In all experiments, both GRULSIF and POOL follow the numerical implementation guidelines described in Sec. 5, which include the CBCGD optimization technique and the Nyström approximation of the RKHS.

6.1 Synthetic experiments

Scenarios. Each instance of a synthetic experiment is generated by the three stages below.

1. *Graph structure.* A random graph is generated according to a standard model:
   - A *Stochastic Block Model* (SBM) with 4 clusters, each containing 25 nodes (intra-cluster edge probability: 0.5; inter-cluster edge probability: 0.01).
   - A *Grid* graph model with 100 nodes arranged in 10 rows and 10 columns.

Table 1: Synthetic scenarios. The scenarios are defined by the graph structure they employ, the node-level distributions (p_v and q_v) generating the data observations at each node, and the location where changes take place in the graph. When the distributions or their parameters remain unchanged between p_v and q_v, this is indicated by '•'.

				Node-level hypotheses		
Experiment	$\mathcal{X}$	**Graph**	**Location**	p_v	**vs.**	q_v
Synth.Ia	$\mathbb{R}^1$	SBM 4 clusters, 25 nodes each	$v \in C_1$ $v \in C_2 \cup C_3$ $v \in C_4$	$\mathrm{N}(\mu=0, \sigma=1)$ $\mathrm{N}(\mu=0, \sigma=1)$ $\mathrm{N}(\mu=0, \sigma=1)$	vs. vs. vs.	$\mathrm{Uniform}(-\sqrt{3}, \sqrt{3})$ • $\mathrm{N}(\mu=1, \sigma=\bullet)$
Synth.Ib	$\mathbb{R}^2$	SBM 4 clusters, 25 nodes each	$v \in C_1 \cup C_2$ $v \in C_3$ $v \in C_4$	$\mathrm{N}(\mu=(0,0)^\mathsf{T}, \Sigma_{1,2}=-\frac{4}{5})$ $\mathrm{N}(\mu=(0,0)^\mathsf{T}, \Sigma_{1,2}=\frac{4}{5})$ $\mathrm{N}(\mu=(0,0)^\mathsf{T}, \Sigma_{1,2}=0)$	vs. vs. vs.	• $\mathrm{N}(\mu=\bullet, \Sigma_{1,2}=0)$ $\mathrm{N}(\mu=(1,1)^\mathsf{T}, \Sigma_{1,2}=\bullet)$
Synth.IIa	$\mathbb{R}^2$	Grid 100 nodes	$v \in V$	$\mathrm{N}(\mu=(0,0)^\mathsf{T}, \Sigma=I)$	vs.	$\mathrm{N}(\mu=(r,c)^\mathsf{T}, \Sigma_{0,0}=1+\lvert r\rvert, \Sigma_{1,1}=1+\lvert c\rvert, \Sigma_{1,2}=\Sigma_{2,1}=0)$ $r=2[(\#\mathrm{row}-5)/\max_{\#\mathrm{row}}(\#\mathrm{row}-5)]$ $c=2[(\#\mathrm{col}-5)/\max_{\#\mathrm{col}}(\#\mathrm{col}-5)]$
Synth.IIb	$\mathbb{R}^4$	Grid 4 quadrants, 25 nodes each	$v \in C_2 \cup C_3 \cup C_4$	$\mathrm{N}(\mu=\mathbf{0}_4, \Sigma=I_4)$	vs.	$\mathrm{N}(\mu=\mathbf{0}_4, \Sigma_{i,i}=\bullet, \Sigma_{i,i+1}=0.8, \Sigma_{i+1,i}=0.8)$
Synth.IIc	$\mathbb{R}^{20}$	Grid 4 quadrants, 25 nodes each	$v \in C_2 \cup C_3 \cup C_4$	$\mathrm{N}(\mu=\mathbf{0}_{20}, \Sigma=I_{20})$	vs.	$\mathrm{N}(\mu=\mathbf{0}_{20}, \Sigma_{i,i}=\bullet, \Sigma_{i,i+1}=0.8, \Sigma_{i+1,i}=0.8)$

2. *Nodes' behavior.* A scheme is considered that first specifies if a node v shall experience a change of measure or not ($p_v \neq q_v$ vs. $p_v = q_v$), and then associates specific pdfs to it. This is a critical design feature since, for the Collaborative LRE to be meaningful, nodes' behavior (expressed as likelihood-ratios) should be explainable by the graph. In each scenario, one of the following two schemes is used:
   ▪ *Cluster-based scheme*: All nodes in a cluster exhibit the same behavior. It is used for SBM graphs that have inherent cluster structure. Clusters are denoted by $C_1, C_2, \ldots$.
   ▪ *Change in most of the graph*: The four quadrants of a Grid graph are considered as being separate clusters of nodes. The nodes of the first quadrant (C_1) remain unchanged, while for the nodes in the other quadrants the change is in the covariance matrix, specifically the off-diagonal elements increase from 0 to 0.8.
   ▪ *Smooth graph variation*: All nodes in a graph experience a change in both their mean vectors and their covariance matrices. The magnitude of a change get larger as the node is more distant to the center of the grid, therefore the distribution q_v is different for each node. This scheme induces smooth node changes across a strong spatial structure, without however corresponding to a multi-cluster structure.

3. *Data observations.* Finally, for each node v, an equal number of $n_v = n'_v = n$ (i.e. same for all nodes) data observations are generated from each associated p_v and q_v.

In each scenario summarized in Tab. 1, the observations have the same dimensionality for all nodes, 1-, 2-, 4- or 20, as our framework requires for nodes to have the same input space $\mathcal{X}$ and RKHS. The scenarios are designed to pose various challenges. In some cases p_v and q_v are different probability models, in others they are the same model with different parametrizations. Moreover, there can be more than one type of change in a scenario; e.g. in Synth.Ia, all p_v's are the same Normal distribution, the cluster C_2 remains unchanged, while C_1 becomes a Uniform with the same first two moments of a standard Normal distribution, and C_3 changes its mean. In Synth.IIa, the change of measure is different for each node.

Table 2: LRE competitors. All the methods included in our experimental evaluation.

Method	Reference	Target function	ϕ-divergence	Graph
KLIEP	Sugiyama et al. (2007)	l.-r.	KL-divergence	No
ULSIF	Sugiyama et al. (2011a)	l.-r.	χ^2-divergence	No
RULSIF	Yamada et al. (2011)	relative l.-r.	χ^2-divergence	No
POOL	this work (Sec. 5.3)	relative l.-r.	χ^2-divergence	No
GRULSIF	this work	relative l.-r.	χ^2-divergence	Yes

Synth.IIb and Synth.IIc aim to illustrate the impact of dimensionality to the performance of GRULSIF, similar experiments were originally presented by Rhodes et al. (2020).

Compared LRE methods. We compare against existing Kernel-based LRE methods built upon an ϕ-divergence variational formulation, namely ULSIF (Sugiyama et al., 2011a), RULSIF (Yamada et al., 2011), and KLIEP (Sugiyama et al., 2007). POOL, RULSIF, and ULSIF rely on the χ^2-divergence; the first two use the relative likelihood-ratio (Eq. 2), and ULSIF uses the classical definition (equiv. to when $\alpha = 0$). KLIEP uses the KL-divergence. We also include POOL (Sec. 5.3) relies on the proposed optimization scheme and the same Nyström approximation as GRULSIF, but disregards the graph. Tab. 2 summarizes the compared methods, while their hyperparameter selection is discussed in Appendix B.1.

Evaluation measures. The variational formulation of ϕ-divergences, described in Sec. 2.3, establishes the connection between estimating the likelihood-ratio between two probability measures and quantifying their dissimilarity via a ϕ-divergence (see Theorem 1). All LRE methods listed in Tab. 2 leverage this relation. In the graph-based extension presented in this paper, an LRE method produces node-level likelihood-ratio estimates, which can approximate the ϕ-divergence between the node-level probability measures p_v and q_v. The connection between both approximations derives from a similar approach to the one described in Sec. 3. Therefore, when knowing p_v and q_v, we can compare the performance of LRE methods along two dimensions: how accurately they approximate the target true node-level likelihood-ratio functions, and how effectively they quantify the true ϕ-divergence between the corresponding probability measures.

The approximation to the true node-level likelihood-ratios is quantified by the average node-level *Mean Squared Error* (MSE):

$$\begin{aligned} P^{\alpha}\left[[\mathbf{f}-\mathbf{r}^{\alpha}]^2\right] &= \frac{1}{N}\sum_{v\in V}\mathbb{E}_{p_v^{\alpha}(y)}[[r_v^{\alpha}-\hat{f}_v]^2(y)] \\ &= \frac{1}{N}\sum_{v\in V}(1-\alpha)\mathbb{E}_{p_v(x)}[[r_v^{\alpha}-\hat{f}_v]^2(x)] + \alpha\mathbb{E}_{q_v(x')}[[r_v^{\alpha}-\hat{f}_v]^2(x')], \end{aligned} \tag{51}$$

where the expected value $\mathbb{E}_{p^{\alpha}(y)}[[f_v - r_v^{\alpha}]^2(y)]$ is computed by averaging $10,000$ independent samples $\{(x_i, x_i')\}_{i=1}^{10000}$ that were not used during the training phase. α equals 0 for the LRE methods whose target function is the usual likelihood-ratio, and different from zero for methods which target the relative likelihood-ratio (See Tab. 2).

The MSE of Eq. 51 refers to the whole graph and it is the quantity in terms of which the convergence guarantees of GRULSIF (Theorem 3) and the theoretical results of Sec. 4 are given. Therefore, measuring the MSE can validate empirically those results. However,

for some applications, such as Hypothesis Testing and Change-Point Detection, the interest may be whether the graph is beneficial when comparing p_v and q_v. In such cases, the node-level ϕ-divergence estimates would be more informative, thus we also compare the node-level approximations with the true ϕ-divergence at each node. This information is summarized by box-plots and heatmaps in Fig. 2–16. The choice of the ϕ-divergence depends on the LRE method being used; details are given in Tab. 2 and in the self-contained figure captions.

Results and findings. The first batch of results for the designed scenarios are in Fig. 2-6. Two line-plots at the top of each figure show the convergence of the methods, in terms of the logarithm of the MSE, as a function of the sample size n, for a graph with $N = 100$ nodes. Recall that the sample size refers to the equal number of available observations from each of the pdfs of a node v, i.e. $n_v = n'_v = n$ from p_v and q_v, respectively. The line-plot on the right zooms into a smaller range of the y-axis to compare GRULSIF, POOL, and RULSIF, therefore to investigate the impact of using the graph. The displayed error band is for one standard deviation computed over 10 instances. Below the line-plots there is a grid of box-plots (sample size $\times$ method) for $n = \{50, 100, 250, 500\}$. Each box-plot shows the ϕ-divergence estimates within meaningful node groups for each scenario, i.e. node clusters C_1 to C_4, or the subset $C(u)$ and its complement $C(u)^\complement$. In these synthetic experiments, the true ϕ-divergence is identical for all nodes belonging to the same cluster, and the dashed lines indicate this value. In Fig. 4, where q_v varies according to the position of v in a Grid graph, the box-plots are replaced by heatmaps. The first row shows the true node-level ϕ-divergence, and the subsequent rows show the approximations for varying sample sizes.

In most of the experiments, POOL and GRULSIF show superior performance to the other methods. Even though POOL disregards the graph, same as ULSIF, RULSIF, and KLIEP do, it still shows better convergence behavior than those methods. Evidently, the introduction of a global non-redundant dictionary, the Nyström approximation, and the joint hyperparameter selection that we propose, boost the LRE performance when multiple sources of information are available and are all approximated by the same RKHS.

Concerning the effect of using the graph structure, when the smoothness hypothesis is satisfied and the likelihood-ratios are smooth over the graph, GRULSIF achieves a consistently better convergence and estimation quality compared to POOL, indicating the importance of the geometry of the problem. The advantage of GRULSIF becomes more clear as the sample size (n) at each node is smaller, which is expected as node-level tasks become more challenging. We see that GRULSIF's estimates have lower bias compared to the other methods, especially for nodes where $p_v = q_v$. This bias gets smaller as the sample size increases. However, as expected, collaboration may bring smaller gains when the available data at each node are sufficient to estimate the LRE independently, e.g. in Synth.Ia and Synth.Ib where POOL and GRULSIF have similar performance for $n = n' = 500$ (Fig. 3). For a thorough discussion, see Sec. 4.

The impact of the dimensionality (d) of the input space to Kernel-based LRE methods can be investigated by comparing Fig. 5 (Synth.IIb, $d = 4$) and Fig. 6 (Synth.IIc, $d = 20$). The box-plots show that the bias of all methods increases as d grows, with GRULSIF and POOL showing a smaller bias. The convergence of GRULSIF and POOL with the sample size is slower in higher dimensions due to requiring larger dictionaries. This is expected as the size of the dictionary built by the strategy we employ (Richard et al., 2009) depends on the covering number of the RKHS. For Gaussian kernels, the covering number scales with

d. This may explain why RULSIF, ULSIF, and KLIEP, which fix the size of the dictionary independently of the kernel, show a higher bias. In conclusion, when the dimensions of the input space reduces the smoothness of the likelihood-ratios w.r.t. the RKHS, Kernel-based LRE methods converge slower, and adaptive dimensionality reduction strategies are necessary. This highlights the need for further research in the high dimensional regime.

The role of α-regularization. To investigate the sensitivity of GRULSIF and POOL to the regularization parameter α, we complement the previous experiments by reporting in Fig. 7-11 results for $\alpha = \{0.01, 0.1, 0.5\}$. We can see that tuning α affects convergence, as suggested by Theorem 3. Low α values make the LRE task harder, hence leads to estimates with higher bias and variance, and slower convergence to the true target quantities (this is more evident in the box-plots). Moreover, graph regularization leads to more robust node-level estimates, i.e. lower variance within sets of connected nodes and faster convergence, especially for nodes where $p_v = q_v$. Finally, when α is closer to 1, GRULSIF and POOL get closer since the target relative likelihood-ratios become easier to estimate even without collaboration. The findings show that the Collaborative LRE is more robust as α gets closer to 0, when targeting a less regularized likelihood-ratio (see also Sec. 5.4).

The role of the graph size (N). The number of nodes affects GRULSIF's performance, both numerically and in terms of its generalization properties. For the first point, notice that under the hypothesis that the size of the dictionary ($\hat{L}$) is fixed and the sample size at each node remains the same ($n_v = n'_v = n$), GRULSIF's complexity scales linearly with N, specifically at a rate of $\mathcal{O}(N\hat{L}^3 + nN\hat{L}^2 + N\hat{L}^2 \log^2(nN\hat{L}))$ (see Sec. 5.2).

Regarding the impact of N on the generalization properties of GRULSIF, notice that the bound of Theorem 3 is dominated by the term: $N^{\frac{-1}{1+\zeta^*}} \left(\mathcal{T}_G \Lambda^2\right)^{\frac{1}{1+\zeta^*}}$, and each term depends on N. Expr. 16 tells us that Λ increases as more likelihood-ratios need to be estimated, while $N^{\frac{-1}{1+\zeta^*}}$ decreases. On the other hand, the new nodes associated to added LRE tasks get wired into the graph and modify its topology, hence affect $\mathcal{T}_G$. As a result, either the term $\left(\mathcal{T}_G \Lambda^2\right)^{\frac{1}{1+\zeta^*}}$ dominates or increases at the same rate as $N^{\frac{-1}{1+\zeta^*}}$, which implies that increasing N would not help GRULSIF to converge faster; or the term $N^{\frac{-1}{1+\zeta^*}}$ dominates, in which case increasing N would be beneficial. These distinct convergence regimes are observed in the example provided in Sec. 4, where the effect of N differs between a completely disconnected graph and a fully connected graph, despite the LRE tasks being the same for all the nodes.

As Λ is an a priori unknown parameter, our current results do not allow the automatic identification of the convergence regime as we increase the number of nodes. To demonstrate the effect of node number, we compare GRULSIF and POOL in synthetic scenarios with varying graph sizes ($N = \{100, 500, 1000\}$). The change of measure at each node is detailed in Tab. 1: in Synth.Ia and Synth.Ib we increase the number of nodes per cluster, while in Synth.IIa, Synth.IIb, and Synth.IIc we extend proportionally a Grid graph. The results are reported in Fig. 12-16. As expected, the size of the graph does not affect all scenarios in the same way. In Synth.IIc (Fig. 16), we observe that increasing N leads to faster GRULSIF convergence. Contrary, in Synth.IIb (Fig. 15), including more nodes increases the bias of GRULSIF. POOL appears unaffected by the variation of N across all scenarios, as it fits a model at each node independently and hence its convergence rate depends only on the sample size (n).

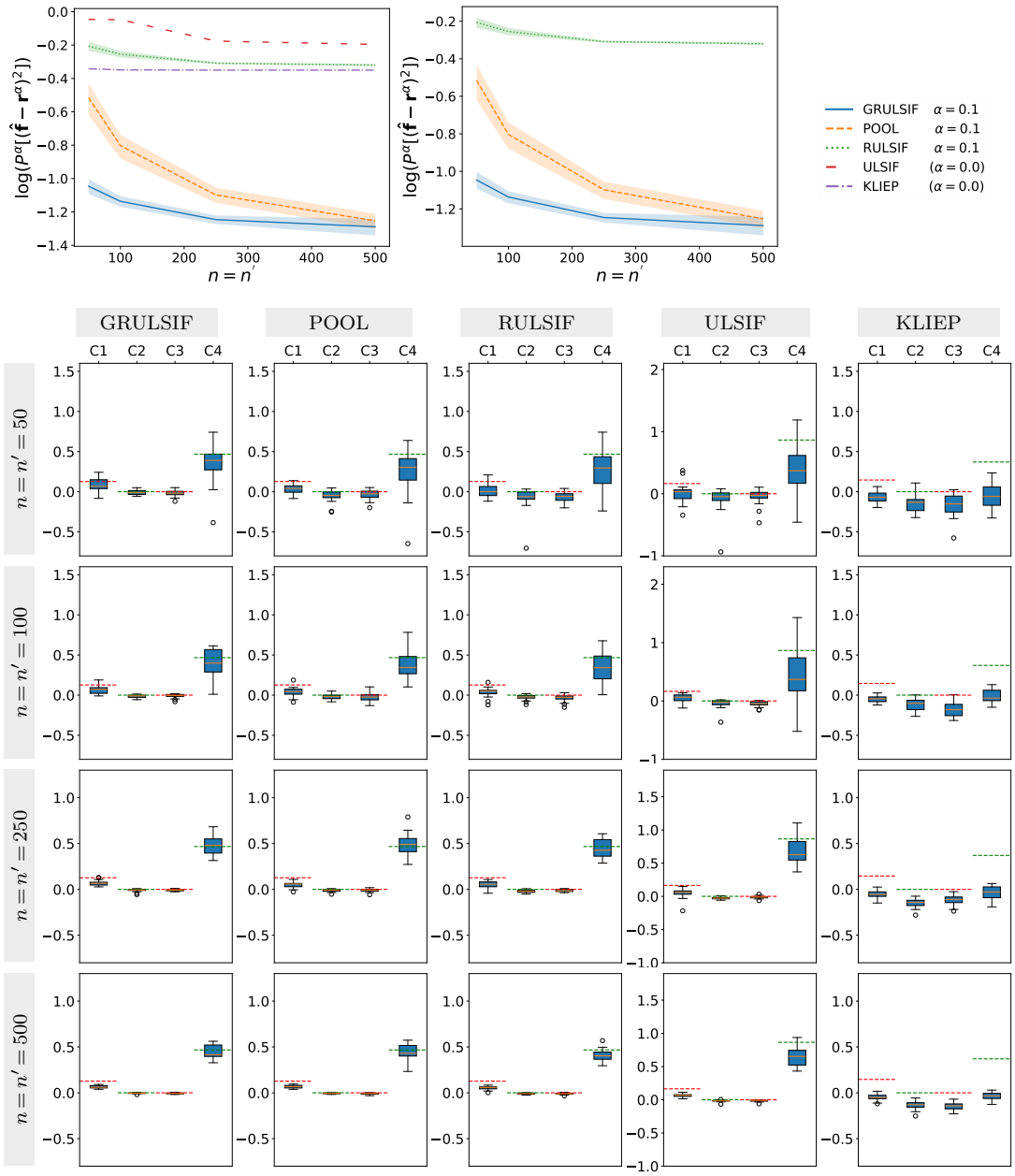

Figure 2: Experiment Synth.Ia. Comparison of GRULSIF and POOL, with fixed $\alpha = 0.1$ and varying sample size $n = \{50, 100, 250, 500\}$ (i.e. n from p_v and n' from q_v, with $n = n'$), against existing LRE approaches. All methods are built upon the χ^2-divergence between p_v^α and q_v, except from KLIEP that uses the KL-divergence. GRULSIF, POOL, and RULSIF target the relative likelihood-ratio with $\alpha = 0.1$, while ULSIF and KLIEP target the original likelihood-ratio ($\alpha = 0$). The graph size is fixed at $N = 100$. **Line-plots:** Convergence to the respective target true likelihood-ratios, in terms of the $\log\left(P^\alpha\left[[\mathbf{f} - \mathbf{r}^\alpha]^2\right]\right)$ (see Eq. 51), as a function of the sample size n (i.e. n from p_v and n' from q_v, with $n = n'$). **Box-plots:** The distribution of node-level ϕ-divergence estimates obtained by each method for varying sample size $n = n'$. The horizontal dashed lines (red and green) indicate the target true ϕ-divergence within each cluster.

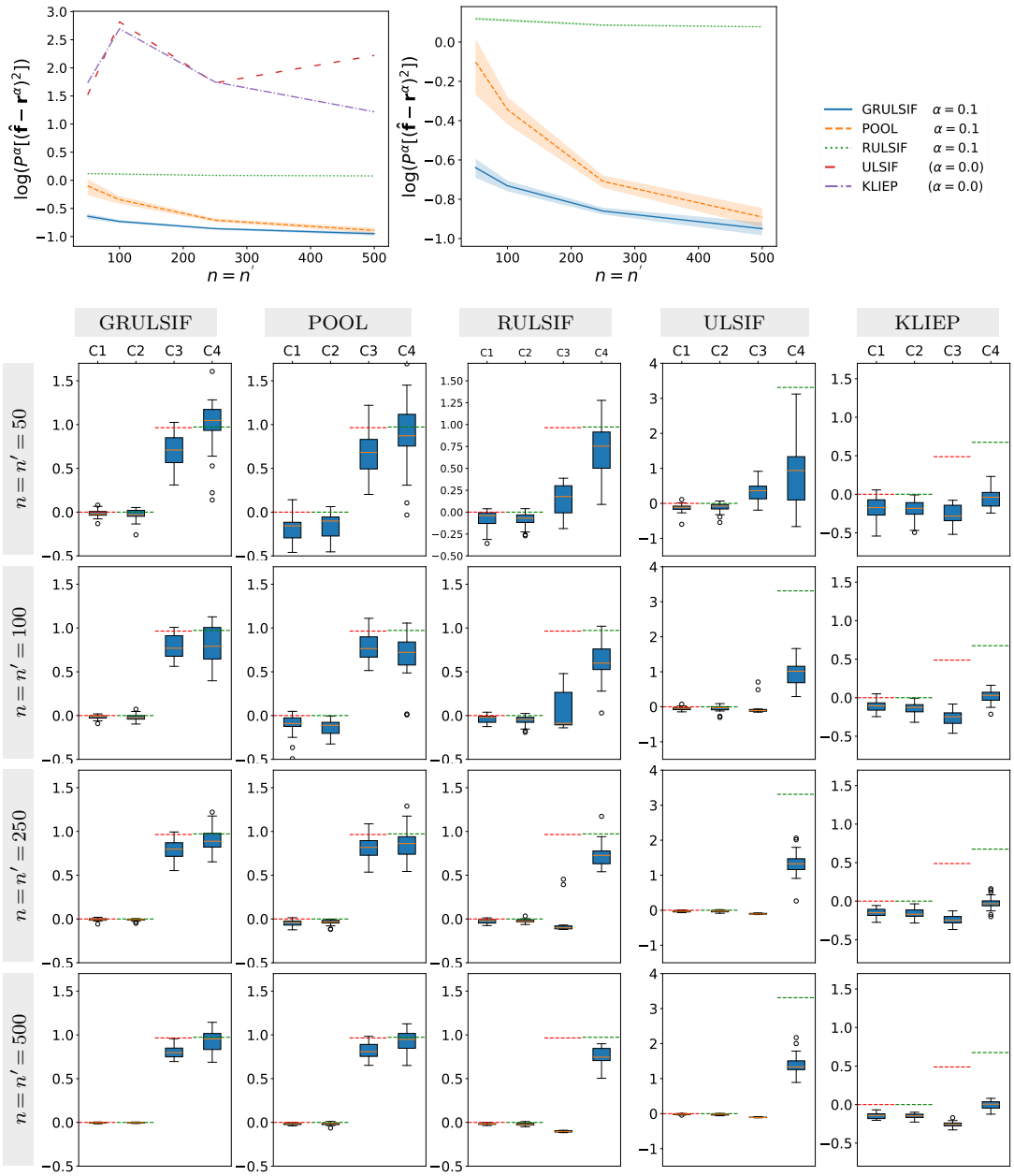

Figure 3: Experiment Synth.Ib. Comparison of GRULSIF and POOL, with fixed $\alpha = 0.1$ and varying sample size $n = \{50, 100, 250, 500\}$ (i.e. n from p_v and n' from q_v, with $n = n'$), against existing LRE approaches. All methods are built upon the χ^2-divergence between p_v^α and q_v, except from KLIEP that uses the KL-divergence. GRULSIF, POOL, and RULSIF target the relative likelihood-ratio with $\alpha = 0.1$, while ULSIF and KLIEP target the original likelihood-ratio ($\alpha = 0$). The graph size is fixed at $N = 100$. **Line-plots:** Convergence to the respective target true likelihood-ratios, in terms of the $\log\left(P^\alpha\left[[\mathbf{f} - \mathbf{r}^\alpha]^2\right]\right)$ (see Eq. 51), as a function of the sample size n (i.e. n from p_v and n' from q_v, with $n = n'$). **Box-plots:** The distribution of node-level ϕ-divergence estimates obtained by each method for varying sample size $n = n'$. The horizontal dashed lines (red and green) indicate the target true ϕ-divergence within each cluster.

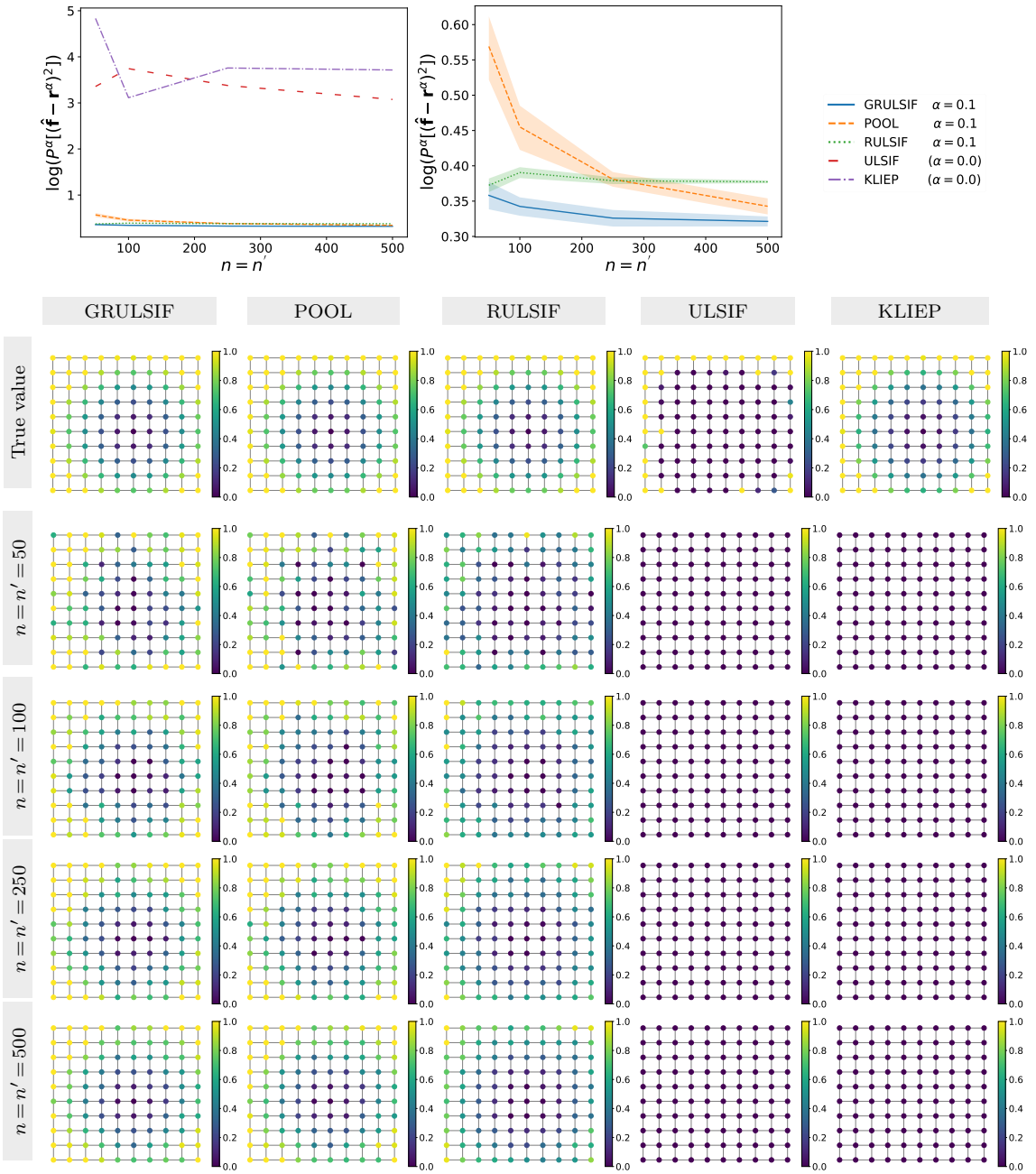

Figure 4: Experiment Synth.IIa. Comparison of GRULSIF and POOL, with fixed $\alpha = 0.1$ and varying sample size $n = \{50, 100, 250, 500\}$ (i.e. n from p_v and n' from q_v, with $n = n'$), against existing LRE approaches. All methods are built upon the χ^2-divergence between p_v^α and q_v, except from KLIEP that uses the KL-divergence. GRULSIF, POOL, and RULSIF target the relative likelihood-ratio with $\alpha = 0.1$, while ULSIF and KLIEP target the original likelihood-ratio ($\alpha = 0$). The graph size is fixed at $N = 10 \times 10$. **Line-plots:** Convergence to the respective target true likelihood-ratios, in terms of the $\log\left(P^\alpha\left[[\mathbf{f} - \mathbf{r}^\alpha]^2\right]\right)$ (see Eq. 51), as a function of the sample size n (i.e. n from p_v and n' from q_v, with $n = n'$). **Heatmaps:** In this experiment, q_v depends on the position of the node in the Grid graph, and so does their true ϕ-divergence value. In the first row, a heatmap shows the true node-level ϕ-divergence associated to each method. The heatmaps in the following rows show the node-level ϕ-divergence estimates obtained by each method.

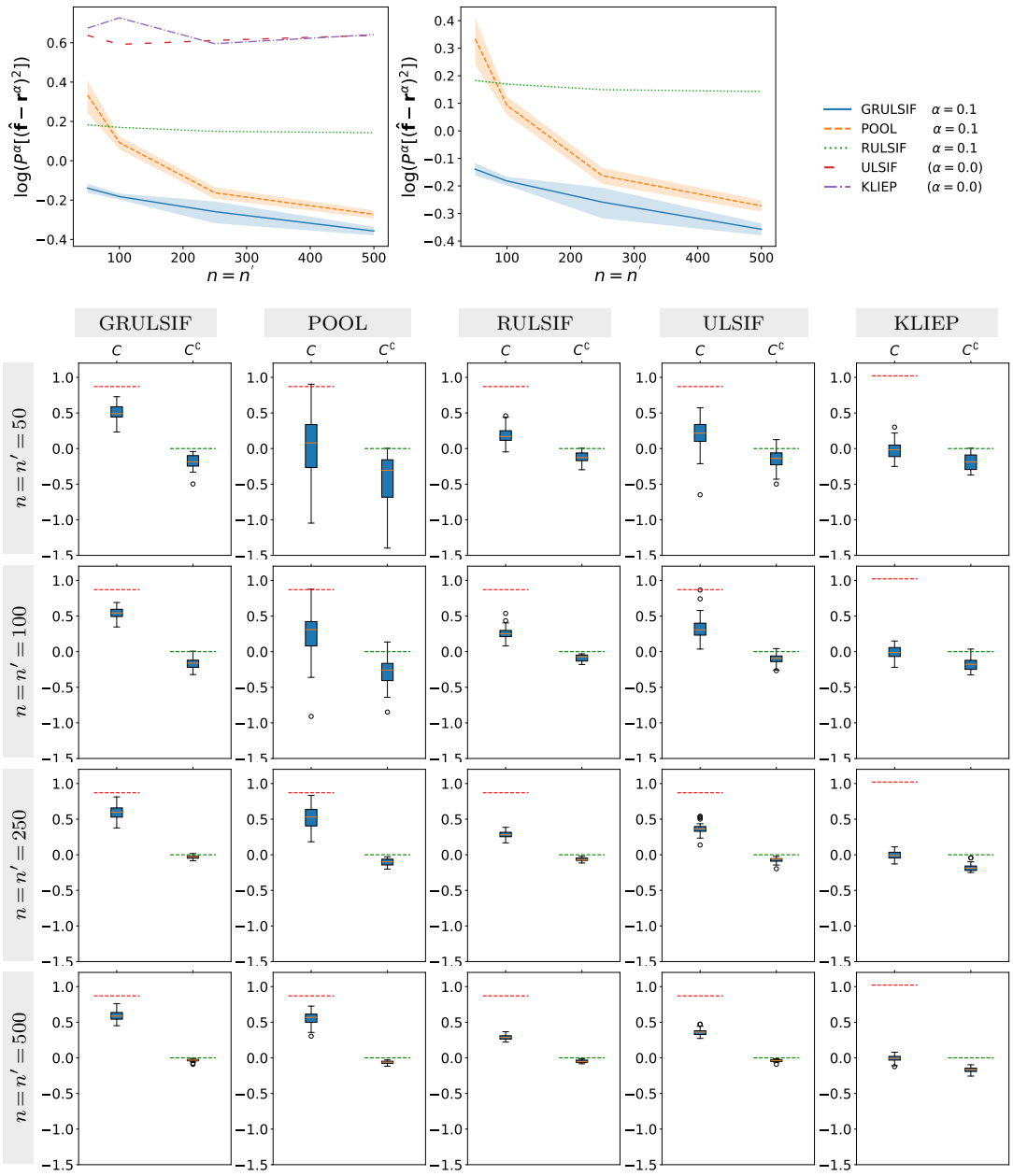

Figure 5: Experiment Synth.IIb. Comparison of GRULSIF and POOL, with fixed $\alpha = 0.1$ and varying sample size $n = \{50, 100, 250, 500\}$ (i.e. n from p_v and n' from q_v, with $n = n'$), against existing LRE approaches. All methods are built upon the χ^2-divergence between p_v^α and q_v, except from KLIEP that uses the KL-divergence. GRULSIF, POOL, and RULSIF target the relative likelihood-ratio with $\alpha = 0.1$, while ULSIF and KLIEP target the original likelihood-ratio ($\alpha = 0$). The graph size is fixed at $N = 100$. **Line-plots:** Convergence to the respective target true likelihood-ratios, in terms of the $\log\left(P^\alpha\left[[\mathbf{f} - \mathbf{r}^\alpha]^2\right]\right)$ (see Eq. 51), as a function of the sample size n (i.e. n from p_v and n' from q_v, with $n = n'$). **Box-plots:** The distribution of node-level ϕ-divergence estimates obtained by each method for varying sample size $n = n'$. The horizontal dashed lines (red and green) indicate the target true ϕ-divergence within each node group of interest.

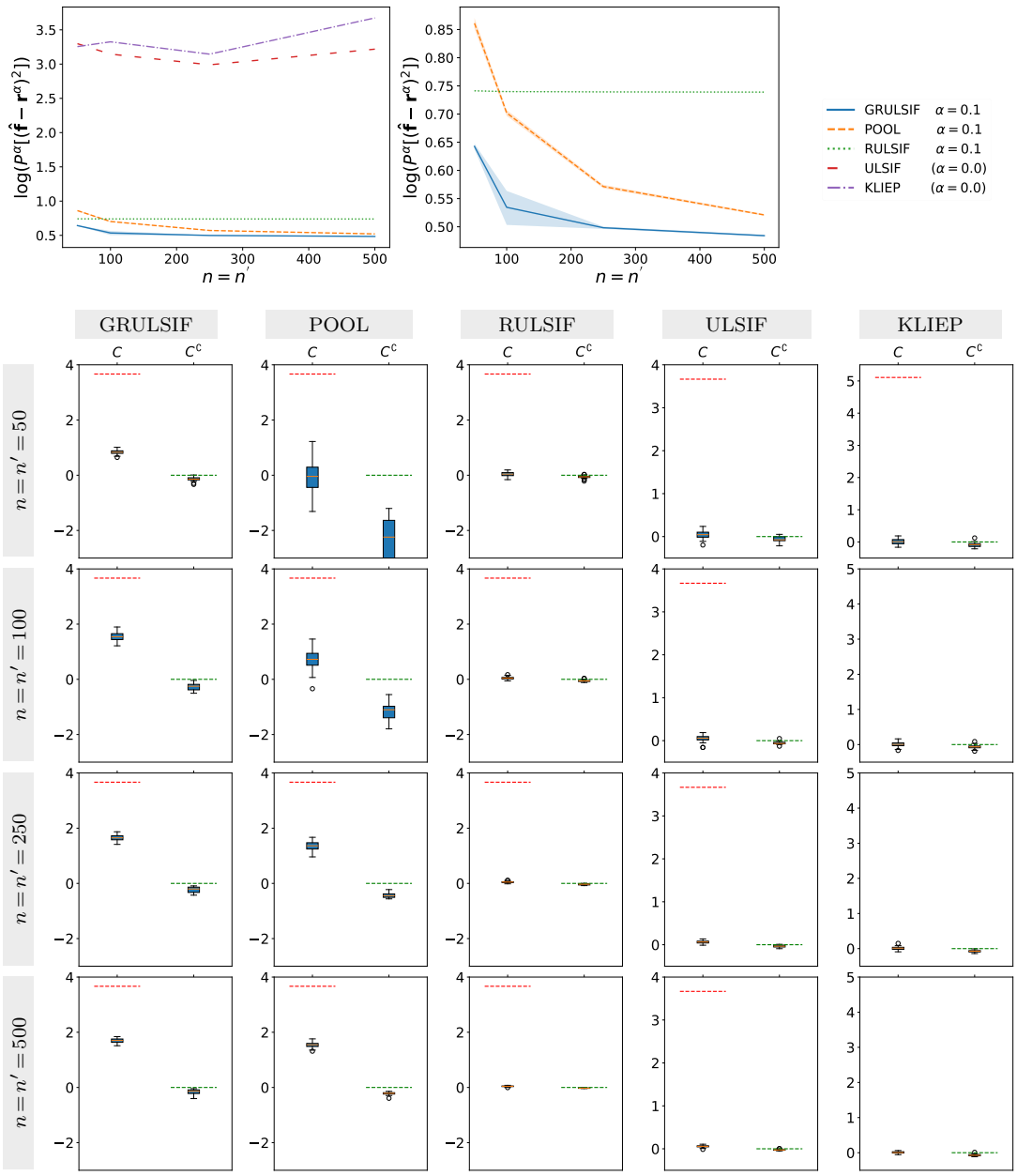

Figure 6: Experiment Synth.IIc. Comparison of GRULSIF and POOL, with fixed $\alpha = 0.1$ and varying sample size $n = \{50, 100, 250, 500\}$ (i.e. n from p_v and n' from q_v, with $n = n'$), against existing LRE approaches. All methods are built upon the χ^2-divergence between p_v^α and q_v, except from KLIEP that uses the KL-divergence. GRULSIF, POOL, and RULSIF target the relative likelihood-ratio with $\alpha = 0.1$, while ULSIF and KLIEP target the original likelihood-ratio ($\alpha = 0$). The graph size is fixed at $N = 100$. **Line-plots:** Convergence to the respective target true likelihood-ratios, in terms of the $\log\left(P^\alpha\left[[\mathbf{f} - \mathbf{r}^\alpha]^2\right]\right)$ (see Eq. 51), as a function of the sample size n (i.e. n from p_v and n' from q_v, with $n = n'$). **Box-plots:** The distribution of node-level ϕ-divergence estimates obtained by each method for varying sample size $n = n'$. The horizontal dashed lines (red and green) indicate the target true ϕ-divergence within each node group of interest.

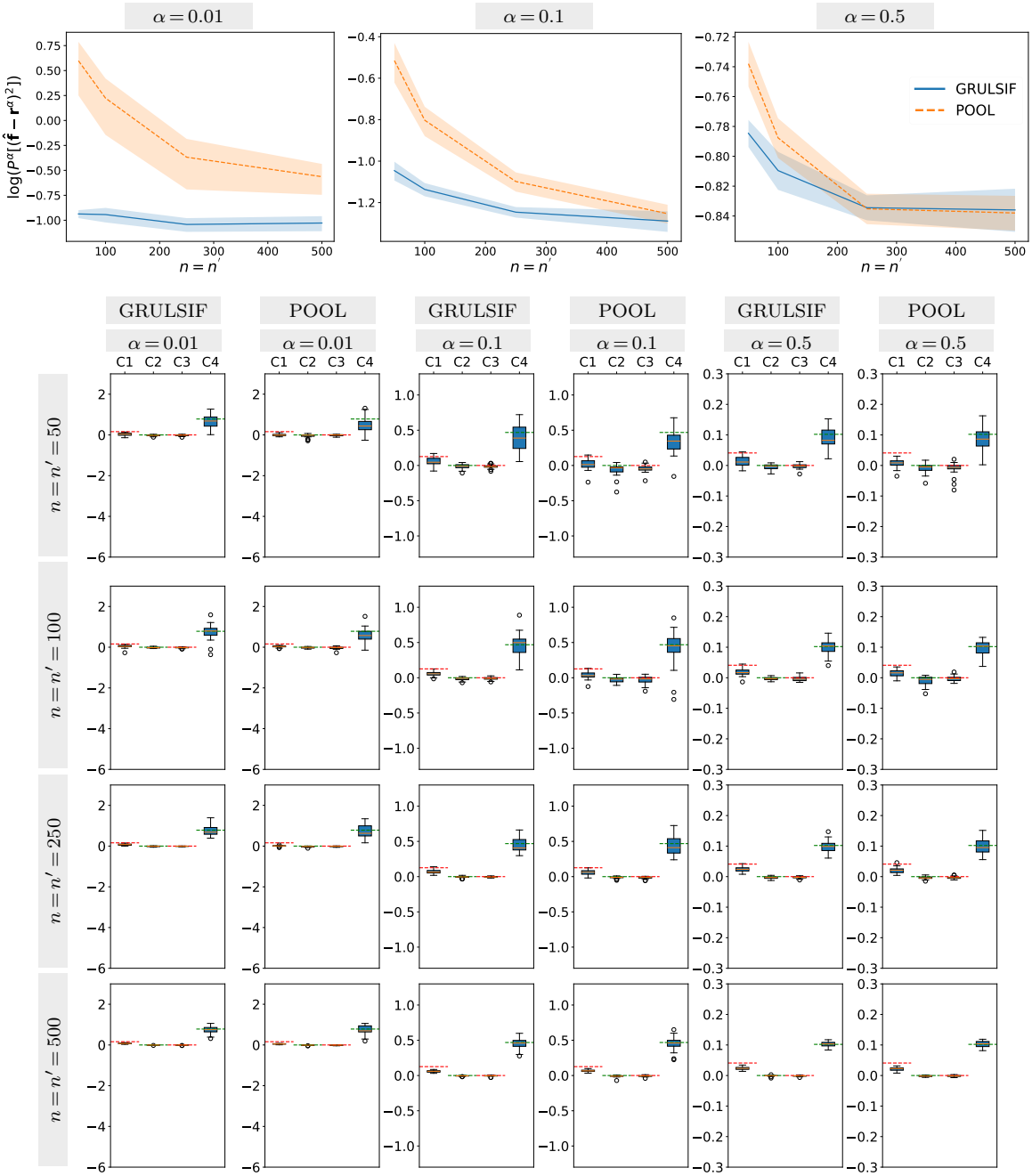

Figure 7: Experiment Synth.Ia for varying α-regularization. Complement of Fig. 2 that focuses on the behavior of GRULSIF and POOL for varying $\alpha = \{0.01, 0.1, 0.5\}$. The graph size is fixed at $N = 100$. **Line-plots:** Convergence to the true relative likelihood-ratio, in terms of the $\log\left(P^{\alpha}\left[[\mathbf{f} - \mathbf{r}^{\alpha}]^2\right]\right)$ (see Eq. 51), as a function of the sample size n (i.e. n from p_v and n' from q_v, with $n = n'$) and α. **Box-plots:** The distribution of node-level estimates, $\{\hat{PE}_v^{\alpha}(X_v \| X'_v)\}_{v \in V}$ (See Eq. 30), obtained for varying α and sample size $n = n'$. The horizontal dashed lines (red and green) indicate the true $PE(p_v^{\alpha} \| q_v)$ (See Eq. 32) within each cluster.

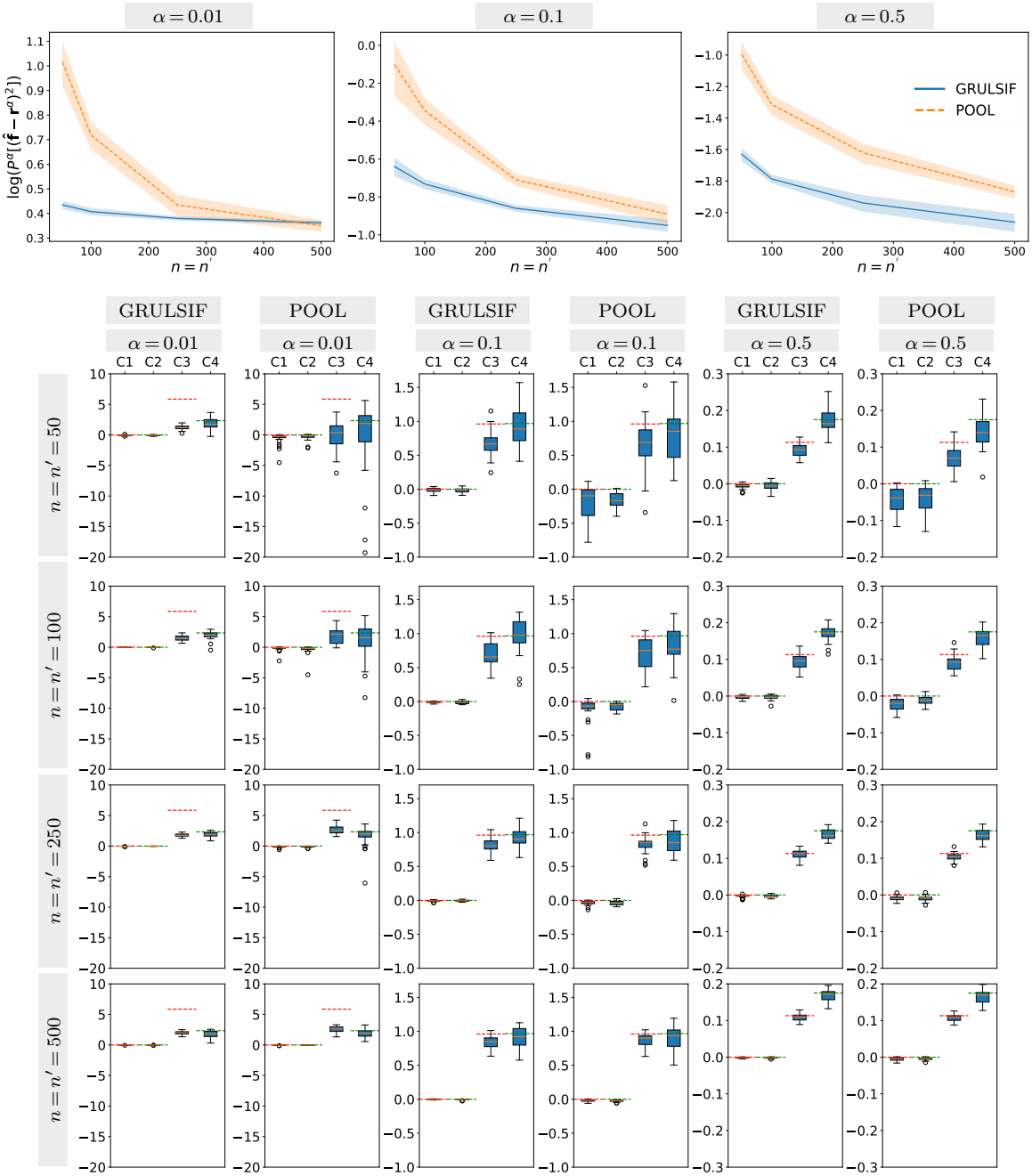

Figure 8: Experiment Synth.Ib for varying α-regularization. Complement of Fig. 3 that focuses on the behavior of GRULSIF and POOL for varying $\alpha = \{0.01, 0.1, 0.5\}$. The graph size is fixed at $N = 100$. **Line-plots:** Convergence to the true relative likelihood-ratio, in terms of the $\log\left(P^{\alpha}\left[[\mathbf{f} - \mathbf{r}^{\alpha}]^2\right]\right)$ (see Eq. 51), as a function of the sample size n (i.e. n from p_v and n' from q_v, with $n = n'$) and α. **Box-plots:** The distribution of node-level estimates, $\{\hat{PE}_v^{\alpha}(X_v \| X'_v)\}_{v \in V}$ (See Eq. 30), obtained for varying α and sample size $n = n'$. The horizontal dashed lines (red and green) indicate the true $PE(p_v^{\alpha} \| q_v)$ (See Eq. 32) within each cluster.

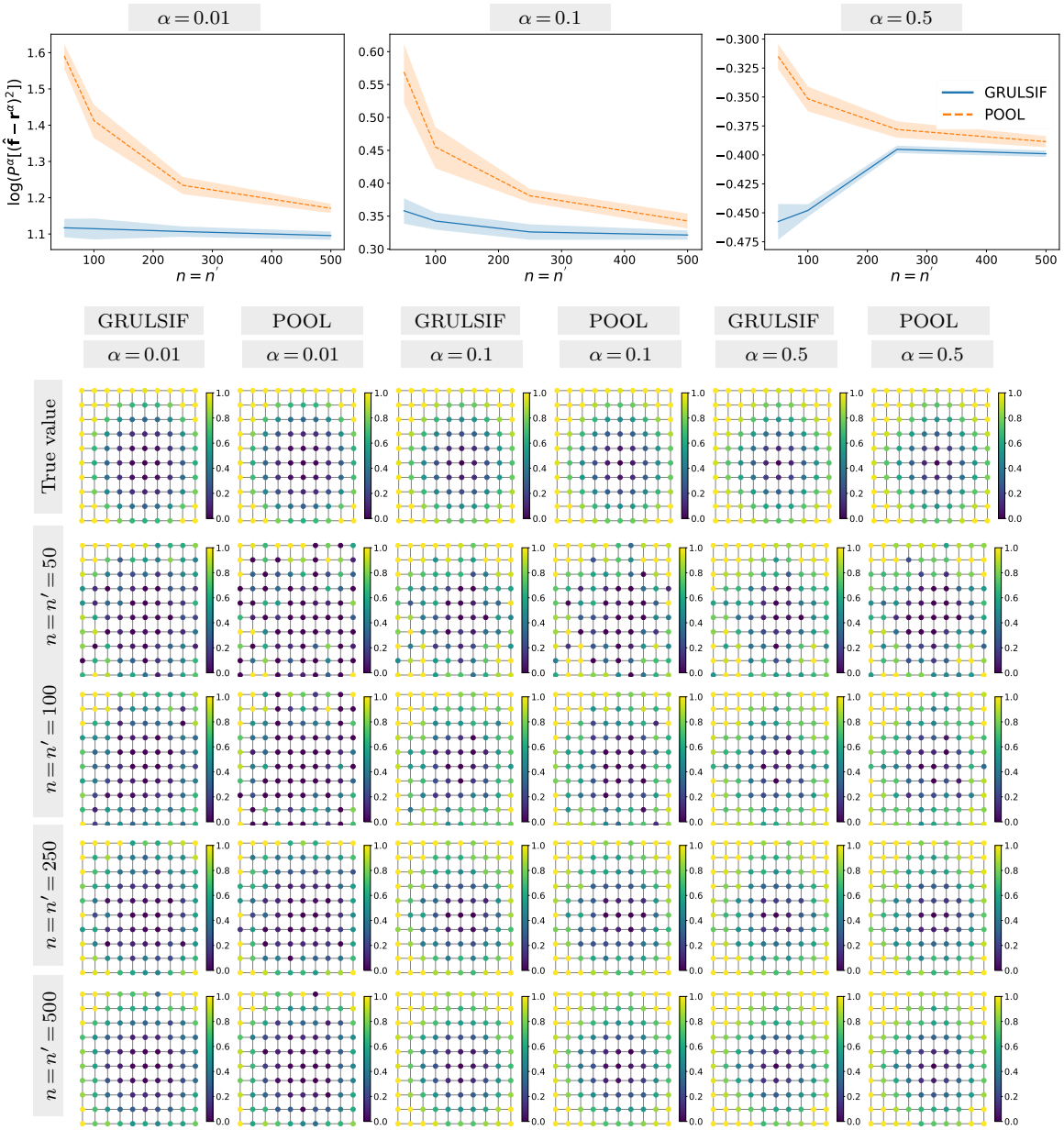

Figure 9: Experiment Synth.IIa for varying α-regularization. Complement of Fig. 4 that focuses on the behavior of GRULSIF and POOL for varying $\alpha = \{0.01, 0.1, 0.5\}$. The graph size is fixed at $N = 100$. **Line-plots:** Convergence to the true relative likelihood-ratio, in terms of the $\log\left(P^{\alpha}\left[[\mathbf{f} - \mathbf{r}^{\alpha}]^2\right]\right)$ (see Eq. 51), as a function of the sample size n (i.e. n from p_v and n' from q_v, with $n = n'$) and α. **Box-plots:** The distribution of node-level estimates obtained for varying α and sample size $n = n'$. **Heatmaps:** In this experiment, q_v depends on the position of the node in the Grid graph, and so does $PE(p_v^{\alpha} \| q_v)$ (See Eq. 32). In the first row, a heatmap shows the true node-level true value of each method. The heatmaps in the following rows show the node-level estimates, $\{\hat{PE}_v^{\alpha}(X_v \| X'_v)\}_{v \in V}$ (See Eq. 30), obtained for varying α and sample size $n = n'$.

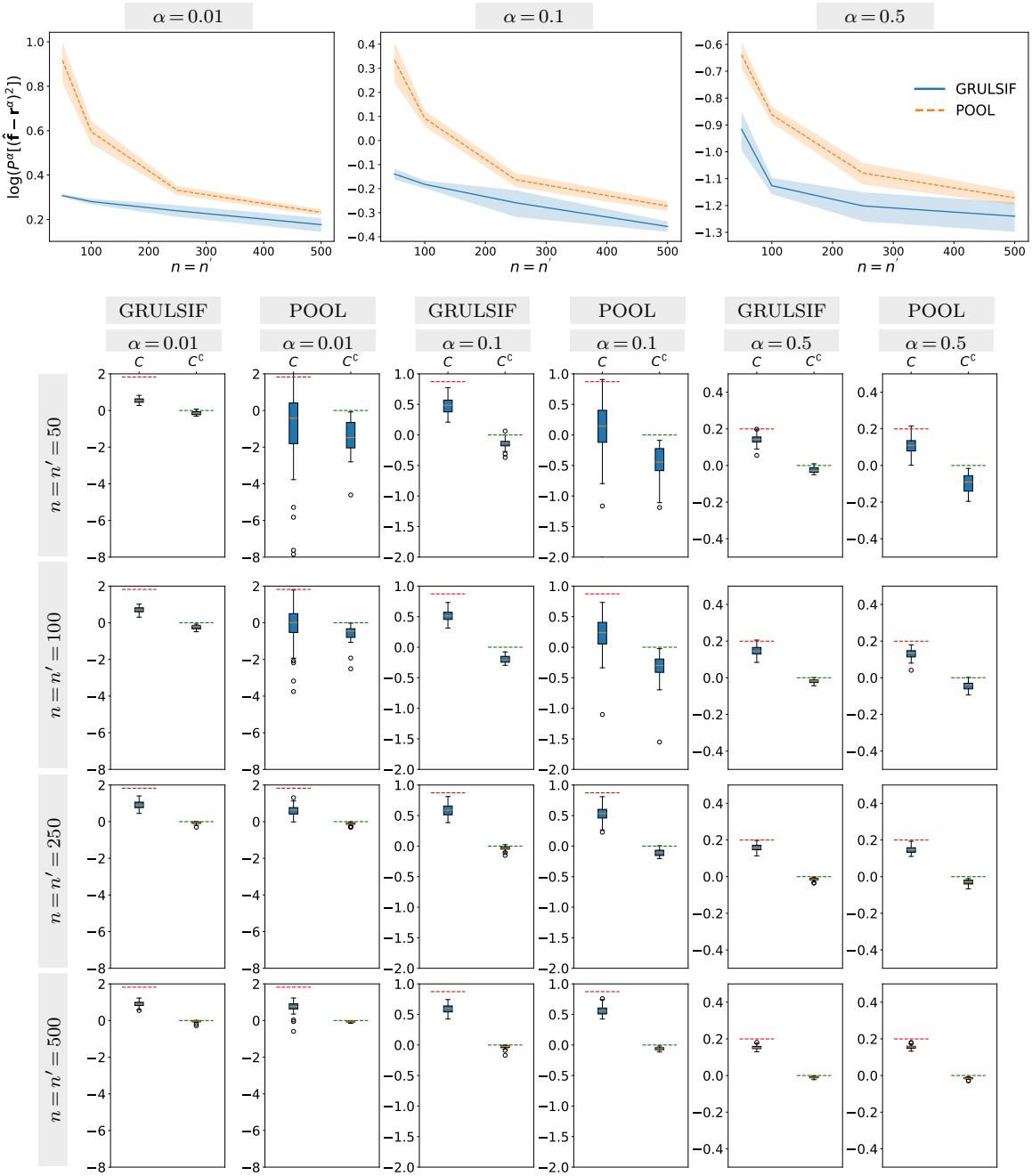

Figure 10: Experiment Synth.IIb for varying α-regularization. Complement of Fig. 5 that focuses on the behavior of GRULSIF and POOL for varying $\alpha = \{0.01, 0.1, 0.5\}$. The graph size is fixed at $N = 100$. **Line-plots:** Convergence to the true relative likelihood-ratio, in terms of the $\log\left(P^{\alpha}\left[[\mathbf{f} - \mathbf{r}^{\alpha}]^2\right]\right)$ (see Eq. 51), as a function of the sample size n (i.e. n from p_v and n' from q_v, with $n = n'$) and α. **Box-plots:** The distribution of node-level estimates, $\{\hat{PE}_v^{\alpha}(X_v \| X'_v)\}_{v \in V}$ (See Eq. 30), obtained for varying α and sample size $n = n'$. The horizontal dashed lines (red and green) indicate the true $PE(p_v^{\alpha} \| q_v)$ (See Eq. 32) within each node group of interest.

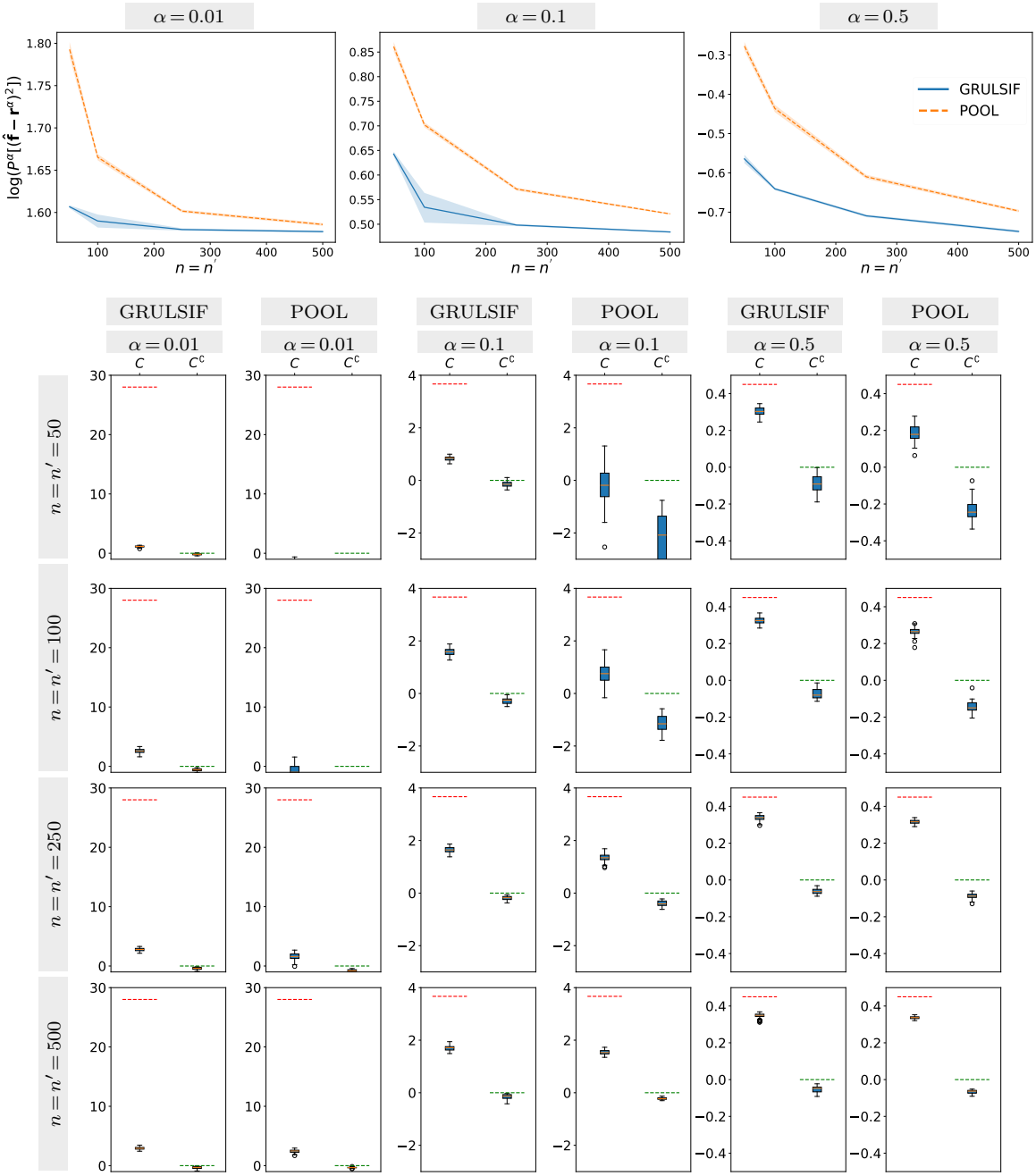

Figure 11: Experiment Synth.IIc for varying α-regularization. Complement of Fig. 6 that focuses on the behavior of GRULSIF and POOL for varying $\alpha = \{0.01, 0.1, 0.5\}$. The graph size is fixed at $N = 100$. **Line-plots:** Convergence to the true relative likelihood-ratio, in terms of the $\log\left(P^{\alpha}\left[[\mathbf{f} - \mathbf{r}^{\alpha}]^2\right]\right)$ (see Eq. 51), as a function of the sample size n (i.e. n from p_v and n' from q_v, with $n = n'$) and α. **Box-plots:** The distribution of node-level estimates, $\{\hat{PE}_v^{\alpha}(X_v \| X_v')\}_{v \in V}$ (See Eq. 30), obtained for varying α and sample size $n = n'$. The horizontal dashed lines (red and green) indicate the true $PE(p_v^{\alpha} \| q_v)$ (See Eq. 32) within each node group of interest.

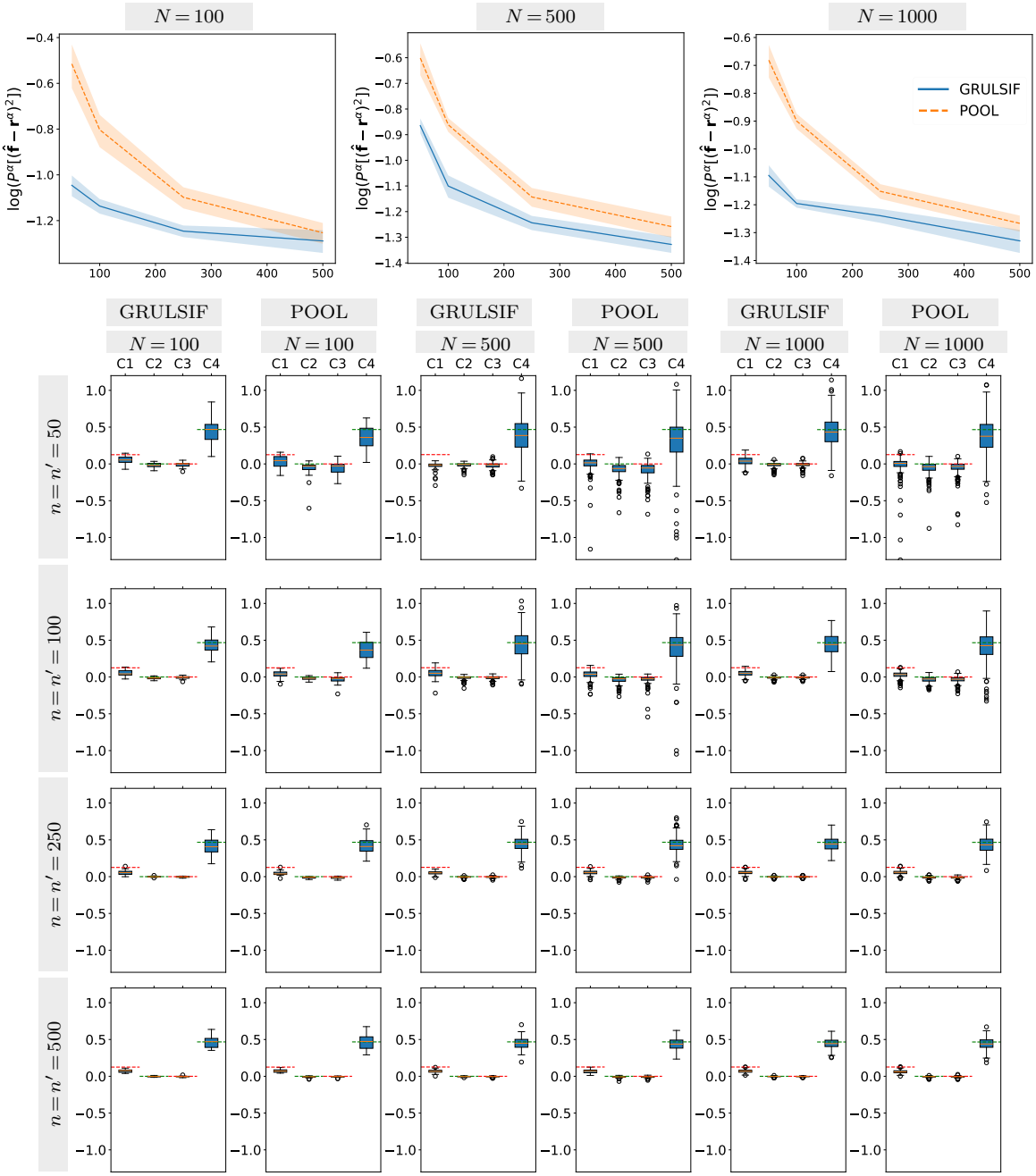

Figure 12: Experiment Synth.Ia for varying graph size. Complement of Fig. 2 that focuses on the behavior of GRULSIF and POOL for varying $N = \{100, 500, 1000\}$. The regularization parameter is fixed at $\alpha = 0.1$. **Line-plots:** Convergence to the true relative likelihood-ratio, in terms of the $\log\left(P^{\alpha}\left[[\mathbf{f} - \mathbf{r}^{\alpha}]^2\right]\right)$ (see Eq. 51), as a function of the sample size n (i.e. n from p_v and n' from q_v, with $n = n'$) and α. **Box-plots:** The distribution of node-level estimates, $\{\hat{PE}_v^{\alpha}(X_v \| X'_v)\}_{v \in V}$ (See Eq. 30), obtained for varying N and sample size $n = n'$. The horizontal dashed lines (red and green) indicate the true value of $PE(p_v^{\alpha} \| q_v)$ (See Eq. 32) within each cluster.

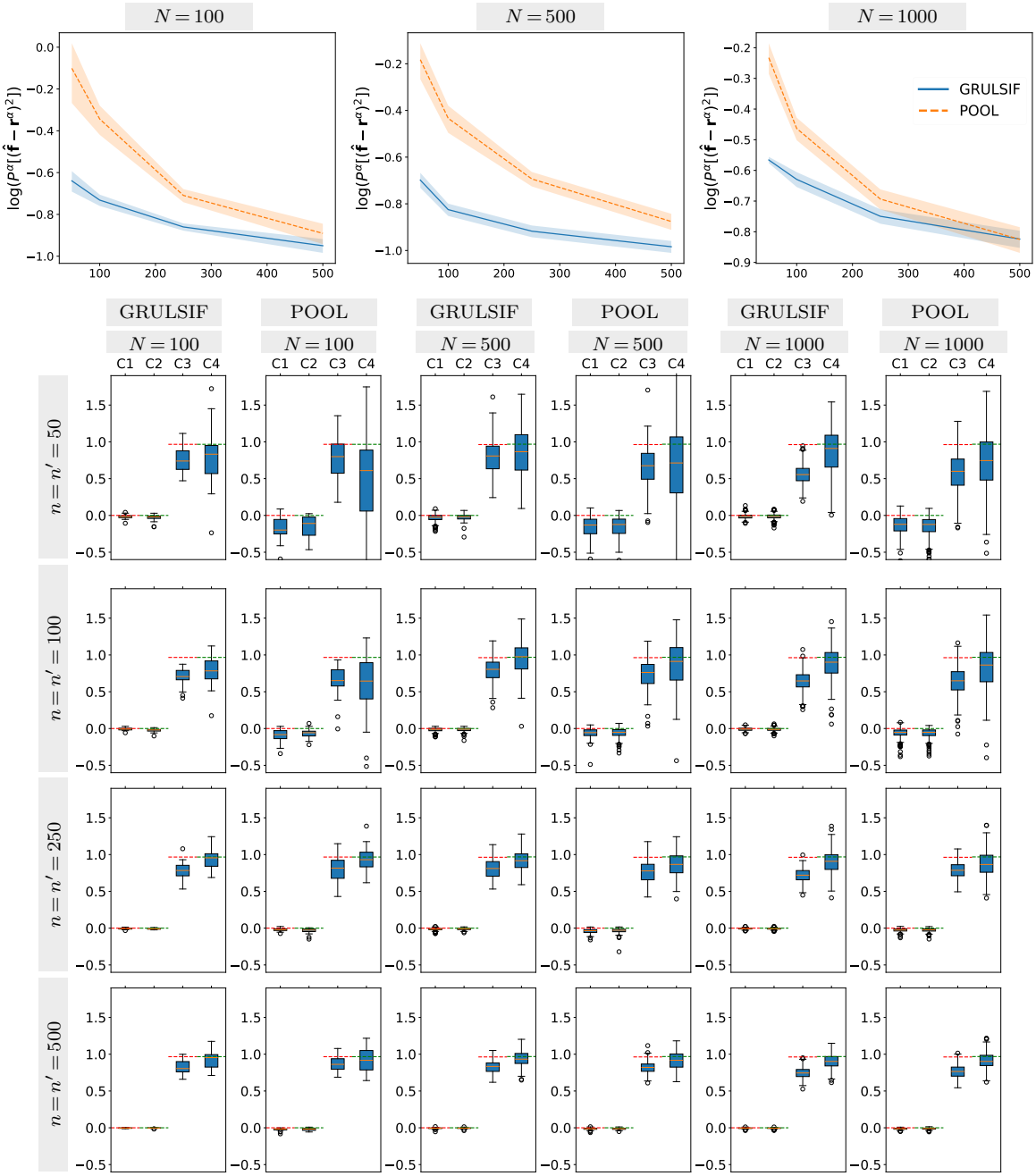

Figure 13: Experiment Synth.Ib for varying graph size. Complement of Fig. 3 that focuses on the behavior of GRULSIF and POOL for varying $N = \{100, 500, 1000\}$. The regularization parameter is fixed at $\alpha = 0.1$. **Line-plots:** Convergence to the true relative likelihood-ratio, in terms of the $\log\left(P^{\alpha}\left[[\mathbf{f} - \mathbf{r}^{\alpha}]^2\right]\right)$ (see Eq. 51), as a function of the sample size n (i.e. n from p_v and n' from q_v, with $n = n'$) and α. **Box-plots:** The distribution of node-level estimates, $\{\hat{PE}_v^{\alpha}(X_v \| X'_v)\}_{v \in V}$ (See Eq. 30), obtained for varying N and sample size $n = n'$. The horizontal dashed lines (red and green) indicate the true value of $PE(p_v^{\alpha} \| q_v)$ (See Eq. 32) within each cluster.

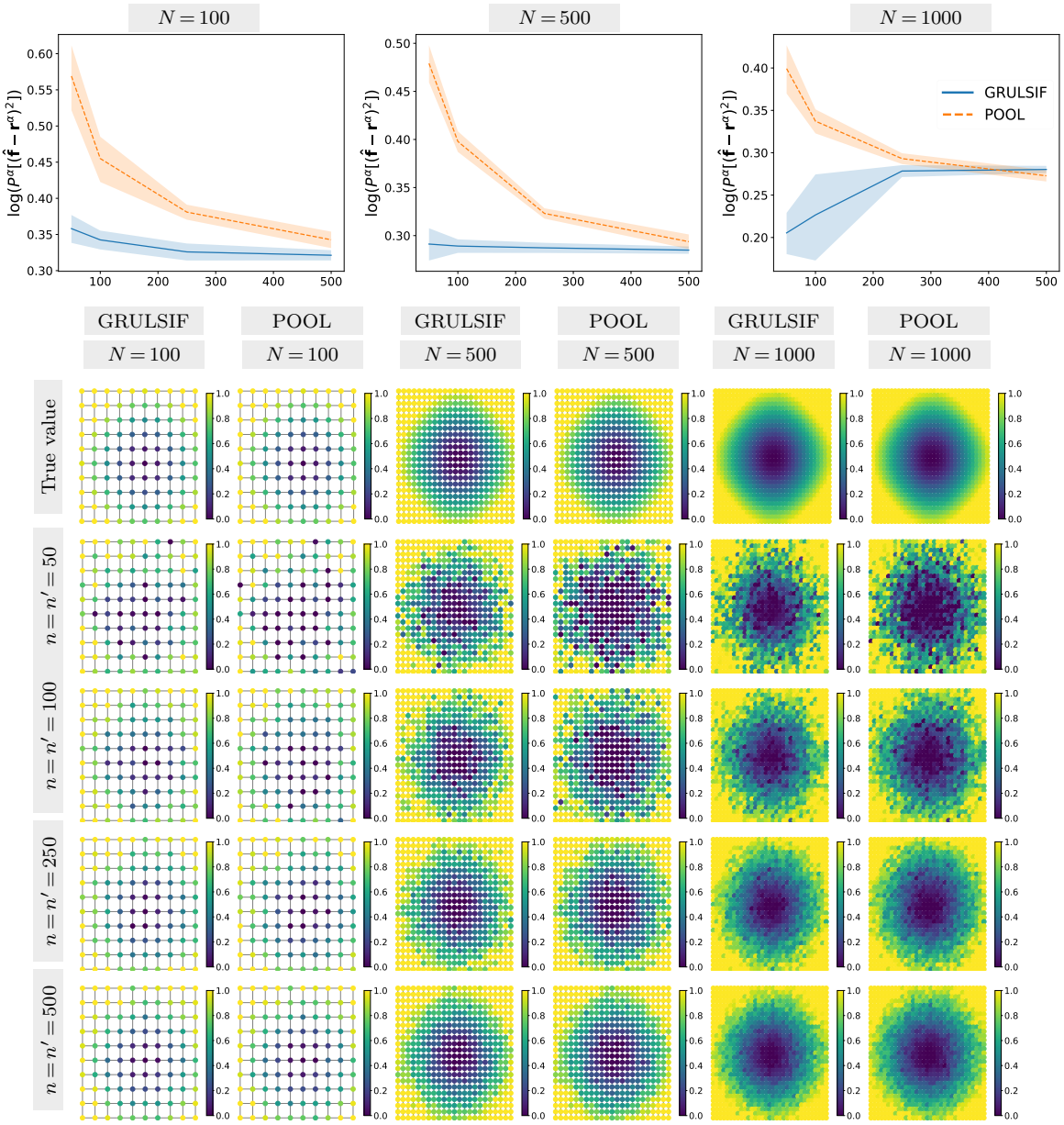

Figure 14: Experiment Synth.IIa for varying graph size. Complement of Fig. 4 that focuses on the behavior of GRULSIF and POOL for varying $N = \{100, 500, 1000\}$ (i.e. from a 10×10 to a 100×100 Grid graph). The regularization parameter is fixed at $\alpha = 0.1$. **Line-plots:** Convergence to the true relative likelihood-ratio, in terms of the $\log\left(P^{\alpha}\left[[\mathbf{f} - \mathbf{r}^{\alpha}]^2\right]\right)$ (see Eq. 51), as a function of the sample size n (i.e. n from p_v and n' from q_v, with $n = n'$) and α. **Heatmaps:** In this experiment, q_v depends on the position of the node in the Grid graph, and so does $PE(p_v^{\alpha} \| q_v)$ (See Eq. 32). In the first row, a heatmap shows the true node-level true value of each method. The heatmaps in the following rows show the node-level estimates, $\{\hat{PE}_v^{\alpha}(X_v \| X'_v)\}_{v \in V}$ (See Eq. 30), obtained for varying N and sample size $n = n'$.

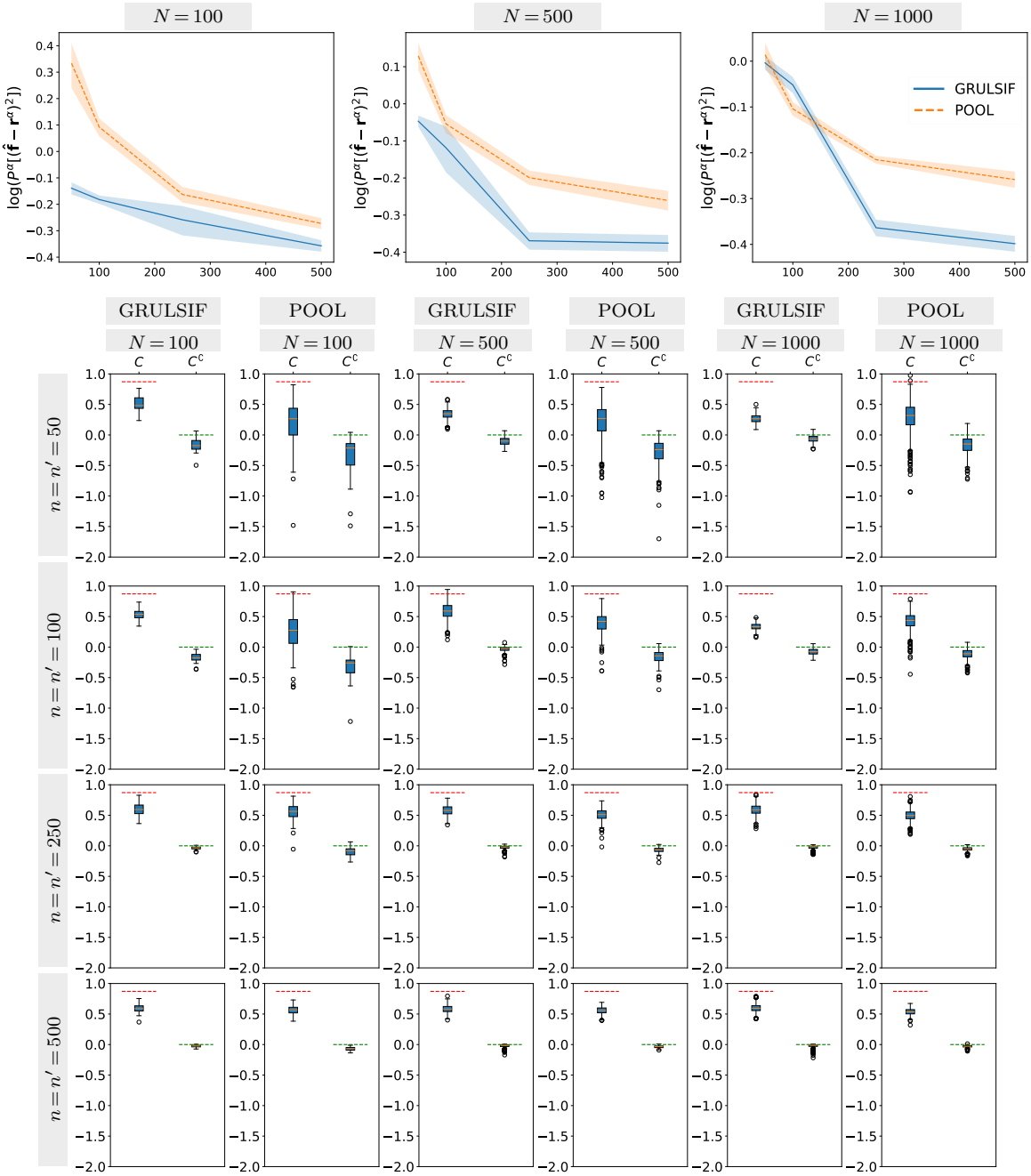

Figure 15: Experiment Synth.IIb for varying graph size. Complement of Fig. 5 that focuses on the behavior of GRULSIF and POOL for varying $N = \{100, 500, 1000\}$. The regularization parameter is fixed at $\alpha = 0.1$. **Line-plots:** Convergence to the true relative likelihood-ratio, in terms of the $\log\left(P^{\alpha}\left[[\mathbf{f} - \mathbf{r}^{\alpha}]^2\right]\right)$ (see Eq. 51), as a function of the sample size n (i.e. n from p_v and n' from q_v, with $n = n'$) and α. **Box-plots:** The distribution of node-level estimates, $\{\hat{PE}_v^{\alpha}(X_v \| X'_v)\}_{v \in V}$ (See Eq. 30), obtained for varying N and sample size $n = n'$. The horizontal dashed lines (red and green) indicate the true value of $PE(p_v^{\alpha} \| q_v)$ (See Eq. 32) within each cluster.

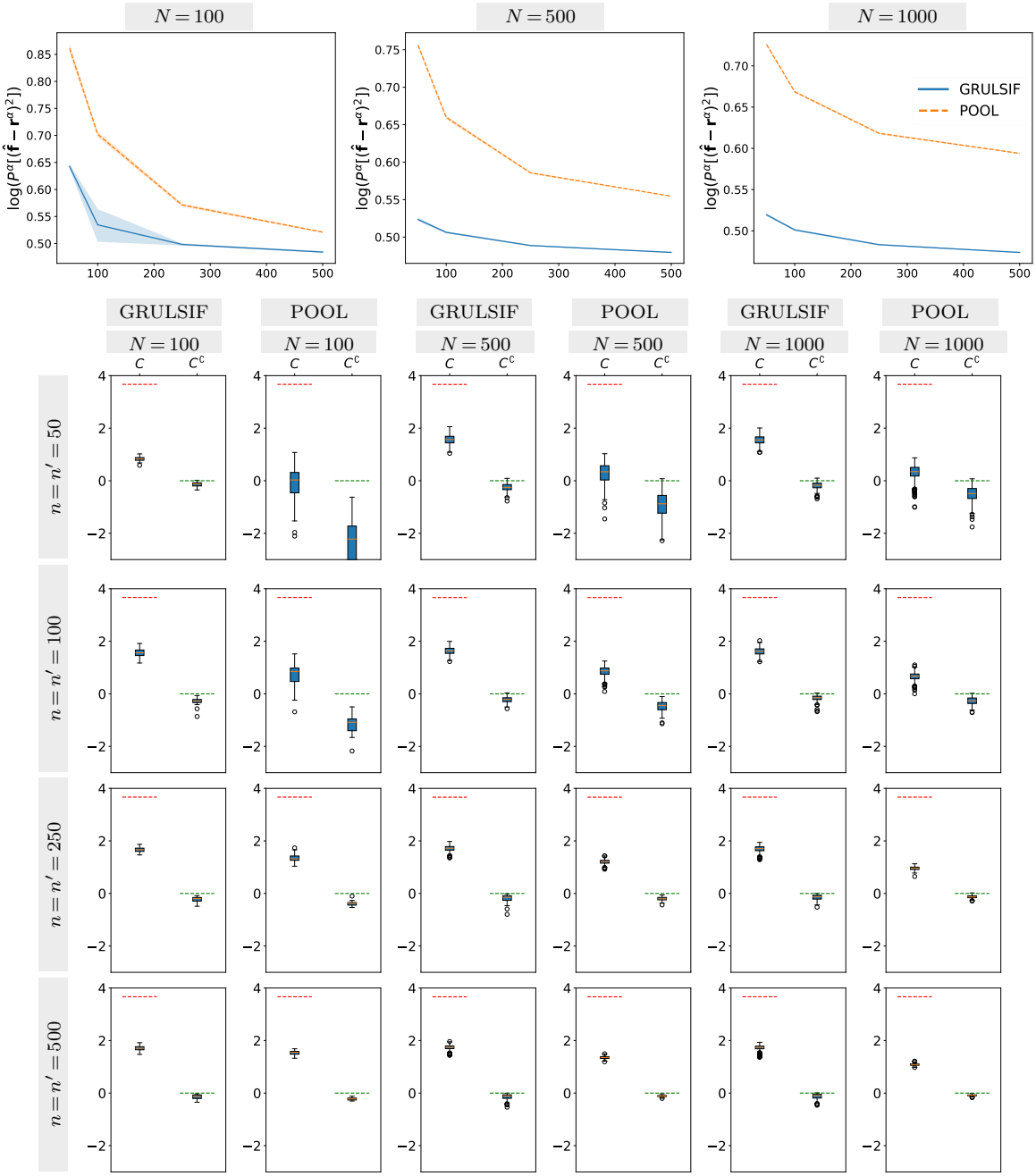

Figure 16: Experiment Synth.IIc for varying graph size. Complement of Fig. 6 that focuses on the behavior of GRULSIF and POOL for varying $N = \{100, 500, 1000\}$. The regularization parameter is fixed at $\alpha = 0.1$. **Line-plots:** Convergence to the true relative likelihood-ratio, in terms of the $\log\left(P^{\alpha}\left[[\mathbf{f} - \mathbf{r}^{\alpha}]^2\right]\right)$ (see Eq. 51), as a function of the sample size n (i.e. n from p_v and n' from q_v, with $n = n'$) and α. **Box-plots:** The distribution of node-level estimates, $\{\hat{PE}_v^{\alpha}(X_v \| X'_v)\}_{v \in V}$ (See Eq. 30), obtained for varying N and sample size $n = n'$. The horizontal dashed lines (red and green) indicate the true value of $PE(p_v^{\alpha} \| q_v)$ (See Eq. 32) within each cluster.

Table 3: GRULSIF's empirical time complexity*. Average running time in seconds in synthetic experiments with $n = 500$ observations at each node. The average is taken over 10 runs of GRULSIF, after the model selection whose computational time is not included.

#nodes (N)	**Synth.Ia**	**Synth.Ib**	**Synth.IIa**	**Synth.IIb**	**Synth.IIc**
100	0.09s	0.24s	0.31s	0.88s	29.19s
500	1.00s	1.25s	2.15s	1.70s	317.19s
1000	2.75s	6.05s	9.33s	5.54s	1395.95s

*On a single machine with 12th Gen Intel(R) Core(TM) i7-12700H processor and 16GB of RAM.

Finally, we compare the empirical time complexity of GRULSIF for different graph sizes. Tab. 3 reports the average running time over 10 GRULSIF instances, after the model selection step. The time complexity increases with N. For the experiments Synth.Ia-Synth.IIb, this increase does not seem to compromise the scalability of the algorithm. Nevertheless, the overhear for Synth.IIc is considerably bigger. The higher input space dimensionality of the Synth.IIc results in a larger dictionary of anchor points for approximating the associated RKHS. Specifically, the size of the dictionary scales at rate $\mathcal{O}(\hat{L}^3)$, which is evident even for graphs with as few as 100 nodes. In conclusion, the scalability of GRULSIF to large graphs depends more on the size of the dictionary used for approximation rather than on the number of nodes.

6.2 Real-life experiments

In this section, we compare GRULSIF to other Kernel-based LRE methods in examples of seismic data. Although this work focuses on the LRE performance, GRULSIF can be used for designing Change-Point Detection and Two-Sample Testing methods that account for graph-structured information. Such applications have been developed in parallel to this work in separate contributions in de la Concha et al. (2023, 2025).

Seismic data fall into the setting described in Sec. 2. Modern geological hazard monitoring systems consist of several stations located across a territory. Each station gathers data on ground noise and shaking at its specific location. It is expected that stations located close to each other will exhibit similar observations, which is a fundamental hypothesis in Spatial Statistics. We follow here a simple approach that considers the stations as nodes within a graph that encodes the similarity induced by their geographical proximity.

The main difficulty in evaluating the approximation of LRE methods using real data is that, in most cases, the likelihood-ratio is unknown. In this application, though, there is a specific case where a likelihood-ratio can be known in advance: the case when there is no seismic activity, i.e. both $p_v = q_v$ for all $v \in V$, implying $r_v^\alpha = 1$ for all $v \in V$, regardless of α. We leverage this observation to illustrate how incorporating geographical proximity within LRE improves the approximation accuracy.

Seismic data preprocessing. We start by identifying three seismic events that occurred in New Zealand. Seism A is of magnitude 5.5R (in Richter scale), occurred on May 31, 2021[1]; Seism B is of magnitude 2.6R, occurred on Oct 2, 2023 [2]; Seism C is of magnitude 2.3R,

1. Public data available by the GeoNet project GNS Science (1970): `https://www.geonet.org.nz/earthquake/2021p405872`
2. `https://www.geonet.org.nz/earthquake/2023p741652`

Figure 17: New Zealand seismographic network. The purple nodes represent the locations of seismic stations. The edges illustrate a 3-nearest neighbors graph, inferred based on stations' geographical coordinates to capture spatial proximity.

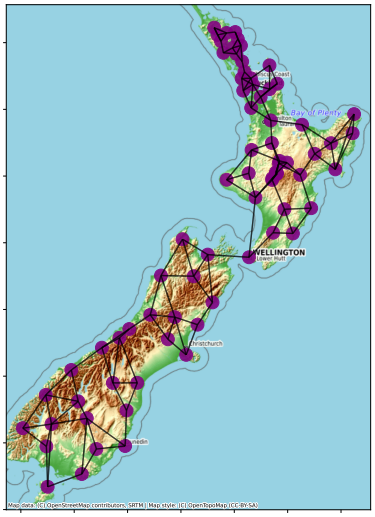

Table 4: Results on seismic data set. We report $P^{\alpha}\left[[\mathbf{f}-\mathbf{r}^{\alpha}]^2\right]$ (see Expr. 51) for each LRE method. To estimate this quantity, we preprocess the data corresponding to a time period where $r_v^{\alpha} \approx 1$ for all $v \in V$, fit the LRE method using 4/5 of the observations as training set, and the remaining 1/5 as test set. By design, a reported value closer to 0 indicates a better approximation.

LRE Method	Seism A	Seism B	Seism C
KLIEP	0.31	0.36	0.37
ULSIF	0.10	0.17	0.14
RULSIF ($\alpha = 0.1$)	0.10	0.16	0.13
POOL ($\alpha = 0.1$)	0.14	0.20	0.08
GRULSIF ($\alpha = 0.1$)	**0.08**	**0.09**	**0.02**

occurred on Oct 30, 2024 [3]. The data is made public available by the GeoNet project that operates a geological hazard monitoring system in that territory. The system is composed by seismic stations equipped with strong-motion accelerometers that provide 3d signals corresponding to the shaking across three perpendicular directions.

To emulate the situation where p_v and q_v are approximately the same, we analyze the waveforms recorded during one minute time, between 30 and 29 minutes before the seismic event at 100Hz frequency. The purpose of this choice is to ensure that observations within this time-window are stationary. The waveforms correspond to the measurements provided by the three perpendicular directions of the accelerometers, defining the input space $\mathcal{X} \subset \mathbb{R}^3$.

The preprocessing is performed independently for each station, and independently for each direction, using the Obspy Python package (Beyreuther et al., 2010). We apply a series of standard preprocessing steps used in seismology: the instrument response is deconvolved, the linear trend is removed, the observations are demeaned, a 2-20 bandpass filter is applied, and the filtered data are downsampled by a factor of 5. To reduce the temporal dependency, we fit an autoregressive model of order 1, and we keep the residuals to analyze further. The output is then standardized so that it has zero mean and unit variance. After completing these steps, we obtain 1200 observations at each location. We then assign the first 600 observations to p_v and the remaining 600 of them to q_v.

To account for spatial similarity, we generate an unweighted spatial graph $G_S = (V, E, W)$ where the nodes represent the seismic stations and the edges are computed in order to form a spatial 3-nearest neighbors graph, as visualized in Fig. 17.

3. `https://www.geonet.org.nz/earthquake/2024p817566`

Results and findings. Tab. 4 compares the Kernel-based LRE methods listed in Tab. 2 in terms of their average node-level MSE (Eq. 51). To compute this measure, we need both the approximated relative likelihood-ratio $\mathbf{f}$ and the true $\mathbf{r}^\alpha$. The former is given by each LRE estimator trained on 80% of the observations, while by design we have $\mathbf{r}^\alpha_v(x) = (1, ..., 1) \in \mathbb{R}^N$ for any α and $x \in \mathcal{X}$. This choice is consistent with the sampling from p_v and q_v, which is done such that $p_v \approx q_v$. The expected value $\mathbb{E}_{p^\alpha(y)}[[f_v - r^\alpha_v]^2(y)]$ is then calculated by averaging the estimation result on the remaining 20% of the observations that were not used during the training phase. In this configuration, we expect $\mathbb{E}_{p^\alpha(y)}[[f_v - r^\alpha_v]^2(y)]$ to be low and close to zero. In the results of Tab. 4, GRULSIF achieves the best performance (lowest MSE) for all the seismic events, which highlights the importance of the graph structure in enhancing the likelihood-ratio approximation.

7. Conclusions

In this paper, we first introduced a graph-based extension to the likelihood-ratio estimation (LRE) problem that we call Collaborative LRE. Then, we presented GRULSIF: a novel collaborative non-parametric LRE framework for multiple data sources whose similarity is encoded in a given fixed graph. We provided a detailed convergence analysis that highlights the conditions under which collaboration is beneficial compared to individual local LRE at each node, but also more specifically the role played by important variables of the problem, such as the complexity of the problem at hand, the amount of available data for each local problem, the expressiveness of the graph structure chosen for estimation, and the number of nodes. The provided distributed GRULSIF implementation scales well for big graphs. This is supported by the computational complexity analysis as well the reported empirical running time. In fact, the results show that the scaling of the algorithm is more affected by the size of the dictionary used for approximation than on the number of nodes.

As future work, there can be applications enabled by GRULSIF, some of which were outlined in this article. The presented theoretical guarantees could be used to study the behavior of task-oriented algorithms built on the top of GRULSIF estimates. Finally, an interesting question that deserves thorough investigation in the future, is how we can choose a graph that would render Collaborative LRE meaningful and more efficient.

Acknowledgments

The authors acknowledge support from the Industrial Data Analytics and Machine Learning Chair hosted at ENS Paris-Saclay, Université Paris-Saclay, and the Ile-de-France Region.

APPENDIX

Appendix A. Methodological aspects

A.1 Connection between Pearson's divergence and likelihood-ratio estimation

This section explains the relationship between χ^2-divergence and likelihood-ratio estimation (LRE), which justifies why we formulate the problem of comparing probabilistic models defined over the nodes of a graph as a LRE problem. In other words, we prove Theorem 2.

Proof of Theorem 2. The χ^2-divergence refers to the ϕ-divergence with $\phi(\zeta)=\frac{(\zeta-1)^2}{2}$, which is strictly convex around 1 and essentially smooth. Its convex conjugate is given by $\phi^*(s)=\frac{s^2}{2}+s$. Furthermore, from Inequality 2, we can conclude that $PE(P^\alpha\|Q)$ is bounded.

As $Q\in\mathcal{P}(\mathcal{X})$, then the total variation measure becomes $|Q|=Q$ and $Q\in\mathcal{M}_{\mathbb{H}}$. To verify the last point, take $f\in\mathbb{H}$, then:

$$\int \big|f(x')\big|\,dQ(x')=\int \big|\langle f,\mathrm{K}(x',\cdot)\rangle_{\mathbb{H}}\big|\,dQ(x')\leq C\,\|f\|_{\mathbb{H}}<\infty,$$

where the first equality is given by the representer property of $\mathbb{H}$, and the second one is a consequence of the Cauchy-Schwarz inequality.

The upper-bound on r^α implies that the function $\phi'(r^\alpha(x))=r^\alpha(x)-1$ belongs to $\mathrm{span}(\mathbb{H}\cup\mathcal{B})$, thus the requirements of Theorem 1 are satisfied and we obtain the expression:

$$PE(P^\alpha\|Q)=\sup_{g\in\mathrm{span}(\mathbb{H}\cup\mathcal{B})}\int g(x')dQ(x')-\int\left[\frac{g^2(y)}{2}+g(y)\right]dP^\alpha(y), \tag{52}$$

which admits the unique optimal solution $g^*=r^\alpha(x)-1$. Let us rewrite Eq. 52 in terms of functions of the form $g=f-1$:

$$\begin{aligned}PE(P^\alpha\|Q)&=\sup_{f\in\mathrm{span}(\mathbb{H}\cup\mathcal{B})}\int(f(x')-1)dQ(x')-\int\left[\frac{(f(y)-1)^2}{2}+(f(y)-1)\right]dP^\alpha(y)\\&=\sup_{f\in\mathrm{span}(\mathbb{H}\cup\mathcal{B})}\int f(x')dQ(x')-\int\frac{f^2(y)}{2}\,dP^\alpha(y)-\frac{1}{2}\\&\geq\sup_{f\in\mathbb{H}}\int f(x')dQ(x')-\int\frac{f^2(y)}{2}\,dP^\alpha(y)-\frac{1}{2}.\end{aligned} \tag{53}$$

■

A.2 Creating an efficient dictionary

In this section, we describe in detail how the implementation of GRULSIF creates the dictionary. Our strategy is an adaptation of the greedy algorithm introduced by Richard et al. (2009) that defines the *dictionary's coherence*, which measures the redundancy in a set of basis functions with regard to their lineal dependency and is computable in linear time to the dataset size. In practice, this approach selects a subset of observations (datapoints) forming a non-redundant subset of basis functions with good approximation performance. This strategy is applicable to a unit-norm kernel (i.e. $\mathrm{K}(x,x)=1,\ x\in\mathcal{X}$). Formally, the

Algorithm 3 – Dictionary creation

Input: $\{\mathbf{X}_v\}_{v\in V}=\{\{x_{v,1},...,x_{v,n_v}\}\}_{v\in V}$, $\{\mathbf{X}'_v\}_{v\in V}=\{\{x'_{v,1},...,x'_{v,n'_v}\}\}_{v\in V}$: the observations indexed by the node of the graph $G=(V,E,W)$ they belong;
$\mu_0\in(0,1)$: a coherence threshold;
σ : kernel hyperparameter (we assume a unit-norm kernel);
Output: D: a dictionary containing selected elements from $\{\mathbf{X}_v\}_{v\in V}$ and $\{\mathbf{X}'_v\}_{v\in V}$.

$D=\{\}$
for $v\in V$ **do**
  ▪ **Select non-redundant elements from** $\mathbf{X}'_v$
  for $i\in\{1,..,n_v\}$ **do**
    if $\max_{x\in D_v}|\mathrm{K}_{\sigma_v}(x_{v,i},x)|\leq\mu_0$ **then**
      $D=D\cup\{x_{v,i}\}$
    end if
  end for
  ▪ **Select non-redundant elements from** $\mathbf{X}'_v$
  for $i\in\{1,..,n'_v\}$ **do**
    if $\max_{x\in D_v}\left|\mathrm{K}_{\sigma_v}(x'_{v,i},x)\right|\leq\mu_0$ **then**
      $D=D\cup\{x'_{v,i}\}$
    end if
  end for
end for
return D

coherence of a dictionary $D_L=\{x_l\}_{l=1}^L$ of size L is defined as:

$$\mu=\max_{l\neq l'}|\langle\varphi(x_l),\varphi(x_{l'})\rangle|=\max_{l\neq l'}|\mathrm{K}(x_l,x_{l'})|\,. \tag{54}$$

This quantity can be read as the largest level of collinearity between pairs of elements of the dictionary. When the basis functions are orthogonal, this quantity equals to zero. The greedy recipe processes the observations one after the other and integrates a new observation x to the current dictionary D if its addition keeps the dictionary coherence bellow a given threshold $\mu_0\in(0,1)$, that is if: $\max_{x_l\in D_L}|\mathrm{K}(x,x_l)|\leq\mu_0$. The strategy is detailed in Alg. 3.

In all the reported experiments, we fix $\mu_0=0.3$. Larger values of μ_0 increased the running time of GRULSIF without offering performance gains.

A.3 Analysis of the proposed optimization algorithm

In this section, we detail the Cyclic Block Coordinate Gradient Descent (CBCGD) strategy described in Sec. 5. In particular, we prove an upper-bound for the number of interactions to attain a given precision ϵ in terms of the size of the dictionary L and the number of nodes N.

Theorem 4 is a particular case of the results appearing in Li et al. (2018). In that work, the convergence of Cyclic Block Coordinate-type algorithms is analyzed. For completeness of the presentation, we present some of their main results. The objective functions analyzed in Li et al. (2018) takes the form:

$$\min_{\mathbf{\Theta}\in\mathbb{R}^M}\Phi(\mathbf{\Theta})=\min_{\theta\in\mathbb{R}^M}Z(\mathbf{\Theta})+R(\mathbf{\Theta}), \tag{55}$$

where Z is a twice differentiable loss function, R is a possibly non-smooth and strongly convex penalty function, and the variable $\mathbf{\Theta}$ is of dimension $M=\sum_{v=1}^{N} M_v$ and is partitioned into disjoint blocks $\mathbf{\Theta}=(\theta_1,\theta_2,...,\theta_N)$ each of them being of dimension M_v. It is supposed that the penalization term can be written as $R(\mathbf{\Theta})=\sum_{v=1}^{N} R_v(\theta_v)$.

Assumption 5 *$Z(\cdot)$ is convex, and its gradient mapping $\nabla Z(\cdot)$ is Lipschitz-continuous and also block-wise Lipschitz-continuous, i.e. there exist positive constants C and C_v such that for any $\mathbf{\Theta},\mathbf{\Theta}'\in\mathbb{R}^M$ and $v=1,...,N$, we have:*

$$\begin{aligned}\left\|\nabla Z(\mathbf{\Theta}')-\nabla Z(\mathbf{\Theta})\right\| &\leq C\left\|\mathbf{\Theta}'-\mathbf{\Theta}\right\| \\ \left\|\nabla_v Z(\theta'_{u<v},\theta_v,\theta'_{u>v})-\nabla_v Z(\mathbf{\Theta}')\right\| &\leq C_v\left\|\theta_v-\theta'_v\right\|.\end{aligned} \tag{56}$$

Assumption 6 *$R(\cdot)$ is strongly convex and blockwise strongly convex, i.e. there exist positive constants μ and μ_v's such that for any $\mathbf{\Theta},\mathbf{\Theta}'\in\mathbb{R}^M$ and $v\in V$, we have for any $\xi\in\nabla R(\mathbf{\Theta}')$:*

$$\begin{aligned}R(\mathbf{\Theta}) &\geq R(\mathbf{\Theta}')+(\mathbf{\Theta}-\mathbf{\Theta}')^{\mathsf{T}}\xi+\frac{\mu}{2}\left\|\mathbf{\Theta}-\mathbf{\Theta}'\right\|^2, \\ R_v(\theta_v) &\geq R(\theta'_v)+(\theta_v-\theta'_v)^{\mathsf{T}}\xi_v+\frac{\mu_v}{2}\left\|\theta_v-\theta'_v\right\|^2.\end{aligned} \tag{57}$$

Under the above assumptions, the CBCGD cycle i for block v is defined as:

$$\hat{\theta}_v^{(i)}=\underset{\theta_v}{\operatorname{argmin}}(\theta_v-\hat{\theta}_v^{(i-1)})^{\mathsf{T}}\nabla_v Z(\hat{\theta}_{u<v}^{(i)},\hat{\theta}_{u\geq v}^{(i-1)})+\frac{\eta_v}{2}\left\|\theta_v-\hat{\theta}_v^{(i-1)}\right\|^2+R_v(\theta_v). \tag{58}$$

Then, Theorem A.1 characterizes the maximum number of interactions required to achieve a pre-specified accuracy ϵ.

Theorem 5 *(Theorem 3 in Li et al. (2018)) – Suppose that Assumptions 5 and 6 hold with $M\geq 2$. And that the optimization point is $\mathbf{\Theta}^*$. We choose $\alpha_v=C_v$ for the CBCGD method. Given a pre-specified accuracy $\epsilon>0$ of the objective value, we need at most*

$$i_{\max}=\left\lceil\frac{\mu C_{\min}^{\mu}+16C^2\log^2(3NM_{\max})}{\mu C_{\min}^{\mu}}\log\Big(\frac{\Phi(\mathbf{\Theta}^{(0)})-\Phi(\mathbf{\Theta}^*)}{\epsilon}\Big)\right\rceil$$

iterations to ensure $\Phi(\mathbf{\Theta}^{(i)})-\Phi(\mathbf{\Theta}^*)<\epsilon$ for $i\geq i_{\max}$, where $C_{\min}^{\mu}=\min_{v\in V}C_v+\mu_v$ and $M_{\max}=\max_{v\in V}M_v$.

Proof of Theorem 4. In this section, we assume that $\mathcal{K}$ is positive-definite,i.e. its minimum eigenvalue is strictly positive: $e_{\min}(\mathcal{K})>0$. This is in not the general case, but we can transform the problem for this condition to hold, e.g. using the approximation $\bar{\mathcal{K}}=\mathcal{K}+cI_L$, with $c>0$. Alternatively, $\mathcal{K}$ can be ensured to be positive-definite by selecting a dictionary with linear independent components (Richard et al., 2009), or via Nyström approximations along with anchor points selected via Kernel-PCA, as described in Sec. 5.2.

Problem 25 takes the form of Expr. 55, where we identify the functions Z and R as:

$$\begin{aligned}
Z(\boldsymbol{\Theta}) &= \min_{\boldsymbol{\Theta}\in\mathbb{R}^{NL}} \frac{1}{N}\left(\frac{(1-\alpha)}{2}\boldsymbol{\Theta}^{\mathsf{T}}\mathbf{H}\boldsymbol{\Theta} + \frac{\alpha}{2}\boldsymbol{\Theta}^{\mathsf{T}}\mathbf{H}'\boldsymbol{\Theta} - \mathbf{h}'^{\mathsf{T}}\boldsymbol{\Theta}\right) \\
&\quad + \frac{\lambda}{2}\boldsymbol{\Theta}^{\mathsf{T}}(I_N\otimes\mathcal{K}^{\frac{1}{2}})^{\mathsf{T}}\left[\mathcal{L}\otimes I_L\right](I_N\otimes\mathcal{K}^{\frac{1}{2}})\boldsymbol{\Theta} \\
&= \frac{1}{N}\sum_{v\in V}\left(\frac{(1-\alpha)}{2}\theta_v^{\mathsf{T}}H_v\theta_v + \frac{\alpha}{2}\theta_v^{\mathsf{T}}H'_v\theta_v - h_v'^{\mathsf{T}}\theta_v\right) + \frac{\lambda}{4}\sum_{u,v\in V}W_{uv}(\theta_v-\theta_u)^{\mathsf{T}}\mathcal{K}(\theta_v-\theta_u), \\
R(\boldsymbol{\Theta}) &= \frac{\lambda\gamma}{2}\sum_{v\in V}R_v(\theta_v) = \frac{\lambda\gamma}{2}\sum_{v\in V}\theta_v^{\mathsf{T}}\mathcal{K}\theta_v.
\end{aligned}$$

That given, it is easy to verify that the updating scheme of Eq. 58 takes the form of Eq. 40. It is clear that, given our hypothesis, $R(\boldsymbol{\Theta})$ and $R_v(\theta_v)$ are stronger convex functions of modulus $\lambda\gamma e_{\min}(K)$. Therefore, Assumption 6 is satisfied.

Second, the full gradient of $Z(\cdot)$ can be written as:

$$\nabla Z(\boldsymbol{\Theta}) = \left(\frac{1-\alpha}{N}\mathbf{H} + \frac{\alpha}{N}\mathbf{H}' + \lambda(I_N\otimes\mathcal{K}^{\frac{1}{2}})^{\mathsf{T}}\left[\mathcal{L}\otimes I_L\right](I_N\otimes\mathcal{K}^{\frac{1}{2}})\right)\boldsymbol{\Theta} - \frac{1}{N}\mathbf{h}', \tag{59}$$

which is Lipschitz-continuous with constant

$$C = e_{\max}\left(\frac{1-\alpha}{N}H + \frac{\alpha}{N}H' + \lambda(I_N\otimes\mathcal{K}^{\frac{1}{2}})^{\mathsf{T}}\left[\mathcal{L}\otimes I_L\right](I_N\otimes\mathcal{K}^{\frac{1}{2}})\right).$$

From the node-level expression, it is easy to derive the partial derivative of $Z(\cdot)$:

$$\nabla_v Z(\boldsymbol{\Theta}) = \frac{1-\alpha}{N}H_v + \frac{\alpha}{N}H'_v\theta_v + \lambda\mathcal{K}\left(d_v\theta_v - \sum_{u\in\mathrm{ng}(v)}W_{uv}\left(\theta_u\mathbf{1}_{u<v} + \theta_u\mathbf{1}_{u>v}\right)\right) - \frac{1}{N}h'_v, \tag{60}$$

where d_v is the degree of node v. This means:

$$\begin{aligned}
\left\|\nabla_v Z(\theta'_{u<v},\theta_v,\theta'_{u>v}) - \nabla_v Z(\boldsymbol{\Theta}')\right\| &\leq \left\|\left(\frac{1-\alpha}{N}H_v + \frac{\alpha}{N}H'_v + \lambda d_v\mathcal{K}\right)\left(\theta_v-\theta'_v\right)\right\| \\
&\leq C_v\left\|(\theta_v-\theta'_v)\right\|,
\end{aligned} \tag{61}$$

where $C_v = e_{\max}\left(\frac{1-\alpha}{N}H_{v,t} + \frac{\alpha}{N}H'_v + \lambda d_v K\right)$. Then, Assumption 5 is satisfied.

With these elements, and by fixing $\eta_v = C_v$, we can apply Theorem 5, where:

$$\begin{aligned}
C &= e_{\max}\left(\frac{1-\alpha}{N}\mathbf{H} + \frac{\alpha}{N}\mathbf{H}' + \lambda(I_N\otimes\mathcal{K}^{\frac{1}{2}})^{\mathsf{T}}\left[\mathcal{L}\otimes I_L\right](I_N\otimes\mathcal{K}^{\frac{1}{2}})\right), \\
C_v &= e_{\max}\left(\frac{(1-\alpha)}{N}H_v + \frac{\alpha}{N}H'_v + \lambda d_v\mathcal{K}\right), \\
C^{\mu}_{\min} &= \min_{v\in V}C_v + \mu, \\
\mu &= \lambda\gamma c.
\end{aligned}$$

After substitution, we get the expression given in Eq. 41. ■

Appendix B. GRULSIF in practice and details for the empirical evaluation

B.1 Further details for the conducted experiments

This section details the hyperparameter selection used in the experiments. For the RULSIF and ULSIF algorithms, we follow Sugiyama et al. (2011a) and Yamada et al. (2011). We run a leave-one-out cross-validation procedure over the parameter associated with the Gaussian kernel and the penalization term γ. The parameter σ is selected from the grid $\{0.6\sigma_{\text{median}}, 0.8\sigma_{\text{median}}, 1\sigma_{\text{median}}, 1.2\sigma_{\text{median}}, 1.4\sigma_{\text{median}}\}$, where σ_{median} is the parameter σ found via the median heuristic over the observations in x'_v. On the other hand, the penalization parameter γ is selected from the grid $\{1e^{-5}, 1e^{-3}, 0.1, 10\}$.

The procedure for KLIEP is similar, but we use instead a 5-fold cross-validation procedure, over the grid $\{0.6\sigma_{\text{median}}, 0.8\sigma_{\text{median}}, 1\sigma_{\text{median}}, 1.2\sigma_{\text{median}}, 1.4\sigma_{\text{median}}\}$ for the width σ of the Gaussian kernel, and over the grid $\{1e^{-5}, 1e^{-3}, 0.1, 10\}$ for the penalization constant.

For GRULSIF and POOL, we apply 5-fold cross-validation for selecting the hyperparameters σ, γ, λ (Alg. 1). Since POOL ignores the graph, we fix $\lambda = 1$, and the penalization term related with the norm of each functional f_v will only depend on the parameter γ. In order to select the width σ of the Gaussian kernel, we first compute $\{\sigma_v\}_{v\in V}$ for each node via the median heuristic applied to the observations of X_v (those are available when creating the dictionary), and we define $\sigma_{\min} = \min\{\sigma_v\}_{v\in V}$, $\sigma_{\text{median}} = \text{median}\{\sigma_v\}_{v\in V}$, and $\sigma_{\max} = \max\{\sigma_v\}_{v\in V}$, we then choose the final parameter from the set $\{\sigma_{\min}, \frac{1}{2}(\sigma_{\min} + \sigma_{\text{median}}), \sigma_{\text{median}}, \frac{1}{2}(\sigma_{\max} + \sigma_{\text{median}}), \sigma_{\max}\}$. γ is selected from the set $\{1e^{-5}, 1e^{-3}, 0.1, 10\}/c$, using $c = \sqrt{\min(n, n')}$ for POOL. For GRULSIF, we select the optimal λ^* from the set $\{1e^{-5}, 5.62e^{-4}, 3.16e^{-2}, 1.77, 1e^{2}\}/c$, and the optimal $\hat{\gamma}^*$ (representing the product of the constants $\gamma \cdot \lambda$) from the set $\{1e^{-5}, 1e^{-3}, 0.1, 10\}/c$, where $c = \sqrt{\min(n, n')N}$.

Appendix C. GRULSIF convergence guarantees

C.1 Auxiliary concepts and results from Multitask Learning

The excess risk bounds for Multitask Learning in Yousefi et al. (2018) depend on the concept of Multitask Local Rademacher Complexity (MTLRC) that aims at quantifying the complexity of classes of vector-valued functions. MTLRC leads to sharper bounds compared to the classical Global Rademacher Complexity, while remains easy to compute for a VV-RKHS that is of our interest. Specifically, the bounds are tight enough to explicit the role of important variables of the problem, such as the number of observations, the number of tasks, the smoothness of the vector-valued function to approximated in terms of the norm in the associated VV-RKHS. For completeness of presentation, and in order to clarify better our results, we adapt to the notation used in the main text and rewrite important concepts and results appearing in the reference papers Bartlett et al. (2005); Yousefi et al. (2018).

Let us denote by $\mathbf{Z} = (z_{v,i})_{v\in V, i=1,\dots,n}$ a set of nN independent observations such that for each $v \in \{1, \dots, N\}$, $\{z_{v,i} \sim p_{z,v}\}_{i=1}^{n}$ are identically distributed according to the measure $p_{z,v}$. Given a vector-valued function $\mathbf{h} = (h_1, \dots, h_v)$, we define the expressions:

$$P_z[\mathbf{h}] = \frac{1}{N}\sum_{v\in V} \mathbb{E}_{p_{z,v}}[h_v(z)] \quad \text{and} \quad P_{z,n}[\mathbf{h}] = \frac{1}{nN}\sum_{v=1}^{N}\sum_{i=1}^{n} h_v(z_{v,i}). \tag{62}$$

We start by introducing the concept of MTLRC (Yousefi et al., 2018) and the sub-root function (Bartlett et al., 2005) that will appear in the upper-bounds of the excess risk.

Definition 6 *(**Multitask Local Rademacher Complexity**) For a vector-value function class $\mathcal{F} = \{\mathbf{f} = (f_1, ..., f_N)\}$, the Multitask Local Rademacher Complexity (MTLRC) for $\rho > 0$, $\mathcal{R}(\mathcal{F}, \rho)$, is defined as:*

$$\mathcal{R}(\mathcal{F}, \rho) = \mathbb{E}_{z,\sigma}\left[\sup_{\substack{V(\mathbf{f}) \leq \rho \\ \mathbf{f}=(f_1,...,f_N)\in\mathcal{F}}} \frac{1}{nN}\sum_{v=1}^{N}\sum_{i=1}^{n}\sigma_{v,i} f_v(z_{v,i})\right], \tag{63}$$

where $\{\sigma_{v,i}\}_{v=1,...,N;i=1,...,n}$ is a sequence of independent Rademacher variables. We denote by $\mathbb{E}_{z,\sigma}[\cdot]$ the expectation w.r.t. all the involved random variables. $V(\mathbf{f})$ is an upper-bound on the variance of the function in $\mathcal{F}$.

Definition 7 *(**Sub-root function**) A function $\varrho : [0, \infty] \to [0, \infty]$ is sub-root iff it is non-decreasing and the function $\frac{\varrho(\rho)}{\sqrt{\rho}}$ is non-increasing for $\rho > 0$.*

Lemma 8 *(Lemma 3.2 in Bartlett et al. (2005)) If ϱ is a nontrivial sub-root function, then it is continuous in $[0, \infty]$ and the equation $\varrho(\rho) = \rho$ has a unique non-zero solution ρ^*, which is known as the fixed point of ϱ. Moreover, for any $\rho > 0$, it holds that $\rho \geq \varrho(\rho)$ iff $\rho^* \leq \rho$.*

The following result is established when the MTLRC is itself a sub-root function.

Lemma 9 *(Lemma 3.4 in Bartlett et al. (2005)) If the class $\mathcal{F}$ is star-shaped around $\mathbf{f}_0$, and $V : \mathcal{F} \to \mathbb{R}_+$ is a function that satisfies $V(a\mathbf{f}) \leq a^2 V(\mathbf{f})$ for any $\mathbf{f} \in \mathcal{F}$ and any $a \in [0, 1]$, then the function ϱ defined for $\rho \geq 0$ by:*

$$\varrho(\rho) = \mathbb{E}_{\sigma}\left[\sup_{\substack{V(\mathbf{f}-\mathbf{f}_0) \leq \rho \\ \mathbf{f}\in\mathcal{F}}} \frac{1}{nN}\sum_{v=1}^{N}\sum_{i=1}^{n}\sigma_{v,i} f_v(z_{v,i})\right] \tag{64}$$

is a sub-root function and $\rho \to \mathbb{E}_z[\varrho(\rho)]$ is also sub-root function.

A function class being star-shaped around $\mathbf{f}_0$, means that $\{\mathbf{f}_0 + a(\mathbf{f} - \mathbf{f}_0) : \mathbf{f} \in \mathcal{F}, a \in [0, 1]\} \subset \mathcal{F}$. Note that the a convex function class is by definition star-shaped around each of its elements.

MTLRC will be used to obtain a global error bound for classes of vector-valued functions for which the variance is bounded, $V(\mathbf{f} - \mathbf{f}_0) \leq \rho$. The goal is to identify models of small generalization error and small variance. There is a tradeoff between the size of the subset we consider (controlled by the parameter ρ) and its complexity, the optimal choice is given by a fixed point of a sub-root function. To formalize the relationship between the generalization error and variance, we need to define the concept of Vector-Valued Bernstein Class.

Definition 10 *(**Vector-Valued Bernstein Class**) Let $0 < \beta \leq 1$ and $B > 0$. A vector-valued function class $\mathcal{F}$ is said to be a (β, B)-Bernstein class w.r.t. the probability measure P if there exists a function $V : \mathcal{F} \to \mathbb{R}_+$ such that*

$$P\mathbf{f}^2 \leq V(\mathbf{f}) \leq B(P\mathbf{f})^{\beta}, \quad \forall \mathbf{f} \in \mathcal{F}. \tag{65}$$

The following result describes the role of MTLRC in obtaining upper-bounds for Multitask Learning and it is the core component in the proof of Theorem 3.

Theorem 11 *(Theorem B.3 in Yousefi et al. (2018)) Let $\mathcal{F}=\{\mathbf{f}=(f_1,...,f_N)\}$ be a class of vector-valued functions satisfying $\max_{v\in V}\sup_{z\in\mathcal{Z}}|f_v(z)|\leq b$. Let $\mathbf{Z}=\{\mathbf{Z}_v\}_{v\in V}=\{\{z_{v,1},...,z_{v,n}\}\}_{v\in V}$ be a vector of nN random variables where for each $v\in V$, $\{z_{v,1},...,z_{v,n}\}$ are identically distributed. Assume that $\mathcal{F}$ is (β,B)-Bernstein class of vector-valued functions with $0<\beta\leq 1$ and $B\geq 1$. Let ϱ be a sub-root function with fixed point ρ^*. If $B\mathcal{R}(\mathcal{F},r)\leq\varrho(\rho)$, $\forall\rho\geq\rho^*$, then for any $C>1$, and $\delta\in(0,1)$, with probability at least $1-\delta$, every $\mathbf{f}\in\mathcal{F}$ satisfies:*

$$\begin{aligned}P_z[\mathbf{f}] \;\leq\; & \frac{C}{C-\beta}P_{z,n}[\mathbf{f}]+(2C)^{\frac{\beta}{2-\beta}}20^{\frac{2}{2-\beta}}\max\left((\rho^*)^{\frac{1}{2-\beta}},(\rho^*)^{\frac{1}{\beta}}\right)\\ &+\left(\frac{2^{\beta+3}B^2C^\beta}{nN}\log\left(\frac{1}{\delta}\right)\right)^{\frac{1}{2-\beta}}+\frac{24Bb}{(2-\beta)nN}\log\left(\frac{1}{\delta}\right).\end{aligned}\tag{66}$$

C.2 Lemmata before Theorem 3

The general idea is to use Theorem 11 to upper-bound the excess risk associated with the Problem 33. To this end, we need to address the following subproblems:

1. Define a class of vector-valued functions satisfying the hypotheses of Theorem 11. (This is the context of Lemma 14.)
2. Identify the sub-root function ϱ that upper-bounds the MTLRC of the class. (Provided in the last point of Lemma 14.)
3. Upper-bound the fixed point of ϱ. (See Lemma 17.)

Let us start by defining the instantaneous loss function for the scalar function $f\in\mathbb{H}$:

$$\ell_v(f)(z_v)=\frac{(1-\alpha)f^2(x_v)+\alpha f^2(x'_v)}{2}-f(x'_v).\tag{67}$$

Here, the variable z_v denotes a pair of observations $z_v=(x_v,x'_v)$, where $x_v\sim p_v$ and $x'_v\sim q_v$.

Given a vector-valued function $\mathbf{f}=(f_1,f_2,...,f_N)\in\mathbb{G}$ and a set of pairs of observations $(z_1,z_2,...,z_N)$, we define the vector-valued loss function:

$$\ell(\mathbf{f})=(\ell_1(f_1)(z_1),\ell_2(f_2)(z_2),...,\ell_N(f_N)(z_N)).\tag{68}$$

To facilitate reading, we introduce the following operators evaluated at vector-valued functions of the form $\mathbf{h}=(h_1,...,h_N)$:

$$P[\mathbf{h}]=\frac{1}{N}\sum_{v\in V}\mathbb{E}_{p_v(x)}\left[h_v(x)\right],\quad Q[\mathbf{h}]=\frac{1}{N}\sum_{v\in V}\mathbb{E}_{q_v(x')}\left[h_v(x')\right],\quad P^\alpha[\mathbf{h}]=\frac{1}{N}\sum_{v\in V}\mathbb{E}_{p_v^\alpha(y)}[h_v(y)].\tag{69}$$

We can easily verify the following expressions:

$$P^\alpha[\mathbf{h}]=(1-\alpha)P[\mathbf{h}]+\alpha Q[\mathbf{h}]\quad\text{and}\quad P^\alpha[\mathbf{r}^\alpha\mathbf{h}]=Q[\mathbf{h}],\tag{70}$$

where, with an abuse of notation, $\mathbf{r}^\alpha\mathbf{h}$ refers to the point-wise multiplication of the vector-valued functions $\mathbf{r}^\alpha$ and $\mathbf{h}$. With this notation, we can define the cost function:

$$L(\mathbf{f}) = \sum_{v\in V} \mathbb{E}_{p_{z,v}}\left[\ell_v(f_v)(z)\right] = P_z[\ell(\mathbf{f})]. \tag{71}$$

The following lemma identifies the connection between the excess risk $L(\mathbf{f}) - L(\mathbf{r}^\alpha)$ and the distance $P^\alpha[\mathbf{f}-\mathbf{r}^\alpha]^2$. In particular, it makes evident the advantages of using the χ^2-divergence as a surrogate loss function for LRE.

Lemma 12 *Let the vector-valued functional space $\mathcal{F}_G$ (Expr. 33) and suppose the value of each scalar function f_v to range in $[-b,b]$, $b\in\mathbb{R}^+$. Then the following statements hold:*

1. *There is a function $\mathbf{f}^* = (f_1^*,...,f_N^*)\in\mathbb{G}$ satisfying:*

$$\begin{aligned}\mathbf{f}^* &= \operatorname*{argmin}_{\mathbf{f}\in\mathcal{F}_G} \frac{1}{N}\sum_{v\in V}\left[\frac{(1-\alpha)}{2}\mathbb{E}_{p_v(x)}[f_v^2(x)] + \frac{\alpha}{2}\mathbb{E}_{q_v(x')}[f_v^2(x')] - \mathbb{E}_{q_v(x')}[f_v(x')]\right] \\ &= \operatorname*{argmin}_{\mathbf{f}\in\mathcal{F}_G} L(f).\end{aligned}$$

2. *For every $\mathbf{f}\in\mathbb{G}$, we have $P^\alpha[\mathbf{f}-\mathbf{f}^*]^2 = 2(L(\mathbf{f}) - L(\mathbf{f}^*))$.*
3. *There exists $B_0 > 0$, such that $\forall\mathbf{f}\in\mathbf{G}$:*

$$P_z[\ell(\mathbf{f}) - \ell(\mathbf{f}^*)]^2 \le \frac{B_0}{2}P^\alpha[\mathbf{f}-\mathbf{f}^*]^2 = B_0 P_z[\ell(\mathbf{f}) - \ell(\mathbf{f}^*)].$$

Proof. ***First point*:** Assumption 3 states that $\mathbf{r}^\alpha\in\mathcal{F}_G$. Following the line of reasoning used to prove Expr. 13, we can conclude:

$$\operatorname*{argmin}_{\mathbf{f}\in\mathcal{F}_G}\frac{1}{N}\sum_{v\in V}\left[\frac{1}{2}\mathbb{E}_{p_v^\alpha(y)}[f_v^2(y)] - \mathbb{E}_{q_v(x')}[f_v(x')]\right] = \operatorname*{argmin}_{\mathbf{f}\in\mathcal{F}_G}\frac{1}{N}\sum_{v\in V}\frac{1}{2}\mathbb{E}_{p_v^\alpha(y)}[(f_v(y) - r_v^\alpha(y))^2],$$

which implies $\mathbf{r}^\alpha = (r_1^\alpha,...,r_N^\alpha)$ is solution to the optimization problem.

***Second point*:** The proof of the previous point implies $\mathbf{f}^* = \mathbf{r}^\alpha$. Then the second point of the lemma can be restated in terms of $L(\mathbf{f}) - L(\mathbf{r}^\alpha)$ for $\mathbf{f}\in\mathcal{F}_G$:

$$\begin{aligned}L(\mathbf{f}) - L(\mathbf{r}^\alpha) &= \frac{1}{N}\sum_{v\in V}\left[\frac{1}{2}\mathbb{E}_{p_v^\alpha(y)}[f_v^2(y) - (r_v^\alpha)^2(y)] - \mathbb{E}_{q_v(x')}[\big(f_v(x') - r_v^\alpha(x')\big)]\right] \\ &= \frac{1}{N}\sum_{v\in V}\mathbb{E}_{p_v^\alpha(y)}\left[\frac{1}{2}\left[f_v^2(y) - (r_v^\alpha)^2(y)\right] - r_v^\alpha(y)(f_v(y) - r_v^\alpha(y))\right] \\ &= \frac{1}{N}\sum_{v\in V}\frac{1}{2}\mathbb{E}_{p_v^\alpha(y)}[[f_v - r_v^\alpha]^2(y)] = \frac{1}{2}P^\alpha[\mathbf{f}-\mathbf{r}^\alpha]^2,\end{aligned}$$

where the second inequality comes from the expression $\mathbb{E}_{p_v^\alpha(y)}[r_v^\alpha(y)g(y)] = \mathbb{E}_{q_v(x')}[g(x')]$.

***Third point*:** Let us define the positive constants: $C_{\alpha,v} = \min\{\frac{1}{\alpha}, \|r_v^\alpha\|_\infty\}$, $C_\alpha = \max_{v\in V} C_{\alpha,v}$. Notice that by hypothesis over the functional space $\mathcal{F}_G$ and the upper-bound of r_v^α with respect to the regularization parameter, we have:

$$\|f_v + r_v^\alpha\|_\infty \le (b + C_{\alpha,v}) \le (b + C_\alpha), \tag{72}$$

$$
\begin{aligned}
P_z\left[\ell(\mathbf{f})-\ell(\mathbf{r}^\alpha)\right]^2 &= P_z\left[\frac{1-\alpha}{2}[\mathbf{f}^2-(\mathbf{r}^\alpha)^2](x)+\frac{\alpha}{2}[\mathbf{f}^2-(\mathbf{r}^\alpha)^2](x')-[\mathbf{f}-\mathbf{r}^\alpha](x')\right]^2 \\
&\leq 2P_z\left[\frac{1-\alpha}{2}[\mathbf{f}^2-(\mathbf{r}^\alpha)^2](x)+\frac{\alpha}{2}[\mathbf{f}^2-(\mathbf{r}^\alpha)^2](x')\right]^2 \\
&\quad +2Q\left[[\mathbf{f}-\mathbf{r}^\alpha]^2(x')\right] \qquad \left(\text{Inequality } (a+b)^2\leq 2a^2+2b^2\right) \\
&\leq 2P_z\left[\frac{(1-\alpha)}{4}\left(\left[\mathbf{f}^2-(\mathbf{r}^\alpha)^2\right](x)\right)^2+\frac{\alpha}{4}\left(\left[\mathbf{f}^2-(\mathbf{r}^\alpha)^2\right](x')\right)^2\right] \\
&\quad +2Q\left[[\mathbf{f}-\mathbf{r}^\alpha]^2(x')\right] \qquad \left(\text{Convexity of } x\to x^2\right) \\
&= \frac{1}{2}P^\alpha\left[\mathbf{f}^2-(\mathbf{r}^\alpha)^2\right]^2+2P^\alpha\left[\mathbf{r}^\alpha(\mathbf{f}-\mathbf{r}^\alpha)^2\right] \qquad (\text{Expr. } 70) \\
&\leq \frac{1}{2}P^\alpha\left[[\mathbf{f}-\mathbf{r}^\alpha]^2[\mathbf{f}+\mathbf{r}^\alpha]^2\right]+\frac{2}{\alpha}P^\alpha\left[[\mathbf{f}-\mathbf{r}^\alpha]^2\right] \\
&\leq \frac{1}{2}\left(\left(b+C_\alpha\right)^2+4C_\alpha\right)P^\alpha\left[[\mathbf{f}-\mathbf{r}^\alpha]^2\right] \qquad (\text{Expr. } 72) \\
&= \frac{1}{2}B_0P^\alpha\left[[\mathbf{f}-\mathbf{r}^\alpha]^2\right].
\end{aligned}
$$

Moreover, the second point implies:

$$
P_z\left[\ell(\mathbf{f})-\ell(\mathbf{r}^\alpha)\right]^2\leq B_0\left[L(\mathbf{f})-L(\mathbf{r}^\alpha)\right]=B_0P_z\left[\ell(\mathbf{f})-\ell(\mathbf{r}^\alpha)\right].
$$

■

Lemma 13 *Let* $\mathbf{Y}=\{\mathbf{Y}_v\}_{v\in V}=\left\{\{y_{v,1},...,y_{v,n}\}\right\}_{v\in V}$ *be a sample of* nN *observations such that* $\forall v,i\colon\ y_{v,i}\overset{i.i.d.}{\sim}p_v^\alpha$. *Then:*

$$
\mathbb{E}_\sigma\left[\sup_{\substack{P^\alpha[\mathbf{f}-\mathbf{r}^\alpha]^2\leq\rho\\ \mathbf{f}\in\mathcal{F}_G}}\frac{1}{nN}\sum_{v\in V}\sum_{i=1}^n\sigma_{v,i}f_v(y_{v,i})r^\alpha(y_{v,i})\right]\leq C_\alpha\mathbb{E}_\sigma\left[\sup_{\substack{P^\alpha[\mathbf{f}-\mathbf{r}^\alpha]^2\leq\rho\\ \mathbf{f}\in\mathcal{F}_G}}\frac{1}{nN}\sum_{v\in V}\sum_{i=1}^n\sigma_{v,i}f_v(y_{v,i})\right]. \tag{73}
$$

Proof Let us define $\mathbb{E}_{\sigma\setminus\sigma_{u,j}}[\cdot]$ to be the expectation with respect to all the Rademacher random variables $\{\sigma_{v,i}\}_{v=1,...,N;i=1,...,n}$ except $\sigma_{u,j}$, then:

$$
\begin{aligned}
&\mathbb{E}_\sigma\left[\sup_{\substack{P^\alpha[\mathbf{f}-\mathbf{r}^\alpha]^2\leq\rho\\ \mathbf{f}\in\mathcal{F}_G}}\frac{1}{nN}\sum_{v\in V}\sum_{i=1}^n\sigma_{v,i}f_v(y_{v,i})r_v^\alpha(y_{v,i})\right] \\
&=\mathbb{E}_{\sigma\setminus\sigma_{u,j}}\left\{\mathbb{E}_{\sigma_{u,j}}\left[\sup_{\substack{P^\alpha[\mathbf{f}-\mathbf{r}^\alpha]^2\leq\rho\\ \mathbf{f}\in\mathcal{F}_G}}\frac{1}{nN}\sum_{v\in V}\sum_{i=1}^n\sigma_{v,i}f_v(y_{v,i})r_v^\alpha(y_{v,i})\right]\right\} \\
&=\frac{1}{nN}\mathbb{E}_{\sigma\setminus\sigma_{u,j}}\left\{\mathbb{E}_{\sigma_{u,j}}\left[\sup_{\substack{P^\alpha[\mathbf{f}-\mathbf{r}^\alpha]^2\leq\rho\\ \mathbf{f}\in\mathcal{F}_G}}U_{V\setminus u;-j}(\mathbf{f})+\sigma_{u,j}f_u(y_{u,j})r_u^\alpha(y_{u,j})\right]\right\},
\end{aligned}
$$

where $U_{V\setminus u;-j}(\mathbf{f}) = \sum_{v\in V}\sum_{i=1}^{n}\sigma_{v,i}f_v(y_{v,i})r_v^{\alpha}(y_{v,i}) - \sigma_{u,j}f_u(y_{u,j})r_u^{\alpha}(y_{u,j})$. By the definition of the supremum, for any $\epsilon>0$, there exists $\mathbf{g},\mathbf{h}\in\mathcal{F}_G$ such that $P^{\alpha}\left[[\mathbf{g}-\mathbf{r}^{\alpha}]^2\right]\leq\rho$ and $P^{\alpha}\left[[\mathbf{h}-\mathbf{r}^{\alpha}]^2\right]\leq\rho$, such that:

$$U_{V\setminus u;-j}(\mathbf{g})+g_u(y_{u,j})r_u^{\alpha}(y_{u,j}) \geq (1-\epsilon)\left[\sup_{\substack{P^{\alpha}[\mathbf{f}-\mathbf{r}^{\alpha}]^2\leq\rho\\ \mathbf{f}\in\mathcal{F}_G}} U_{V\setminus u;-j}(\mathbf{f})+f_u(y_{u,j})r_u^{\alpha}(y_{u,j})\right]$$

$$U_{V\setminus u;-j}(\mathbf{h})-h_u(y_{u,j})r_u^{\alpha}(y_{u,j}) \geq (1-\epsilon)\left[\sup_{\substack{P^{\alpha}[\mathbf{f}-\mathbf{r}^{\alpha}]^2\leq\rho\\ \mathbf{f}\in\mathcal{F}_G}} U_{V\setminus u;-j}(\mathbf{f})-f_u(y_{u,j})r_u^{\alpha}(y_{u,j})\right].$$

This latter implies:

$$\begin{aligned}
&(1-\epsilon)\mathbb{E}_{\sigma_{u,j}}\left[\sup_{\substack{P^{\alpha}[\mathbf{f}-\mathbf{r}^{\alpha}]^2\leq\rho\\ \mathbf{f}\in\mathcal{F}_G}} U_{V\setminus u;-j}(\mathbf{f})+\sigma_{u,j}f_u(y_{u,j})r_u^{\alpha}(y_{u,j})\right]\\
&=\frac{(1-\epsilon)}{2}\left[\sup_{\substack{P^{\alpha}[\mathbf{f}-\mathbf{r}^{\alpha}]^2\leq\rho\\ \mathbf{f}\in\mathcal{F}_G}}\left[U_{V\setminus u;-j}(\mathbf{f})+f_u(y_{u,j})r_u^{\alpha}(y_{u,j})\right]+\sup_{\substack{P^{\alpha}[\mathbf{f}-\mathbf{r}^{\alpha}]^2\leq\rho\\ \mathbf{f}\in\mathcal{F}_G}}\left[U_{V\setminus u;-j}(\mathbf{f})-f_u(y_{u,j})r_u^{\alpha}(y_{u,j})\right]\right]\\
&\leq\frac{1}{2}\left[U_{V\setminus u;-j}(\mathbf{g})+g_u(y_{u,j})r_u^{\alpha}(y_{u,v})\right]+\frac{1}{2}\left[U_{V\setminus u;-j}(\mathbf{h})-h_u(y_{u,j})r_u^{\alpha}(y_{u,v})\right]\\
&\leq\frac{1}{2}\left[U_{V\setminus u;-j}(\mathbf{g})+U_{V\setminus u;-j}(\mathbf{h})+sr_u^{\alpha}(y_{u,j})\left(g_u(y_{u,j})-h_u(y_{u,j})\right)\right],
\end{aligned}$$

where $s=\operatorname{sgn}(g_u(y_{u,j})-h_u(y_{u,j}))$. Then, the upper-bound on r_u^{α} implies:

$$\begin{aligned}
&(1-\epsilon)\mathbb{E}_{\sigma_{u,j}}\left[\sup_{\substack{P^{\alpha}[\mathbf{f}-\mathbf{r}^{\alpha}]^2\leq\rho\\ \mathbf{f}\in\mathcal{F}_G}} U_{V\setminus u;-j}(\mathbf{f})+\sigma_{u,j}f_{u,j}(y_{u,j})r_u^{\alpha}(y_{u,j})\right]\\
&\leq\frac{1}{2}\left[U_{V\setminus u;-j}(\mathbf{g})+U_{V\setminus u;-j}(\mathbf{h})+C_{\alpha}s\left(g_u(y_{u,j})-h_u(y_{u,j})\right)\right]\\
&=\frac{1}{2}\left[U_{V\setminus u;-j}(\mathbf{g})+C_{\alpha}sg_u(y_{u,j})\right]+\frac{1}{2}\left[U_{V\setminus u;-j}(\mathbf{h})-C_{\alpha}sh_u(y_{u,j})\right]\\
&\leq\frac{1}{2}\sup_{\substack{P^{\alpha}[\mathbf{f}-\mathbf{r}^{\alpha}]^2\leq\rho\\ \mathbf{f}\in\mathcal{F}_G}}\left[U_{V\setminus u;-j}(\mathbf{f})+C_{\alpha}sf_u(y_{u,j})\right]+\frac{1}{2}\sup_{\substack{P^{\alpha}[\mathbf{f}-\mathbf{r}^{\alpha}]^2\leq\rho\\ \mathbf{f}\in\mathcal{F}_G}}\left[U_{V\setminus u;-j}(\mathbf{f})-C_{\alpha}sf_u(y_{u,j})\right]\\
&=\mathbb{E}_{\sigma_{u,j}}\left[\sup_{\substack{P^{\alpha}[\mathbf{f}-\mathbf{r}^{\alpha}]^2\leq\rho\\ \mathbf{f}\in\mathcal{F}_G}} U_{V\setminus u;-j}(\mathbf{f})+C_{\alpha}\sigma_{u,j}f_{u,j}(y_{u,j})\right],
\end{aligned}$$

where in the last equality, we have used the definition of $\sigma_{u,j}$. As the inequality is satisfied for all $\epsilon>0$, we have:

$$\mathbb{E}_{\sigma_{u,j}}\left[\sup_{\substack{P^{\alpha}[\mathbf{f}-\mathbf{r}^{\alpha}]^2\leq\rho\\ \mathbf{f}\in\mathcal{F}_G}} U_{V\setminus u;-j}(\mathbf{f})+\sigma_{u,j}f_{u,j}(y_{u,j})r_u^{\alpha}(y_{u,j})\right]\leq\mathbb{E}_{\sigma_{u,j}}\left[\sup_{\substack{P^{\alpha}[\mathbf{f}-\mathbf{r}^{\alpha}]^2\leq\rho\\ \mathbf{f}\in\mathcal{F}_G}} U_{V\setminus u;-j}(\mathbf{f})+C_{\alpha}\sigma_{u,j}f_{u,j}(y_{u,j})\right]. \tag{74}$$

Using the same argument for the rest $\sigma_{v,i}$ terms for $v \neq u$, $i \neq j$, we get the final result. ■

Lemma 14 identifies the vector-valued function class satisfying the hypotheses of Theorem 11.

Lemma 14 *Let us define the class of functions:*

$$\mathcal{H}_{\mathbb{G}} = \{h_{\mathbf{f}} = (h_{f_1}, ..., h_{f_N}), h_{f_v} : (x_v, x'_v) \to \ell_v(f_v)(x_v, x'_v) - \ell_v(r_v^{\alpha})(x_v, x'_v), \mathbf{f} \in \mathcal{F}_G\}, \tag{75}$$

which satisfies the following points:

1. $\max_{v \in V} \sup_{(x,x') \in \mathcal{X}} |h_{f_v}(x, x')| \leq B_1$.
2. $\mathcal{H}_{\mathbb{G}}$ *is a* (β, B)*-Bernstein class with* $\beta = 1$ *and* $B = B_0 = \frac{1}{2}\left((b + C_{\alpha})^2 + 4C_{\alpha}\right)$.
3. $\mathcal{H}_{\mathbb{G}}$ *satisfies the following inequality:*

$$\mathcal{R}(\mathcal{H}_{\mathbb{G}}, \rho) \leq 2\,(b + C_{\alpha})\,\mathcal{R}\left(\mathcal{F}_G, \frac{\rho}{2B_0}\right), \tag{76}$$

where the two involved MTLRCs are defined by:

$$\begin{aligned} \mathcal{R}(\mathcal{H}_{\mathbb{G}}, \rho) &= \mathbb{E}_{z,\sigma}\left[\sup_{V(h_{\mathbf{f}}) \leq \rho, \mathbf{f} \in \mathcal{F}_G} \frac{1}{nN} \sum_{v \in V} \sum_{i=1}^{n} \sigma_{v,i} h_{f_v}(x_{v,i}, x'_{v,i})\right], \\ \mathcal{R}(\mathcal{F}_G, \rho) &= \mathbb{E}_{p^{\alpha},\sigma}\left[\sup_{\substack{P^{\alpha}\mathbf{f}^2 \leq \rho \\ \mathbf{f} \in \mathcal{F}_G}} \frac{1}{nN} \sum_{v \in V} \sum_{i=1}^{n} \sigma_{v,i} f_v(y_{v,i})\right]. \end{aligned} \tag{77}$$

Proof. ***First point*:**

$$\begin{aligned} &\max_{v \in V} \sup_{(x,x') \in \mathcal{X}} \left|h_{f_v}(x, x')\right| = \max_{v \in V} \sup_{(x,x') \in \mathcal{X} \times \mathcal{X}} \left|\ell_v(f_v)(x, x') - \ell_v(r_v^{\alpha})(x, x')\right| \\ &\leq \max_{v \in V} \sup_{(x,x') \in \mathcal{X} \times \mathcal{X}} \frac{(1-\alpha)}{2} \left|\left[f_v^2 - (r_v^{\alpha})^2\right](x)\right| + \frac{\alpha}{2} \left|\left[f_v^2 - (r_v^{\alpha})^2\right](x')\right| + \left|\left[f_v - r_v^{\alpha}\right](x')\right| \\ &= \frac{1}{2}(b + C_{\alpha})^2 + (b + C_{\alpha}) =: B_1. \qquad \text{(Expr. 72)} \end{aligned}$$

***Second point*:** Due to the properties listed in Lemma 12, we have the following inequalities:

$$P_z\left[h_{\mathbf{f}}\right]^2 = P_z\left[\ell(\mathbf{f}) - \ell(\mathbf{r}^{\alpha})\right]^2 \leq \frac{B_0}{2} P^{\alpha}\left[\mathbf{f} - \mathbf{r}^{\alpha}\right]^2 = B_0 P_z\left[\ell(\mathbf{f}) - \ell(\mathbf{r}^{\alpha})\right], \tag{78}$$

which means $\mathcal{H}_{\mathbb{G}}$ is a (β, B)-Bernstein class of vector-value functions, with $\beta = 1$ and $B = B_0$, and the function controlling the variance of the class is defined as: $V(h_{\mathbf{f}}) = \frac{B_0}{2} P^{\alpha}(\mathbf{f} - \mathbf{r}^{\alpha})^2$.

***Third point*:** By fixing $\rho \in \mathbb{R}_+$, we can verify:

$$
\begin{aligned}
B_0\mathcal{R}(\mathcal{H}_{\mathbb{G}},\rho) &= B_0\mathbb{E}_{z,\sigma}\left[\sup_{V(h_{\mathbf{f}})\leq\rho,\mathbf{f}\in\mathcal{F}_G}\frac{1}{nN}\sum_{v\in V}\sum_{i=1}^{n}\sigma_{v,i}\ell_v(f_v)(x_{v,i},x'_{v,i})\right] \qquad \text{(Eq. 63)}\\
&= B_0\mathbb{E}_{z,\sigma}\left[\sup_{V(h_{\mathbf{f}})\leq\rho,\mathbf{f}\in\mathcal{F}_G}\frac{1}{nN}\sum_{v\in V}\sum_{i=1}^{n}\sigma_{v,i}\left(\frac{(1-\alpha)}{2}f_v^2(x_{v,i})+\frac{\alpha}{2}f_v^2(x'_{v,i})-f_v(x'_{v,i})\right)\right]\\
&\leq B_0\mathbb{E}_{z,\sigma}\left[\sup_{V(h_{\mathbf{f}})\leq\rho,\mathbf{f}\in\mathcal{F}_G}\frac{1}{nN}\sum_{v\in V}\sum_{i=1}^{n}\sigma_{v,i}\left(\frac{(1-\alpha)}{2}f_v^2(x_{v,i})+\frac{\alpha}{2}f_v^2(x'_{v,i})\right)\right]\\
&+ B_0\mathbb{E}_{z,\sigma}\left[\sup_{V(h_{\mathbf{f}})\leq r,\mathbf{f}\in\mathcal{F}_G}\frac{1}{nN}\sum_{v\in V}\sum_{i=1}^{n}\sigma_{v,i}f_v(x'_{v,i})\right] \qquad \left(\begin{smallmatrix}\text{Subadditivity of the supremum and}\\ \text{symmetry of the Redemacher variables}\end{smallmatrix}\right)\\
&= B_0\mathbb{E}_{p^\alpha,\sigma}\left[\sup_{V(h_{\mathbf{f}})\leq\rho,\mathbf{f}\in\mathcal{F}_G}\frac{1}{nN}\sum_{v\in V}\sum_{i=1}^{n}\frac{1}{2}\sigma_{v,i}f_v^2(y_{v,i})\right]\\
&+ B_0\mathbb{E}_{p^\alpha,\sigma}\left[\sup_{V(h_{\mathbf{f}})\leq r,\mathbf{f}\in\mathcal{F}_G}\frac{1}{nN}\sum_{v\in V}\sum_{i=1}^{n}\sigma_{v,i}f_v(y_{v,i})r_v^\alpha(y_{v,i})\right],
\end{aligned}
$$

where the last expression is a consequence of $\mathbb{E}_{p_v^\alpha(y)}[h(y)] = (1-\alpha)\mathbb{E}_{p_v(x)}[h(x)]+\alpha\mathbb{E}_{q_v(x')}[h(x')]$ and $\mathbb{E}_{p_v^\alpha(y)}[f(y)r^\alpha(y)] = \mathbb{E}_{q_v(x')}[g(x')]$. Notice, x^2 is a Lipschitz function with Lipschitz constant $2b$ when $x \in [-b,b]$. We can apply the Contraction property of Rademacher Complexity, which holds also for the Local Rademacher Complexity for vector-valued function classes (Theorem 17 in Maurer (2006a)). The latter result leads to the inequality:

$$
\mathbb{E}_{p^\alpha,\sigma}\left[\sup_{V(h_{\mathbf{f}})\leq\rho,\mathbf{f}\in\mathcal{F}_G}\frac{1}{nN}\sum_{v\in V}\sum_{i=1}^{n}\sigma_{v,i}f_v^2(y_{v,i})\right] \leq 2b\mathbb{E}_{p^\alpha,\sigma}\left[\sup_{V(h_{\mathbf{f}})\leq\rho,\mathbf{f}\in\mathcal{F}_G}\frac{1}{nN}\sum_{v\in V}\sum_{i=1}^{n}\sigma_{v,i}f_v(y_{v,i})\right].
$$

By combining the above inequality and Lemma 13 we obtain:

$$
\begin{aligned}
B_0\mathcal{R}(\mathcal{H}_{\mathbb{G}},\rho) &\leq B_0\,(b+C_\alpha)\,\mathbb{E}_{p^\alpha,\sigma}\left[\sup_{V(h_{\mathbf{f}})\leq\rho,\mathbf{f}\in\mathcal{F}_G}\frac{1}{nN}\sum_{v\in V}\sum_{i=1}^{n}\sigma_{v,i}f_v(y_{v,i})\right]\\
&= B_0\,(b+C_\alpha)\,\mathbb{E}_{p^\alpha,\sigma}\left[\sup_{\substack{\frac{B_0}{2}P^\alpha[\mathbf{f}-\mathbf{r}^\alpha]^2\leq\rho\\ \mathbf{f}\in\mathcal{F}_G}}\frac{1}{nN}\sum_{v\in V}\sum_{i=1}^{n}\sigma_{v,i}f_v(y_{v,i})\right]\\
&= B_0\,(b+C_\alpha)\,\mathbb{E}_{p^\alpha,\sigma}\left[\sup_{\substack{\frac{B_0}{2}P^\alpha[\mathbf{f}-\mathbf{r}^\alpha]^2\leq\rho\\ \mathbf{f}\in\mathcal{F}_G}}\frac{1}{nN}\sum_{v\in V}\sum_{i=1}^{n}\sigma_{v,i}[f_v(y_{v,i})-r_v^\alpha(y_{v,i})]\right] \left(\text{By the independence of } \sigma,\ \mathbf{Y}\right)\\
&\leq B_0\,(b+C_\alpha)\,\mathbb{E}_{p^\alpha,\sigma}\left[\sup_{\substack{\frac{B_0}{2}P^\alpha[\mathbf{f}-\mathbf{g}]^2\leq\rho\\ \mathbf{f},\mathbf{g}\in\mathcal{F}_G}}\frac{1}{nN}\sum_{v\in V}\sum_{i=1}^{n}\sigma_{v,i}[f_v(y_{v,i})-g_v(y_{v,i})]\right]\\
&= B_0\,(b+C_\alpha)\,\mathbb{E}_{p^\alpha,\sigma}\left[\sup_{\substack{2B_0P^\alpha\mathbf{f}^2\leq\rho\\ \mathbf{f}\in\mathcal{F}_G}}\frac{1}{nN}\sum_{v\in V}\sum_{i=1}^{n}\sigma_{v,i}f_v(y_{v,i})\right] = 2B_0\,(b+C_\alpha)\,\mathcal{R}\Big(\mathcal{F}_G,\frac{\rho}{2B_0}\Big).
\end{aligned}
\tag{79}
$$

In the last inequality we have used the symmetry of the Rademacher variables and the fact that $f \in \mathcal{F}_G$ is symmetric and convex. ■

Lemma 9 implies that $2B_0(b+C_\alpha)\mathcal{R}(\mathcal{F}_G, \frac{\rho}{2B_0})$ is a sub-root function. The goal now is to upper-bound its fixed point ρ^*. This requires exploiting the properties of the graph regularization and the capacity condition associated with the covariance operators $\{\Sigma_v\}_{v\in V}$. Big part of this analysis has been done in Yousefi et al. (2018). We rewrite their most relevant results, and rework the upper-bounds to obtain clearer expressions for our problem.

Theorem 15 *(Theorem 11 in Yousefi et al. (2018)) Let the regularizer be $\|\mathbf{f}\|^2_{\mathbb{G}}$ as defined in Eq. 16, and denote its dual norm by $\|\cdot\|_*$. Let the kernels be uniformly bounded, and define the sample $\mathbf{Y} = \{\mathbf{Y}_v\}_{v\in V} = \big\{\{y_{v,1},...,y_{v,n}\}\big\}_{v\in V}$, and where $\forall v \in V$, $\{y_{v,1},...,y_{v,n}\}$ is a i.i.d. sample drawn from p_v^α. Assume that for each $v \in V$, the associated covariance operator admits an eigenvector decomposition $\Sigma_v = \mathbb{E}_{p_v^\alpha(y)}[\varphi(y)\otimes\varphi(y)] = \sum_{i\in\mathbb{N}} \mu_{v,i}\tilde{\varphi}_{v,i}\otimes\tilde{\varphi}_{v,i}$, where $\{\tilde{\varphi}_{v,i}\}_{v\in V}$ forms an orthonormal basis of $\mathbb{H}$ and $\{\mu_{v,i}\}_{i=1}^\infty$ are the corresponding eigenvalues in non-increasing order. Then, for any given positive operator $\mathcal{D}$ on $\mathbb{R}^N$, any $\rho > 0$ and any non-negative integers $h_1,...,h_N$:*

$$R(\mathcal{F}_G,\rho) \leq \sqrt{\frac{\rho\sum_{v\in V} h_v}{nN}} + \frac{\sqrt{2}\Lambda}{N}\mathbb{E}_{p^\alpha,\sigma}\left[\|\mathcal{D}^{-\frac{1}{2}}\mathbf{V}\|_*\right], \tag{80}$$

where $\mathbf{V} = \Big\{\sum_{j>h_v}\langle\frac{1}{n}\sum_{i=1}^n \sigma_{v,i}\phi(y_{v,i}), \tilde{\varphi}_{v,j}\rangle_{\mathbb{H}}\tilde{\varphi}_{v,j}\Big\}_{v\in V}$.

Lemma 16 *(Expr. C.9 of Corollary 22 in Yousefi et al. (2018)) Under the hypotheses of the previous theorem, we have the following inequality:*

$$\mathbb{E}_{p^\alpha,\sigma}\left[\|\mathcal{D}^{-\frac{1}{2}}\mathbf{V}\|_*\right] \leq \sqrt{\frac{1}{n}\sum_{v\in V}\Big|\mathcal{D}_{vv}^{-1}\sum_{j>h_v}\mu_{v,j}\Big|}, \tag{81}$$

where $\{\mathcal{D}_{vv}^{-1}\}_{v\in V}$ are the diagonal elements of $\mathcal{D}^{-1}$.

Lemma 17 *If Assumptions 2-3 are satisfied and $\mathcal{F}_G$ is a class of functions ranging in $[-b,b]$, $b\in\mathbb{R}^+$, then $2B_0(b+C_\alpha)\mathcal{R}(\mathcal{F}_G,\frac{\rho}{2B_0})$ is a sub-root function fixed point ρ^* satisfying the upper-bound:*

$$\rho^* \leq 8B_0\sqrt{\frac{\zeta^*+1}{\zeta^*-1}}\left(\mathcal{T}_G\Lambda^2(b+C_a)^{2\zeta^*}\right)^{\frac{1}{1+\zeta^*}} n^{-\frac{\zeta^*}{1+\zeta^*}} N^{-\frac{1}{1+\zeta^*}} s_{\max}^{\frac{1}{1+\zeta^*}}, \tag{82}$$

where

$$\mathcal{T}_G = \beta\gamma^{-1} + (1-\beta)\lambda_{\min^+}^{-1} \tag{83}$$

encodes the graph topology, $\zeta^* = \min_{v\in V}\zeta_v$, $s_{\max} = \max_{v\in V} s_v$, $\beta = \frac{\#C(G)}{N} \in [0,1]$ is the ratio between the number of connected components and the number of nodes N, $\lambda_{\min^+}$ denotes the lowest nonzero eigenvalue of $\mathcal{L}$, and by convention, we set $(1-\beta)\lambda_{\min^+}^{-1} = 0$ when $\#C(G) = N$.

Proof Combining Theorem 15 and Lemma 16 leads us to the following inequality:

$$\begin{aligned}
2B_0(b+C_\alpha)\mathcal{R}(\mathcal{F}_G,\frac{\rho}{2B_0}) &\leq 2B_0(b+C_\alpha)\left[\sqrt{\frac{\rho\sum_{v\in V}h_v}{2B_0nN}}+\sqrt{\frac{2\Lambda^2}{N^2n}\sum_{v\in V}\left|\mathcal{D}_{vv}^{-1}\sum_{j>h_v}\mu_{v,j}\right|}\right] \\
&= \sqrt{2B_0\left(b+C_\alpha\right)^2\frac{\rho\sum_{v\in V}h_v}{nN}}+2B_0\left(b+C_\alpha\right)\sqrt{\frac{2\Lambda^2}{N^2n}\sum_{v\in V}\left|\mathcal{D}_{vv}^{-1}\sum_{j>h_v}\mu_{v,j}\right|} \\
&= \sqrt{a_1\rho}+a_2,
\end{aligned} \tag{84}$$

where in the last expression we have introduced the variables:

$$a_1 = 2B_0\left(b+C_\alpha\right)^2\left(\frac{\sum_{v\in V}h_v}{nN}\right), \quad a_2 = 2B_0(b+C_\alpha)\sqrt{\frac{2\Lambda^2}{N^2n}\sum_{v\in V}\left|\mathcal{D}_{vv}^{-1}\sum_{j>h_v}\mu_{v,j}\right|}. \tag{85}$$

Now, we will look for the solution to the equation $\sqrt{a_1\rho}+a_2=\rho$, which is equivalent to solve $\rho^2-(a_1+2a_2)\rho+a_2^2=0$, that is

$$\rho = \frac{(a_1+2a_2)\pm\sqrt{a_1^2+4a_2}}{2} \leq a_1+2a_2. \tag{86}$$

As ρ^* is the fixed point of $2B_0(b+C_\alpha)\mathcal{R}(\mathcal{F}_G,\frac{\rho}{2B_0})$, then by Lemma 8, we have:

$$\rho^* \leq \rho \leq a_1+2a_2. \tag{87}$$

The goal now is to upper-bound both the terms a_1 and a_2 by exploiting Assumption 4. Observe that by the capacity condition (Assumption 4), we have:

$$\sum_{j>h_v}\mu_{v,j} \leq \sum_{j>h_v}s_v^2 j^{-\zeta_v} \leq s_v\int_{h_v}^{\infty}x^{-\zeta_v}dx = -\frac{s_v}{1-\zeta_v}h_v^{1-\zeta_v},$$

which implies: $a_2 \leq 2B_0\left(b+C_\alpha\right)\sqrt{\frac{-2\Lambda^2}{N^2n}\sum_{v\in V}\left|\mathcal{D}_{vv}^{-1}\right|\frac{s_v}{1-\zeta_v}h_v^{1-\zeta_v}}$.

Moreover, by the Cauchy–Schwarz inequality we have:

$$a_1 \leq 2B_0\left(b+C_\alpha\right)^2\sqrt{N}\sqrt{\frac{\sum_{v\in V}h_v^2}{(nN)^2}} = 2B_0(b+C_\alpha)^2\sqrt{\frac{\sum_{v\in V}h_v^2}{n^2N}}.$$

After putting together both inequalities, we get:

$$\begin{aligned}
\rho^* &\leq a_1+2a_2 \\
&\leq 2B_0\left(b+C_\alpha\right)\left[\sqrt{(b+C_\alpha)^2\left(\frac{\sum_{v\in V}h_v^2}{n^2N}\right)}+\sqrt{\frac{-8\Lambda^2}{N^2n}\sum_{v\in V}\left|\mathcal{D}_{vv}^{-1}\right|\frac{s_v}{1-\zeta_v}h_v^{1-\zeta_v}}\right] \\
&\leq 2B_0\left(b+C_\alpha\right)\sqrt{\sum_{v\in V}2(b+C_\alpha)^2\left(\frac{h_v^2}{n^2N}\right)-\frac{16\Lambda^2}{N^2n}\left|\mathcal{D}_{vv}^{-1}\right|\frac{s_v}{1-\zeta_v}h_v^{1-\zeta_v}} \\
&= 2B_0\left(b+C_\alpha\right)\sqrt{\sum_{v\in V}ch_v^2-c_vh_v^{1-\zeta_v}},
\end{aligned}$$

where

$$c = \frac{2(b+C_\alpha)^2}{n^2 N}, \quad c_v = \frac{16\Lambda^2}{N^2 n} \left|\mathcal{D}_{vv}^{-1}\right| \frac{s_v}{1-\zeta_v}.$$

Taking the partial derivative w.r.t. h_v and setting it to zero, yields the optimal value:

$$h_v^* = \left(\frac{(1-\zeta_v)c_v}{2c}\right)^{\frac{1}{1+\zeta_v}} = \left(\frac{4\Lambda^2 \left|\mathcal{D}_{vv}^{-1}\right| s_v n}{(b+C_\alpha)^2 N}\right)^{\frac{1}{1+\zeta_v}}. \tag{88}$$

Then, after substitution:

$$\begin{aligned}
\rho^* &\leq 2B_0(b+C_\alpha)\sqrt{\sum_{v\in V} c(h_v^*)^2 - c_v(h_v^*)^{1-\zeta_v}} \\
&= 2B_0\,(b+C_\alpha)\sqrt{\sum_{v\in V}(h_v^*)^2\left(c - \frac{2c}{(1-\zeta_v)}\right)} \\
&= 2B_0\,(b+C_\alpha)\sqrt{c\sum_{v\in V}\left(\frac{\zeta_v+1}{\zeta_v-1}\right)(h_v^*)^2}.
\end{aligned} \tag{89}$$

If we define $\zeta^* = \min_{v\in V}\zeta_v \geq 1$, $s_{\max} = \max_{v\in V} s_v$, we get:

$$\begin{aligned}
\rho^* &\leq \frac{2B_0(b+C_a)^2}{n}\sqrt{\frac{\zeta^*+1}{\zeta^*-1}}\sqrt{\frac{2}{N}\sum_{v\in V}\left(\frac{4\Lambda^2\left|D_{vv}^{-1}\right| s_v n}{(b+C_\alpha)^2 N}\right)^{\frac{2}{1+\zeta^*}}} \\
&\leq 8B_0\sqrt{\frac{\zeta^*+1}{\zeta^*-1}}\left[\Lambda^2(b+C_a)^{2\zeta^*}\right]^{\frac{1}{1+\zeta^*}} n^{-\frac{\zeta^*}{1+\zeta^*}} s_{\max}^{\frac{1}{1+\zeta^*}}\sqrt{N^{\frac{-(3+\zeta^*)}{1+\zeta^*}}\sum_{v\in V}\left|D_{vv}^{-1}\right|^{\frac{2}{1+\zeta^*}}} \\
&\leq 8B_0\sqrt{\frac{\zeta^*+1}{\zeta^*-1}}\left[\Lambda^2(b+C_a)^{2\zeta^*}\right]^{\frac{1}{1+\zeta^*}} n^{-\frac{\zeta^*}{1+\zeta^*}} s_{\max}^{\frac{1}{1+\zeta^*}} N^{-\frac{1}{1+\zeta^*}}\sqrt{\frac{1}{N}\sum_{v\in V}\left|D_{vv}^{-1}\right|^{\frac{2}{1+\zeta^*}}}.
\end{aligned} \tag{90}$$

Let us consider the last term. Since $D^{-1} = (\mathcal{L}+\gamma I)^{-1}$ is positive definite, its diagonal elements D_{vv}^{-1} are strictly positive. In addition, $g(x) = x^{\frac{2}{1+\zeta^*}}$ is a concave function for $x \geq 0$ and $\zeta^* > 1$, then, by applying Jensen's inequality, it follows that:

$$\frac{1}{N}\sum_{v\in V}\left|D_{vv}^{-1}\right|^{\frac{2}{1+\zeta^*}} \leq \left(\frac{1}{N}\sum_{v\in V} D_{vv}^{-1}\right)^{\frac{2}{1+\zeta^*}} = \left(\frac{1}{N}\operatorname{tr}\left((\mathcal{L}+\gamma I)^{-1}\right)\right)^{\frac{2}{1+\zeta^*}} = \left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\lambda_i+\gamma}\right)^{\frac{2}{1+\zeta^*}} \tag{91}$$

where $\{\lambda_1, ..., \lambda_n\}$ are the eigenvalues of the Laplacian matrix $\mathcal{L}$ in an increasing order. Our hypothesis on the graph G and the properties $\mathcal{L}$ imply $\lambda_i \geq 0$, for all $i \in \{1, ..., N\}$ and $\lambda_1 = 0$. Moreover, the multiplicity of λ_1 is equal to the number of connected components of G. Then, we can upper-bound the last term of Expr. 91 as follows:

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\lambda_i+\gamma} &\leq \begin{cases} \left(\frac{\#C(G)}{N}\right)\frac{1}{\gamma} + \left(\frac{N-\#C(G)}{N}\right)\frac{1}{\lambda_{\min+}} & \text{if } \#C(G) < N \\ \frac{1}{\gamma} & \text{if } \#C(G) = N \end{cases} \\
&= \beta\gamma^{-1} + (1-\beta)\lambda_{\min+}^{-1},
\end{aligned}$$

where $\#C(G)$ is the number of connected components of G, and $\lambda_{\min^+}$ is the smallest nonzero eigenvalue of the Laplacian. In the last equality, we define $\beta = \frac{\#C(G)}{N} \in [0,1]$ and use the convention that $(1-\beta)\lambda_{\min^+}^{-1} = 0$ when $\#C(G) = N$, which corresponds to a completely disconnected graph. With these elements, we can conclude:

$$\rho^* \leq 8B_0\sqrt{\frac{\zeta^*+1}{\zeta^*-1}}\left(\mathcal{T}_G\Lambda^2(b+C_a)^{2\zeta^*}\right)^{\frac{1}{1+\zeta^*}} n^{-\frac{\zeta^*}{1+\zeta^*}} N^{-\frac{1}{1+\zeta^*}} s_{\max}^{\frac{1}{1+\zeta^*}}.$$

■

C.3 Proof Theorem 3

Proof Lemma 14 implies that $\mathcal{H}_{\mathbb{G}}$ is a (β, B)-Bernstein class of vector-valued functions with $\beta = 1$ and $B = B_0$, and $\max_{v\in V}\sup_{(x,x')\in\mathcal{X}}|h_{f_v}(x,x')| \leq B_1$. By Lemma 17, we have that there exists a sub-root function such that $B\mathcal{R}(\mathcal{H}_{\mathbb{G}},\rho) \leq \varrho(\rho)$. Then, the hypotheses of Theorem 11 are satisfied, implying that with probability at least $1-\delta$, every $f \in \mathbb{G}$ satisfies:

$$\begin{aligned}
&\frac{1}{N}\sum_{v\in V}\mathbb{E}_{p_{z,v}}\left[\ell_v(f_v)(z) - \ell_v(r_v^\alpha)(z)\right] \\
&\leq \frac{B_1}{B_1-1}\left[\frac{1}{N}\sum_{v\in V}\left(\frac{1-\alpha}{n}\sum_{i=1}^{n}\frac{\left[f_v^2-(r_v^\alpha)^2\right](x_{v,i})}{2} + \frac{\alpha}{n}\sum_{i=1}^{n}\frac{\left[f_v^2-(r_v^\alpha)^2\right](x'_{v,i})}{2} - \frac{1}{n}\sum_{i=1}^{n}[f_v - r_v^\alpha](x'_{v,i})\right)\right] \\
&+ 2C(20^2)B_1\rho^* + \frac{16B_0^2C}{nN}\log\left(\frac{1}{\delta}\right) + \frac{24B_0B_1}{nN}\log\left(\frac{1}{\delta}\right). \qquad (92)
\end{aligned}$$

In particular for the minimum $\hat{\mathbf{f}}$ of Problem 34, we have that the term involving the empirical expectations is less than zero.

As detailed in Appendix A.1, we can easily verify that $PE(p^\alpha\|q) = \mathbb{E}_{p_{z,v}}\left[-\ell_v(r_v^\alpha)(z)\right] - \frac{1}{2}$, and by Expr. 32 $PE_v^\alpha(f_v) = \mathbb{E}_{p_{z,v}}\left[-\ell_v(f_v)(z)\right] - \frac{1}{2}$. Then for $\hat{\mathbf{f}}$ we can rewrite Expr. 92 as:

$$\frac{1}{N}\sum_{v\in V}\left[PE(p_v^\alpha\|q_v) - PE_v^\alpha(\hat{f}_v)\right] \leq 2C(20^2)B_1\rho^* + \frac{16B_0^2C}{nN}\log\left(\frac{1}{\delta}\right) + \frac{24B_0B_1}{nN}\log\left(\frac{1}{\delta}\right).$$

Alternatively, after applying the second point of Lemma 12 we can conclude:

$$\frac{1}{N}\sum_{v\in V}\mathbb{E}_{p_v^\alpha(y)}\left[\left[\hat{f}_v - r_v^\alpha\right]^2(y)\right] \leq 4C(20^2)B_1\rho^* + \frac{32B_0^2C}{nN}\log\left(\frac{1}{\delta}\right) + \frac{48B_0B_1}{nN}\log\left(\frac{1}{\delta}\right),$$

where Lemma 17 implies:

$$\rho^* \leq 8B_0\sqrt{\frac{\zeta^*+1}{\zeta^*-1}}\left(\mathcal{T}_G\Lambda^2(b+C_a)^{2\zeta^*}\right)^{\frac{1}{1+\zeta^*}} n^{-\frac{\zeta^*}{1+\zeta^*}} N^{-\frac{1}{1+\zeta^*}} s_{\max}^{\frac{1}{1+\zeta^*}}.$$

■

References

Rohit Agrawal and Thibaut Horel. Optimal bounds between f-divergences and integral probability metrics. *Journal of Machine Learning Research*, 22(128):1–59, 2021.

Mauricio A. Álvarez, Lorenzo Rosasco, and Neil D. Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

Francis Bach. Sum-of-Squares Relaxations for Information Theory and Variational Inference. *Foundations of Computational Mathematics*, 2024.

Francis Bach and Michael Jordan. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48, 2002.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Amir Beck and Luba Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23:2037–2060, 2013.

Moritz Beyreuther, Robert Barsch, Lion Krischer, Tobias Megies, Yannik Behr, and Joachim Wassermann. ObsPy: A Python Toolbox for Seismology. *Seismological Research Letters*, 81(3):530–533, 2010.

Jeremiah Birrell, Paul Dupuis, Markos A Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet. (f, γ)-divergences: Interpolating between f-divergences and integral probability metrics. *Journal of Machine Learning Research*, 23(39):1–70, 2022a.

Jeremiah Birrell, Markos A Katsoulakis, and Yannis Pantazis. Optimizing variational representations of divergences and accelerating their statistical estimation. *IEEE Transactions on Information Theory*, 68(7):4553–4572, 2022b.

Michel Broniatowski and Amor Keziou. Minimization of ϕ-divergences on sets of signed measures. *Studia Scientiarum Mathematicarum Hungarica*, 43(4):403–442, 2006.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2006.

Claudio Carmeli, Ernesto De Vito, and Alessandro Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 4 (04):377–408, 2006

Imre Csiszár. On topological properties of f-divergences. *Studia Scientiarum Mathematicarum Hungarica*, 2:329—-339, 1967.

Alejandro de la Concha, Nicolas Vayatis, and Argyris Kalogeratos. Online centralized non-parametric change-point detection via graph-based likelihood-ratio estimation. *arXiv:2301.03011*, 2023.

Alejandro de la Concha, Nicolas Vayatis, and Argyris Kalogeratos. Online non-parametric likelihood-ratio estimation by Pearson-divergence functional minimization. In *Proceedings*

of the International Conference on Artificial Intelligence and Statistics, pages 1189–1197, 2024.

Alejandro de la Concha, Nicolas Vayatis, and Argyris Kalogeratos. Collaborative non-parametric two-sample testing. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2025.

Aymeric Dieuleveut. *Stochastic approximation in Hilbert spaces*. Theses, Université Paris sciences et lettres, 2017.

André Ferrari, Cédric Richard, Anthony Bourrier, and Ikram Bouchikhi. Online change-point detection with kernels. *Pattern Recognition*, 133:109022, 2023.

GNS Science. GeoNet Aotearoa New Zealand Earthquake Catalogue, 1970.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, 2006.

Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86(3):335–367, 2011.

Masahiro Kato and Takeshi Teshima. Non-negative bregman divergence minimization for deep direct density ratio estimation. In *Proceedings of the International Conference on Machine Learning*, pages 5320–5333, 2021.

Solomon Kullback. *Information Theory and Statistics*. Wiley, 1959.

Erich L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Nature, Cham, Switzerland, 4 edition, 2022.

Xingguo Li, Tuo Zhao, Raman Arora, Han Liu, and Mingyi Hong. On faster convergence of cyclic block coordinate descent-type methods for strongly convex minimization. *Journal of Machine Learning Research*, 18(184):1–24, 2018.

Song Liu, Makoto Yamada, Nigel Collier, and Masashi Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72–83, 2013.

Nan Lu, Tianyi Zhang, Tongtong Fang, Takeshi Teshima, and Masashi Sugiyama. *Rethinking Importance Weighting for Transfer Learning*, pages 185–231. Springer, 2023.

Andreas Maurer. The Rademacher complexity of linear transformation classes. In *Learning Theory*, pages 65–78. Springer, 2006a.

Andreas Maurer. The rademacher complexity of linear transformation classes. In Gábor Lugosi and Hans Ulrich Simon, editors, *Learning Theory*, pages 65–78. Springer, 2006b.

Charles A. Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural Computation*, 17(1):177–204, 2005.

Roula Nassif, Stefan Vlaski, Cedric Richard, Jie Chen, and Ali H. Sayed. Multitask learning over graphs: An approach for distributed, streaming machine learning. *IEEE Signal Processing Magazine*, 37(3):14–25, 2020a.

Roula Nassif, Stefan Vlaski, Cédric Richard, and Ali H. Sayed. Learning over multitask graphs—part i: Stability analysis. *IEEE Open Journal of Signal Processing*, 1:28–45, 2020b.

Roula Nassif, Stefan Vlaski, Cédric Richard, and Ali H. Sayed. Learning over multitask graphs—part ii: Performance analysis. *IEEE Open Journal of Signal Processing*, 1:46–63, 2020c.

Duc Hoan Nguyen, Werner Zellinger, and Sergei Pereverzyev. On regularized radon-nikodym differentiation. *Journal of Machine Learning Research*, 25(266):1–24, 2024.

XuanLong Nguyen, Martin J. Wainwright, and Michael Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems*, 2008.

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, 2016.

Karl Pearson. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, 2007.

Benjamin Rhodes, Kai Xu, and Michael U. Gutmann. Telescoping density-ratio estimation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, 2020.

Cédric Richard, José Carlos M. Bermudez, and Paul Honeine. Online prediction of time series data with kernels. *IEEE Transactions on Signal Processing*, 57(3):1058–1067, 2009.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10(5):1299–1319, 1998.

Danuel Sheldon. Graphical Multi-Task Learning. Technical report, Cornell University, 2008.

David I. Shuman, Sunil K. Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013.

Alex J. Smola and Bernhard Schökopf. Sparse greedy matrix approximation for machine learning. In *International Conference on Machine Learning*, pages 911–918, 2000.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R.G. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(70):2389–2410, 2011. URL `http://jmlr.org/papers/v12/sriperumbudur11a.html`.

Ingo Steinwart, Don R Hush, Clint Scovel, et al. Optimal rates for regularized least squares regression. In *Conference on Learning Theory*, pages 79–93, 2009.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*, 2007.

Masashi Sugiyama, Taiji Suzuki, Yuta Itoh, Takafumi Kanamori, and Manabu Kimura. Least-squares two-sample test. *Neural networks*, 24:735–51, 2011a.

Masashi Sugiyama, Taiji Suzuki, Yuta Itoh, Takafumi Kanamori, and Manabu Kimura. Least-squares two-sample test. *Neural Networks*, 24(7):735–751, 2011b.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

Ameet Talwalkar, Sanjiv Kumar, and Henry Rowley. Large-scale manifold learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

Alexander Tartakovsky, Igor V. Nikiforov, and Michele Basseville. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. Taylor & Francis, CRC Press, 2014.

Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, 2000.

Stephen J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1): 3–34, 2015.

Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. In *Advances in Neural Information Processing Systems*, 2011.

Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, 25(5):1324–1370, 2013.

Yiming Ying and Massimiliano Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, 2007.

Niloofar Yousefi, Yunwen Lei, Marius Kloft, Mansooreh Mollaghasemi, and Georgios C. Anagnostopoulos. Local rademacher complexity-based learning guarantees for multi-task learning. *Journal of Machine Learning Research*, 19(38):1–47, 2018.

Werner Zellinger, Stefan Kindermann, and Sergei V Pereverzyev. Adaptive learning of density ratios in rkhs. *Journal of Machine Learning Research*, 24(395):1–28, 2023.

Kai Zhang, Ivor W. Tsang, and James T. Kwok. Improved nyström low-rank approximation and error analysis. In *International Conference on Machine Learning*, pages 1232–1239, 2008.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021.