Pontifícia Universidade Católica
do Rio de Janeiro

Kleyton Vieira Sales da Costa

Learning on Graphs via Generalized Divergence Measures

Dissertação de Mestrado

Dissertation presented to the Programa de Pós–graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Informática.

Advisor : Prof. Hélio Côrtes Vieira Lopes
Co-advisor: Prof. Ivan Fabio Mota de Menezes

Rio de Janeiro
April 2025

Pontifícia Universidade Católica
do Rio de Janeiro

Kleyton Vieira Sales da Costa

Learning on Graphs via Generalized Divergence Measures

Dissertation presented to the Programa de Pós–graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Informática. Approved by the Examination Committee:

Prof. Hélio Côrtes Vieira Lopes
Advisor
Departamento de Informática – PUC-Rio

Prof. Ivan Fabio Mota de Menezes
Co-advisor
Departamento de Engenharia Mecânica – PUC-Rio

Prof. Marcus Vinicius Soledade Poggi de Aragao
Departamento de Informática – PUC-Rio

Prof. Bernardo Andrade Lyrio Modenesi
Division of Biostatistics - University of Utah

Dr. Georges Miranda Spyrides
Departamento de Informática – PUC-Rio

Rio de Janeiro, April 10th, 2025

Kleyton Vieira Sales da Costa

BSc in Economics at Federal Rural University of Rio de Janeiro (UFRRJ). Worked as AI researcher at ExACTa (PUC-Rio) and Holistic AI.

Bibliographic Data

Costa, Kleyton Vieira Sales da

Learning on Graphs via Generalized Divergence Measures / Kleyton Vieira Sales da Costa; advisor: Hélio Côrtes Vieira Lopes; co-advisor: Ivan Fabio Mota de Menezes. – 2025.

78 f: il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2025.

Inclui bibliografia

1. Unsupervised Learning on Graphs – Teses. 2. Variational Autoencoders – Teses. 3. Generalized Divergence Measures – Teses. 4. Aprendizado Não-supervisionado em Grafos. 5. Inferência Aproximada. 6. Autoencoders Variacionais. 7. Generalizações para Medidas de Divergência . I. Lopes, Hélio Côrtes Vieira. II. Menezes, Ivan Fabio Mota de. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. IV. Título.

CDD: 004

To those who are living, those who are dead, and those who are to be born.

Acknowledgments

He who strives on and lives to strive Can earn redemption still.
— Faust, Part II, Act V

Education is the most effective instrument for building a fair society with equal opportunities for individuals, independently of their background. The ability to understand aspects of natural and social phenomena is rewarding. In a country like Brazil, advancing along this education path is both a privilege and a responsibility. With that said, I begin my acknowledgments by dedicating this work to the Brazilian people and the research funding institutions that made the development of this dissertation possible.

I dedicate this dissertation to my family, who provided all the emotional support needed for me to complete this further stage in my journey. I also thank Jessica Marques for all the care and patience during the years required to construct this dissertation.

I thank my advisors, Hélio Lopes, Ivan Menezes, and Bernardo Modenesi. Without their patient and committed collaboration, the progress of this work would have been much more stressful.

I want to thank the comments and suggestions of Professor Constantino Tsallis, who kindly contributed to the intuition of this work.

Finally, I list long-time friends and partners, and those I have had the privilege to interact with over the past few years (any omissions are my memory's fault): Adriano Koshiyama, Alexandres Cabús, André Modenesi, Fellipe Tavares, Antonio Jose Alves Junior, Felipe Leite, Laércio Lucchesi, Josef Kamysek, Jonatas Grosman, Fernando Freitas, Georges Spyrides, Carlos Vinicios, Rafael Stutz, Carlos Miranda, Camila Alves, Dayana, João Gallas, Marcos Kalinowski, Marco Molinaro, Markus Endler, Adriano Branco, Simone Barbosa, Pedro Caldeiras, Enio Blay, Igor Caetano, Thiago Lamenza, Umar Mohammed, Cristian Munoz, Franklin Fernandez, Zekun Wu, André Novaes, Tarciso Gouveia, Rodrigo de Lamare.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Abstract

Costa, Kleyton Vieira Sales da; Lopes, Hélio Côrtes Vieira (Advisor); Menezes, Ivan Fabio Mota de (Co-Advisor). **Learning on Graphs via Generalized Divergence Measures**. Rio de Janeiro, 2025. 78p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

This master dissertation investigates the effectiveness of generalized information measures for learning on graphs (LoG). The variational graph autoencoders framework proposed by Kipf and Welling (2016b) was modified by generalized divergence measures as part of the learning objective to delimit the research scope. Then, the main contributions of this work are: (i) the κ-divergences - a unified representation for generalized divergence measures; (ii) two novel families of divergences, δ and η; and (iii) the generalized graph variational autoencoders (GGVA) - a variational graph autoencoders framework based on κ-divergences. The experiments on LoG, using five citation network datasets and a Brazilian power grid network dataset, indicate that GGVA outperforms baseline models in node classification and link prediction, considering time efficiency and average precision. The qualitative analysis of the learned embeddings of GGVA indicates a good enough capacity to distinguish classes.

Keywords

Unsupervised Learning on Graphs; Approximate Inference; Variational Autoencoders; Generalized Divergence Measures.

Resumo

Costa, Kleyton Vieira Sales da; Lopes, Hélio Côrtes Vieira (Orientador); Menezes, Ivan Fabio Mota de (Co-orientador). **Aprendizagem em Grafos via Medidas de Divergência Generalizadas**. Rio de Janeiro, 2025. 78p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Esta dissertação de mestrado investiga a efetividade de generalizações de medidas de informação para aprendizado em grafos. Para delimitar o escopo de pesquisa, a função de custo do *variational graph autoeconders* proposto por Kipf and Welling (2016b) foi modificada por meio da incorporação de generalizações de medidas de divergência. Dessa maneira, as principais contribuições deste trabalho são: (i) κ-divergências - uma representação unificada para generalizações de medidas de divergência; (ii) duas novas famílias de divergências, δ e η; e (iii) o desenvolvimento do *generalized variational graph autoencoders* (GGVA), um arcabouço de *variational graph autoencoders* baseado em κ-divergências. Os experimentos realizados em tarefas de aprendizado em grafos, utilizando cinco conjuntos de dados de redes de citação e um conjunto de dados para a rede de distribuição de energia elétrica do Brasil, indicam que o GGVA supera os modelos de referência em dois tipos de tarefas: classificação de nós e previsão de relacionamentos, considerando tempo de execução e precisão média. Os resultados qualitativos para os *embeddings* treinados do GGVA indicam uma capacidade satisfatória para distinguir classes.

Palavras-chave

Aprendizado Não-supervisionado em Grafos; Inferência Aproximada; Autoencoders Variacionais; Generalizações para Medidas de Divergência .

Table of Contents

List of Figures

List of Tables

List of Abbreviations

LoG – Learning on graphs

VI – Variational inference

VAE – Variational autoencoders

VGAE – Variational graph autoencoders

GGVA – Generalized variational autoencoders

GNN – Graph neural networks

NESM – Non-extensive statistical mechanics

To imagine is the characteristic act, not of the poet's mind, or the painter's, or the scientist's, but of the mind of man.

Jacob Bronowski, *The Ascent of Man.*

1
Introduction

This chapter describes this work's motivation, objectives, and document organization.

1.1
Motivation

1.1.1
Graphs are everywhere

A graph is a simple yet powerful mathematical structure representing relationships in numerous natural and artificial systems. This flexible representation appears in various domains (Newman, 2018; Easley et al., 2010; Bullmore and Sporns, 2009) - from biological networks modeling protein interactions and neural pathways to social networks depicting friendships and collaborations to technological systems like computer networks and transportation infrastructure. The flexibility of graph structures allows them to capture complex relationships in ways that linear or hierarchical data structures cannot.

The mathematical foundation of graph theory, pioneered by Leonhard Euler in 1736 with the Seven Bridges of Königsberg problem (Euler, 1741), has evolved into a rich field with applications in computer science, operations research, chemistry, linguistics, and numerous other disciplines. In this work, our focus is on the intersection between deep learning and graphs. As Figure 1.2 shows, we can define three basic learning tasks on graphs: graph classification, node classification, and link prediction. In the following, we describe the challenges and contributions of this field based on examples extracted from the literature.

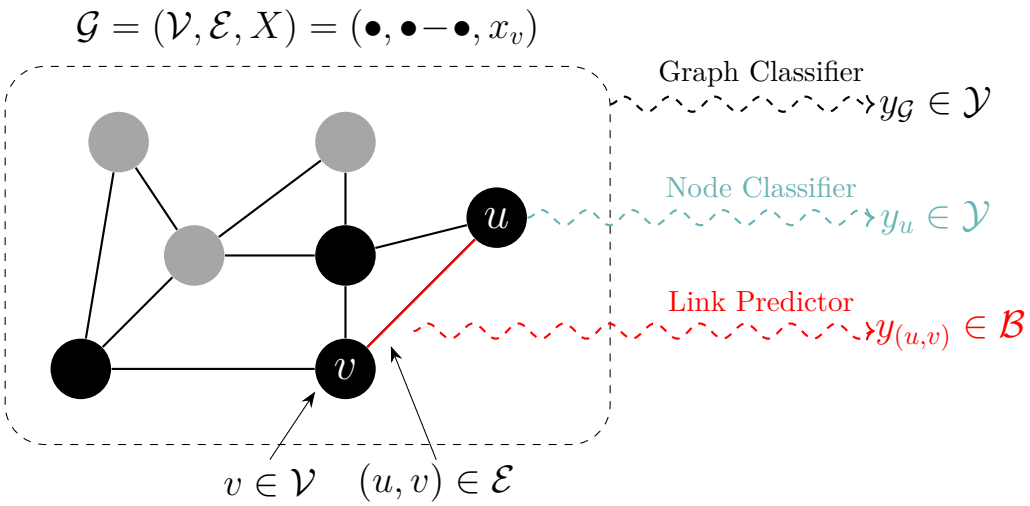

Figure 1.1: Illustration of graph tasks. A simple graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, X)$ consists of nodes $v \in \mathcal{V}$, edges $(u, v) \in \mathcal{E}$, and node features X. The framework enables multiple learning tasks: graph classification ($y_{\mathcal{G}} \in \mathcal{Y}$), node classification ($y_u \in \mathcal{Y}$), and link prediction ($y_{(u,v)} \in \mathcal{B}$)

1.1.2 Learning on graphs

Modern deep learning allows models with multiple processing layers to learn data representations with several levels of abstraction (LeCun et al., 2015). These models have been accumulating state-of-the-art performance with architectures designed for simple sequences (Hochreiter and Schmidhuber, 1997; Vaswani et al., 2017) and grids (Fukushima, 1980; LeCun et al., 1998; Krizhevsky et al., 2017; Dosovitskiy et al., 2020).

Working with graphs presents unique challenges due to their inherent complexity[1]. As Figure 1.2 illustrates, graphs have intricate topological structures without spatial locality as a reference, with a single graph having multiple representations, which makes the definition of effective deep learning architectures challenging. In this case, a desired characteristic to define deep learning over graphs is that they should be permutation invariant: the function that guides the architecture does not depend on the arbitrary ordering of the graph (Hamilton, 2020).

[1]Considering a Geometric Deep Learning blueprint (Bronstein et al., 2021), sequences, grids, and graphs are functions of the same framework, only changing the domain and the symmetry group.

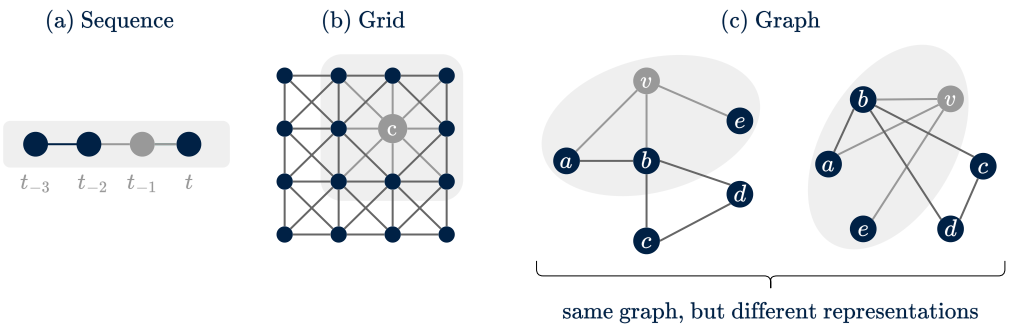

Figure 1.2: Comparison of three fundamental data structures in deep learning. (a) sequences, (b) grids, and (c) graphs

The field of learning on graphs (LoG)[2] has emerged as a powerful framework for addressing deep learning on graphs. Some relevant applications of LoG techniques are polypharmacy side effects prediction (Zitnik et al., 2018; Johnson et al., 2024), drug discovery (Wieder et al., 2020; Stokes et al., 2020; Zhang et al., 2022; Liu et al., 2023), traffic prediction (Derrow-Pinion et al., 2021), global weather forecasting (Lam et al., 2023), flood prediction (Roudbari et al., 2024), load forecasting in power grid networks (Hansen et al., 2022; Liao et al., 2021), and football tactics (Wang et al., 2024).

Recent advancements in large language models also employ graph-based approaches to improve model reasoning and decrease hallucination. Two promising approaches involve (i) leveraging graph retrieval augmentation techniques, which integrate knowledge graphs to increase the model results with graph-structured relationships (Edge et al., 2024), and (ii) using graph neural networks (GNN) models to enhance the quality and reliability of generated outputs (Mavromatis and Karypis, 2024).

This literature shows the robust characteristics of LoG in addressing solutions to complex scientific and social challenges. This aspect of learning in complex systems motivates us to explore the connections between LoG and *non-extensive statistical mechanics*.

1.1.3 Learning at the edge of chaos

In deep learning architectures generally, and LoG architectures specifically, the predominant approach relies on learning objectives derived from the Boltzmann-Gibbs-Shannon entropic functional (e.g., cross-entropy, Kullback-Leibler divergence, Jensen-Shannon divergence). Based on the characteristics of weak chaotic systems described below, we argue that *non-extensive statis-*

[2]More details about LoG techniques are presented in the Appendix A.

tical mechanics (NESM) and *generalized measures of information* could potentially enhance the solution of LoG problems. These alternative frameworks may better capture the complex dynamics (like long-range correlations) inherent in many graph-structured data, potentially leading to more robust and accurate learning algorithms, as observed in other deep learning applications.

The Boltzmann–Gibbs (BG) statistical mechanics (Boltzmann, 1872; Gibbs, 1902) represents one of the fundamental pillars of modern theoretical physics. Its theoretical foundation rests upon the optimization of the BG additive entropic functional $H_{BG} = -k \sum_i p_i \ln p_i$, which has profound connections to Shannon's information theory (Shannon, 1948). Indeed, the mathematical isomorphism between thermodynamic entropy and information-theoretic uncertainty sets a framework for understanding physical systems in terms of their information content and statistical properties. This formalism has been experimentally validated in numerous contexts. It remains mathematically legitimate for broad classes of nonlinear dynamical systems characterized by positive maximal Lyapunov exponents (strong chaos).. Strongly chaotic dynamical systems are characterized by their exponential sensitivity to initial conditions, quantified by a positive maximal Lyapunov exponent ($\lambda_{\max} > 0$) (Benettin et al., 1980). This exponential divergence implies that nearby trajectories in phase space separate, on average, exponentially fast. A direct consequence of this rapid divergence is a swift loss of information regarding the system's past state. This manifests statistically as a rapid decay of time correlation functions, often exhibiting an exponential decrease with increasing time lag. Consequently, the system possesses a short "memory"; its state at a given time is significantly correlated only with its states in the very recent past, while correlations with distant past states quickly become negligible. This behavior defines short-range dependencies, where the system's dynamics are effectively influenced only by proximate events in its history, a feature distinguishing strong chaos from weakly chaotic regimes exhibiting slower, power-law correlation decay and long-range memory effects.

When applied to systems exhibiting weak chaos, the BG framework demonstrates significant limitations where the maximal Lyapunov exponent vanishes - a condition prevalent in complex natural, artificial, and social systems (Tsallis, 2023). To address these limitations, a generalization was introduced by Tsallis (1988) based on the non-additive entropic functional $H_q = k\frac{1-\sum_i p_i^q}{q-1}$ with $q \in \mathbb{R}$ and $H_1 \to H_{BG}$, which extends statistical mechanics (and information theory) beyond its conventional boundaries. This generalized entropy, often called Tsallis entropy, introduces the entropic index q that quantifies the degree of non-extensivity and must be calculated from first

principles for specific classes of weakly chaotic systems. The transition from strong to weak chaos marks the boundary where Boltzmann-Gibbs statistics becomes inadequate and NESM become necessary (Tsallis et al., 2005; Tirnakli et al., 2007).

The NESM paradigm has been applied in many probabilistic approaches for deep learning. Silva et al. (2024) evaluates the performance of Tsallis divergence as a learning objective for generative flow networks. Zhu et al. (2023) applies Tsallis divergence in policy optimization approaches in reinforcement learning. Zimmert and Seldin (2021) presents a general analysis of online mirror descent algorithms regularized by Tsallis entropy.

This work investigates the connection between learning on graphs and generalized information measures. Our intuition starts with the following induction: If generalized measures of information (like NESM) are accurate for modeling complex networks (Wen and Jiang, 2019; de Oliveira et al., 2021; Robledo and Velarde, 2022; Tsallis, 2023), then *can generalized measures of information be used effectively in learning problems on graphs?*

1.2 Objectives

The scope to investigate if generalized measures of information are effective for LoG tasks is vast. Based on this, we delimited our research scope, applying generalized divergence measures as part of the learning objective of the variational graph autoencoders proposed by Kipf and Welling (2016b).

The main objectives of this master's thesis are:

1. To propose a possible unified representation for generalized divergence measures and a novel set of generalized divergence measures to be used as part of learning objectives in LoG tasks;

2. To propose and evaluate a new unsupervised learning framework, the generalized graph variational autoencoders. The evaluation is based on benchmark datasets commonly used in the literature and a new dataset extracted from the Brazilian power grid network.

1.3 Major Contributions

1. We propose three novel divergence measures that can be used as learning objectives in deep learning models. These divergences generalize previously used ones and allow for tuning the divergence to prior beliefs,

considering an extra parameter. The parameter tuning increases the exploration for more flexible and accurate results. Experiment I evaluates their behavior to approximate complex distributions based on a simple prior. The results indicate that our divergences are mass-covering (inclusive) in approximation inference scenarios;

2. We proposed an effective framework for unsupervised learning on graphs. In Experiment II, we observe that our proposed unsupervised learning framework for graphs achieves competitive performance - in terms of accuracy and computational efficiency - against the baseline models when applied for node classification tasks;

3. We present a novel dataset for learning on graphs - the Brazilian power grid network (BPGN). In Experiment III, we show that GGVA framework outperforms the baseline models in link prediction applied in BPGN;

1.4 Master Thesis Organization

This master's thesis is organized as follows:

– **Chapter 2** describes the methodological basis of our work. We present an optimization perspective of the approximate inference problem. We describe how the literature has advanced to (i) deal with intractable joint densities via variational inference methods, (ii) possible generalizations proposed to these advances, and (iii) how to use these advances in LoG using variational graph autoencoders.

– **Chapter 3** describes our proposed approach for generalized divergence measures. We propose a formal notation representing a wide class of generalized divergences, called κ-divergences. Within this class, we introduce two novel families of divergences: δ and η divergences and the behavior of this divergence in a variational inference problem;

– **Chapter 4** presents a novel unsupervised learning framework for graphs, named *generalized graph variational autoencoder* (GGVA). We apply this LoG methodology in two experiments: (i) a benchmark for graph problems (link prediction and node classification) to evaluate our GGVA framework against other traditional approaches and (ii) a link prediction task based on the Brazilian power grid network;

– **Chapter 5** presents the conclusion and future work.

- **Appendix A** presents a high-level overview of learning on graphs. We describe the main architectures for graph neural networks, the traditional tasks in LoG, and some of the theoretical aspects that support graph neural networks;
- **Appendix B** presents a high-level overview of divergence measures. Here, we describe the main divergence measures and some important properties that support the use of these measures in theoretical and applied setups.
- **Appendix C** presents detailed results for the experiments.

2
Background and Literature

A well-established approach to unsupervised LoG using generative modeling is the *variational graph autoencoders* (VGAE) framework introduced by Kipf and Welling (2016b). This chapter presents the background for the VGAE framework and for our generalized approach (described in detail in Chapter 4). In summary, this chapter describes:

1. an optimization perspective of the approximate inference problem. We describe how the literature has advanced to (i) deal with intractable joint densities via variational inference methods, (ii) possible generalizations proposed to these advances, and (iii) how to use these advances in a learning setup using variational autoencoders;

2. how to use variational autoencoder for learning on graphs. We present a seminal model that introduces how to employ an optimization perspective of the approximate inference problem as a tool for unsupervised learning on graphs. This approach has three components: an inference model, a generative model, and a loss function.

2.1
An Optimization View of Approximate Inference

2.1.1
Motivation

When we think about machine learning problems, two main schools of thought define how the modeling process will occur: the *discriminative* and the *generative* (or Bayesian) (Jebara, 2004). For a set of data instances X and labels Y, the discriminative approach defines a direct attempt to map an input-output relationship for classification or regression tasks, capturing the conditional probability $P(Y|X)$. For the second type of models, a generative probabilistic model approach captures the joint probability $P(X, Y)$ if the labels are available or the $P(X)$ if not. Figure 2.1 summarizes the basic difference between these two views considering a graph-based problem. In essence, generative models try to model how the data is distributed in the

latent space, and discriminative models define a clever decision boundary to separate the data in the latent space.

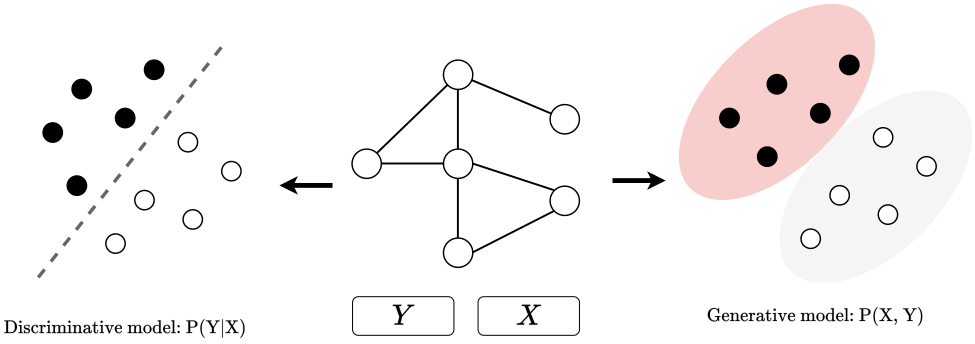

Figure 2.1: Generative and discriminative models for graph-based problems

Modern Bayesian statistics depend on models for which the posterior is challenging to compute, and algorithms are designed to approximate it (Blei et al., 2017; Järvenpää and Corander, 2023). Historically, a well-established family of approximate algorithms is the *sampling* methods, which produce answers by repeatedly generating random numbers from a distribution of interest. The Markov chain Monte Carlo (MCMC) (Geyer, 1992; Geman and Geman, 1984) method is a famous example of this family. The main drawbacks of sampling-based methods are related to running time, mainly suffering from the curse of dimensionality. The methods are guaranteed to find a good enough solution given a sufficient amount of time and regularity conditions. However, given the limited time available, it is difficult to tell how close they are to a good solution. Another point is that MCMC methods, for example, require an appropriate sampling technique, which is challenging to build.

As a fast alternative for sampling methods, variational inference (VI) methods (Jordan et al., 1999) approximates an intractable distribution p by a tractable distribution $q \in \mathcal{Q}$. With this approach, an optimization problem is formulated to minimize information loss between p and q. Given their significant advantages in computational efficiency and scalability (Ganguly and Earp, 2021) compared to traditional sampling techniques, particularly for large datasets, we explore learning methods that use VI as the backbone.

2.1.2 Problem setup of approximate inference

Let $x = x_{1:n}$ be a set of observed variables and $z = z_{1:m}$ be a set of latent variables, with *joint density* $p(z, x)$. Given the observations, the inference problem is to compute the *conditional density* of the latent variables. The

conditional can produce point or interval estimates of the latent variables and form predictive densities of new data. We can define the *conditional density* as:

$$p(z|x) = \frac{p(z,x)}{p(x)} \tag{2-1}$$

where $p(x)$ is the marginal density of the observations (the *evidence*). We compute it by marginalizing out the latent variables from the joint density:

$$p(x) = \int p(z,x)dz \tag{2-2}$$

This *evidence* integral is often unavailable in closed form or requires exponential time to compute. The *evidence* is what we need to calculate the conditional from the joint, which is why inference in such models is a challenging task.

2.1.3 Evidence lower bound

In information theory (Cover, 1999), the Kullback-Leibler divergence D_{KL} is used to measure the information related within two distributions. Let P and Q be two continuous probability distributions, $D_{KL}(P \parallel Q)$ is given by

$$D_{KL}(P \parallel Q) = \int_{\mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right) dx \tag{2-3}$$

We define a family $\mathcal{Q}$ of densities over the latent variables. Each $q(z) \in \mathcal{Q}$ represents a candidate approximation to the exact conditional distribution. The objective is to identify the optimal candidate, which minimizes D_{KL} to the same conditional. Through this approach, the inference problem is formulated as the following optimization problem:

$$q^*(z) = \arg\min_{q(z)\in\mathcal{Q}} D_{KL}(q(z) \parallel p(z|x)) \tag{2-4}$$

After being determined, $q^*(\cdot)$ constitutes the most accurate approximation of the conditional within the family $\mathcal{Q}$ - the structural complexity of the family determines the computational complexity of this optimization. However, this objective function is not directly computable because it necessitates calculating the logarithm of the *evidence*. The D_{KL} divergence is expressed as:

$$D_{KL}(q(z) \parallel p(z|x)) = \mathbb{E}[\log q(z)] - \mathbb{E}[\log p(z|x)], \tag{2-5}$$

where all expectations are taken for $q(z)$.

Expanding the conditional probability, we obtain:

$$D_{KL}(q(z) \parallel p(z|x)) = \mathbb{E}[\log q(z)] - \mathbb{E}[\log p(z,x)] + \log p(x). \tag{2-6}$$

Equation 2-6 demonstrates the explicit dependence on $\log p(x)$. An alternative approach is to optimize an objective function equivalent to the D_{KL} up to an additive constant. This function is called the *evidence lower bound* (ELBO), defined as:

$$ELBO(q) = \mathbb{E}[\log p(z, x)] - \mathbb{E}[\log q(z)] \tag{2-7}$$

where the ELBO is the negative D_{KL} from Equation 2-6 plus $\log p(x)$, which remains constant to $q(z)$[1]. We can decompose the ELBO as the sum of the expected log-likelihood of the data and the negative D_{KL} between the prior $p(z)$ and $q(z)$:

$$\begin{aligned} ELBO(q) = &\ \mathbb{E}[\log p(z)] + \mathbb{E}[\log p(x|z)] - \mathbb{E}[\log q(z)] \\ = &\ \mathbb{E}[\log p(x|z)] - KL(q(z) \parallel p(z)) \end{aligned} \tag{2-8}$$

where the first term represents an expected likelihood that favors densities that concentrate their mass on configurations of the latent variables that effectively explain the observed data. The second term constitutes the negative divergence between the variational density and the prior distribution. Then, the variational objective reflects the fundamental balance between likelihood and prior characteristics of Bayesian inference.

As already mentioned, the complexity of the variational family $\mathcal{Q}$ determines the complexity of the optimization problem. In *mean-field variational family*, the latent variables are mutually independent, each governed by a distinct factor in the variational density. We can have more complex families adding dependencies between variables (Saul and Jordan, 1995; Barber and Wiegerinck, 1998) or consider mixtures of the variational family (Bishop et al., 1997).

2.1.4 Possible generalizations of variational inference

A limitation of the D_{KL} arises from its sensitivity to tail behavior. If the approximating distribution Q assigns a vanishingly small probability $q(x) \approx 0$ to events or regions where the true or target distribution P with non-negligible probability mass $p(x) > 0$, the ratio $\frac{p(x)}{q(x)}$ becomes extremely large. Consequently, the logarithmic term $\log(\frac{p(x)}{q(x)})$ contributes a disproportionately large positive value to the divergence, even if $p(x)$ itself is not large. This means that the D_{KL} heavily penalizes models (Q) that fail to cover the probability mass of the target (P), especially in the tails. In the extreme case where $q(x) = 0$ for some x where $p(x) > 0$, the D_{KL} becomes infinite, reflecting

[1] In this context, we have the duality that maximizing the ELBO is mathematically equivalent to minimizing the D_{KL}.

its strict requirement for the absolute continuity of P for Q.

We argue that generalized information measures, emphasizing divergence measures, offer a flexible framework for learning in complex networks with long-range correlations (Tsallis, 2009; Liang et al., 2025). Our analysis of VI literature reveals promising directions for these generalized approaches, demonstrating their theoretical and empirical advantages for optimization in high-dimensional spaces.

Blei et al. (2017) describes that a promising avenue of research is to develop VI methods that optimize other measures, such as α-divergence. Following the same intuition, the literature proposed approaches to address other divergence measures in the VI problem. Based on our current knowledge, the literature presents reasonable generalization approaches for VI (Li and Turner, 2016; Minka, 2005; Knoblauch et al., 2019; Regli and Silva, 2018; Hernández-Lobato et al., 2016; Wang et al., 2021). Following, we briefly comment on three of these approaches.

Alpha-Beta VI Regli and Silva (2018) introduced an extended sAB divergence (Cichocki et al., 2011) and its relationship with VI. One advantage is that the minimization directly via the divergence, without any extra definition of an equivalent lower bound. The formulation of a $\alpha\beta$ divergence presented in Regli and Silva (2018) is,

$$D_{AB}^{\alpha,\beta}(p||q) = -\frac{1}{\alpha\beta}\int\left(p(\theta)^{\alpha}q(\theta)^{\beta} - \frac{\alpha}{\alpha+\beta}p(\theta)^{\alpha+\beta} - \frac{\beta}{\alpha+\beta}q(\theta)^{\alpha+\beta}\right)d\theta. \tag{2-9}$$

The ELBO associated with this divergence is defined as,

$$\begin{aligned} D_{AB}^{\alpha,\beta}(q(\theta)||p(\theta|\mathbf{X})) = & -\tfrac{1}{\alpha\beta}\int\left(q(\theta)^{\alpha}p(\theta|\mathbf{X})^{\beta} - \tfrac{\alpha}{\alpha+\beta}q(\theta)^{\alpha+\beta} - \tfrac{\beta}{\alpha+\beta}p(\theta|\mathbf{X})^{\alpha+\beta}\right)d\theta \\ = & -\tfrac{1}{\alpha\beta}\int\left(q(\theta)^{\alpha}\left(\tfrac{p(\theta,\mathbf{X})}{p(\mathbf{X})}\right)^{\beta} - \tfrac{\alpha}{\alpha+\beta}q(\theta)^{\alpha+\beta} - \tfrac{\beta}{\alpha+\beta}\left(\tfrac{p(\theta,\mathbf{X})}{p(\mathbf{X})}\right)^{\alpha+\beta}\right)d\theta \\ = & -\tfrac{1}{\alpha\beta}[p(\mathbf{X})^{-\beta}\int q(\theta)^{\alpha}p(\theta,\mathbf{X})^{\beta}d\theta - \tfrac{\alpha}{\alpha+\beta}\int q(\theta)^{\alpha+\beta}d\theta - \tfrac{\beta}{\alpha+\beta} \\ & p(\mathbf{X})^{-(\alpha+\beta)}\int p(\theta,\mathbf{X})^{\alpha+\beta}d\theta] \end{aligned} \tag{2-10}$$

Generalized VI Knoblauch et al. (2019) introduces a novel framework that extends traditional Bayesian inference by addressing key limitations of the standard approach. The first step is reframe the Bayesian inference as an infinite-dimensional optimization problem where the posterior $q_B^*(\theta)$ minimizes $\mathbb{E}_{q(\theta)}[\sum_{i=1}^{n}\ell(\theta,x_i)] + \mathrm{KLD}(q\|\pi)$ over all probability measures $\mathcal{P}(\Theta)$. Building on this insight, they developed the Rule of Three (RoT), defined as

$$q^*(\theta) = \arg\min_{q\in\Pi}\{\mathbb{E}_{q(\theta)}[\sum_{i=1}^{n}\ell(\theta, x_i)] + \mathcal{D}(q\|\pi)\} \stackrel{\text{def}}{=} P(\ell, D, \Pi), \tag{2-11}$$

where ℓ is a loss function, $\mathcal{D}$ is a divergence measuring uncertainty, and Π is a feasible space of posteriors. This formulation provides modularity to address three challenges in traditional Bayesian inference: misspecified priors, misspecified likelihood models, and computational constraints. GVI is a tractable case where $\Pi = \mathcal{Q} = \{q(\theta|\gamma) : \gamma \in \Gamma\} \subset \mathcal{P}(\Theta)$ is a parameterized subset. They prove several theoretical properties for GVI, including frequentist consistency and an interpretation as an approximate evidence lower bound for certain divergences. Empirical evaluations on Bayesian Neural Networks and Deep Gaussian Processes demonstrate that appropriately chosen GVI posteriors can significantly outperform standard VI and alternative approximation methods, particularly in handling misspecified models and prior distributions.

f-divergence VI Wan et al. (2020) introduces f-divergence VI (f-VI), which generalizes VI to all f-divergences. Starting with a surrogate f-divergence that shares statistical consistency with the original f-divergence, the framework unifies existing VI methods like KL-VI, Rényi's α-VI, and χ-VI while providing a standardized toolkit for VI with arbitrary f-divergences. The authors derive a general f-variational bound $L_f(q, D) = \mathbb{E}_{q(z)}\left[f^*\left(\frac{p(z,D)}{q(z)}\right)\right] \geq f^*(p(D))$ that serves as a sandwich estimate of marginal likelihood. They develop optimization schemes using reparameterization tricks, importance weighting, and Monte Carlo approximation, formalized as

$$\nabla_\theta L_f^{rep}(q_\theta, D) = \nabla_\theta \mathbb{E}_{p(\varepsilon)}\left[f^*\left(\frac{p(g_\theta(\varepsilon), D)}{q_\theta(g_\theta(\varepsilon))}\right)\right] \tag{2-12}$$

Additionally, they propose a mean-field approximation that generalizes coordinate ascent VI (CAVI) for f-VI, with update rules depending on whether $f \in \mathcal{F}_0$ or $f \in \mathcal{F}_1$. The paper demonstrates the effectiveness of f-VI through experiments on synthetic data, Bayesian neural networks, and variational autoencoders, showing that it sometimes outperforms state-of-the-art variational methods while offering greater flexibility in choosing divergence measures.

These three approaches demonstrate that applying generalized divergence measures provides a principled theoretical foundation for VI while enhancing flexibility, robustness, and performance. These frameworks enable tailored inference procedures that adapt to model misspecification, complex posterior geometries, and computational constraints by decoupling the optimization objective from specific divergence metrics. Our work explores these

insights into learning on graph-structured data.

In the next section, we present the backbone of our proposed method for unsupervised learning on graphs using variational autoencoders.

2.2 Learning on Graphs via Variational Autoencoders

Variational autoencoders (VAE) are a practical implementation and extension of VI principles within a neural network framework. This section briefly introduces the VAE framework and how to use it to learn graph-structured data using the node's structure and associated features.

2.2.1 Motivation

The VAE framework (Kingma and Welling, 2013; Rezende et al., 2014) provides a method for jointly learning deep latent-variable models and corresponding inference models using gradient-based optimization methods. Figure 2.2 represents the VAE framework (Kingma and Welling, 2019). A VAE learns stochastic mappings between an observed space $\mathcal{X}$ - in many cases, $\mathcal{X}$ have a complicated empirical distribution $q(\mathcal{X})$ - and a latent space $\mathcal{Z}$ - where $\mathcal{Z}$ have a simple distribution. The *generative model* learns a joint distribution $p(X, Z)$ that is in general factorized as $p(\mathcal{X}, Z) = p(\mathcal{Z})p(\mathcal{X}|\mathcal{Z})$, with a prior distribution over latent space $p(\mathcal{Z})$, and a stochastic decoder $p(\mathcal{X}|\mathcal{Z})$. The *inference model* $q(\mathcal{Z}, \mathcal{X})$, or stochastic encoder, approximates the generative model's true (and intractable) posterior $p(\mathcal{Z}|\mathcal{X})$.

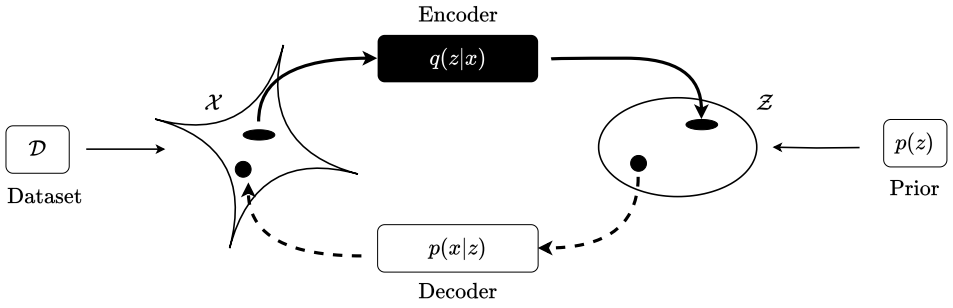

Figure 2.2: A simplified representation of the learning process in a VAE framework

2.2.2 Three learning components

Kipf and Welling (2016b) introduced the extension of VAE in unsupervised learning for graphs. As presented in Figure 2.3, VGAE is a framework

with three learning components: an inference model based on a graph convolutional network (GCN)(Kipf and Welling, 2016a) encoder, a generative model based on a simple inner product decoder, and a learning objective based on a variational lower bound.

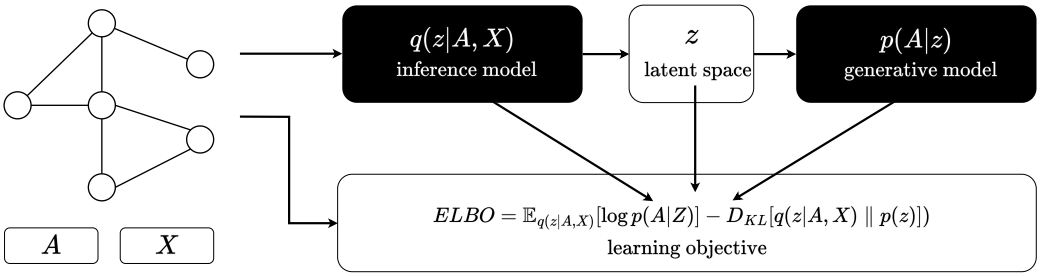

Figure 2.3: A simplified representation of the three learning components in a VGAE framework. The inference mode, the generative model, and the learning objective

To apply the VGAE framework, we consider a graph $\mathcal{G}$ along with its adjacency matrix A and degree matrix D. The basic definitions of these concepts are presented below.

Definition 2.1 (Undirected Graph) *An undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, X)$ consists of a set of nodes $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$, a set of edges $\mathcal{E} \subseteq \{\{u, v\} \mid u, v \in \mathcal{V}, u \neq v\}$, and a feature matrix $X \in \mathbb{R}^{n \times d}$ where each row x_i corresponds to a d-dimensional feature vector of node v_i.*

Definition 2.2 (k-Neighborhood) *The k-neighborhood of a node u, denoted as $\mathcal{N}_u^{(k)}$, is the set of all nodes that are exactly k hops away from u. Formally:*

$$\mathcal{N}_u^{(1)} = \{v \in \mathcal{V} \mid \{u, v\} \in \mathcal{E}\} \tag{2-13}$$

$$\mathcal{N}_u^{(k)} = \{v \in \mathcal{V} \mid \textit{shortest path from } u \textit{ to } v \textit{ has length } k\} \tag{2-14}$$

Definition 2.3 (Adjacency Matrix) *The adjacency matrix $A \in \mathbb{R}^{n \times n}$ of $\mathcal{G}$ represents the connectivity pattern between nodes, where:*

$$a_{uv} = \begin{cases} 1, & \textit{if } \{u, v\} \in \mathcal{E} \textit{ (or equivalently, if } v \in \mathcal{N}_u^{(1)}\textit{)} \\ 0, & \textit{otherwise} \end{cases} \tag{2-15}$$

For an undirected graph, A is symmetric, i.e., $a_{uv} = a_{vu} \quad \forall u, v \in \mathcal{V}$.

Definition 2.4 (Degree Matrix) *The degree matrix $D \in \mathbb{R}^{n\times n}$ of $\mathcal{G}$ is a diagonal matrix where each diagonal element d_{ii} represents the degree of node v_i, i.e., the number of edges incident to v_i:*

$$d_{ii} = \sum_{j=1}^{n} a_{ij} = |\mathcal{N}_{v_i}^{(1)}| \tag{2-16}$$

All off-diagonal elements of D are zero: $d_{ij} = 0$ for $i \neq j$.

Considering a two-layer GCN that parametrizes the inference model in VGAE,

$$q(Z|A, X) = \prod_{i=1}^{N} q(z_i|A, X) \tag{2-17}$$

where $q(z_i|A, X) = \mathcal{N}(z_i|\mu_i, diag(\sigma_i^2))$, $\mu = GCN_\mu(A, X)$ is the matrix of mean vector μ_i, and $\log\sigma = \Psi_\sigma(A, X)$. The two layer GCN is defined as $GCN(X, A) = \tilde{A}\text{ReLU}(\tilde{A}XW_0W_1)$ with weight matrices W_i.

There are some considerations about the implementation of this inference model. First, GCN_μ and GCN_σ share the first layer parameters W_0. The activation function follows the traditional formulation $ReLU(x) = max(0, x)$. And $\tilde{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ is the symmetrically normalized adjacency matrix.

The generative model learns the latent variables by an inner product defined as

$$p(A|Z = \prod_{i=1}^{N}\prod_{i=1}^{N} p(A_{ij}|z_i, z_j), \tag{2-18}$$

where $p(A_{ij} = 1|z_i, z_j) = \sigma(z_i^\top z_j)$, A_{ij} are the elements of A and $\sigma(\cdot)$ is the logistic sigmoid function.

Finally, the framework defines a learning objective based on the variational lower bound of the variational parameter W_i. This formulation can be defined as

$$\mathcal{L}(p, q) = \mathbb{E}_{q(Z|A,X)}\left[\log P(A|Z)\right] - D_{KL}(q(Z|A, X) \parallel p(Z)) \tag{2-19}$$

where we can take the Gaussian prior $p(Z) = \prod_i \mathcal{N}(z_i|0, I)$.

2.3 Summary

This chapter presents the background for understanding unsupervised learning on graphs using generative models, specifically leading up to the variational graph autoencoders framework, the backbone of our proposed framework described in Chapter 4. We present the challenge of approximate Bayesian inference as an optimization problem. We explain how VI deals with

intractable posterior calculations by introducing a simple prior distribution and maximizing the evidence lower bound, equivalent to minimizing the KL divergence. The chapter notes potential VI generalizations using alternative divergence measures. The next chapter will describe a proposed representation for generalized divergence measures and novel divergences.

3
κ-Generalized Divergence Measures

This chapter presents:

1. A generalized representation of divergence measures called κ-divergences, which nests all of the f-divergences;

2. A comprehensive summary of generator functions for various κ-divergences found in literature, demonstrating their mathematical relationships and properties; and two novel families of divergences: δ-divergences and η-divergences, which we propose as specific instantiations within the κ-divergence framework;

3. A evaluation of κ-divergences in a distribution approximation problem for a mixture of Gaussians. This experiment aims to observe the behavior of the divergences in terms of mode-seeking and mass-covering.

3.1
Introduction

Divergence measures between distributions are crucial in machine learning. Some examples of its use are deep generative models based on variational inference (Song et al., 2020; Ho et al., 2020), generative adversarial networks (Nowozin et al., 2016), and mutual information neural estimation (Belghazi et al., 2018). Many of these applications use the class of f-divergence as the default. This class has well-established divergence measures with strong properties and empirical results, like D_{KL}, D_{α}, D_{JS} (Jensen-Shannon), D_{HD} (Hellinger distance), and D_{TV} (total variation).

In this work, we propose a representation for generalized divergence measures, including the non-additive divergence family based on Tsallis statistics (Tsallis et al., 1998; Tsallis and Cirto, 2013), f-divergences (Csiszár, 1967; Morimoto, 1963), and others.

3.2 A κ-Parameterized Generalization of Divergence Measures

The κ-divergence represents a possible generalization that unifies many classical generalized divergence measures through a parameterized approach (Figure 3.1). By modulating the parameter κ, we control the sensitivity of the divergence to different aspects of distributional differences - emphasizing either the tails, central mass, or specific regions of interest. The κ-divergence framework encompasses several well-known parametrized divergence measures while maintaining desirable mathematical properties such as convexity, monotonicity, and adherence to the data processing inequality.

This section formalizes the definition of κ-divergences, establishes their theoretical foundations, and explores their mathematical properties.

Definition 3.1 (κ-divergence) *Let P and Q be probability measures on a measurable space $\mathcal{X}$ with density functions p and q, respectively. The κ-divergence family is defined as:*

$$D_\kappa(P \parallel Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}; \kappa\right) dx, \quad \kappa \in [1, \infty) \tag{3-1}$$

where $f(\cdot; \kappa)$ is a convex function of the likelihood ratio $\phi(\mathbf{x}) = \frac{\mathbf{p(x)}}{\mathbf{q(x)}}$ that satisfies $f(1; \kappa) = 0$ for all $\kappa \in [1, \infty)$. The parameter κ modulates how the function f transforms the likelihood ratio, allowing different sensitivities to various regions of distribution discrepancy.

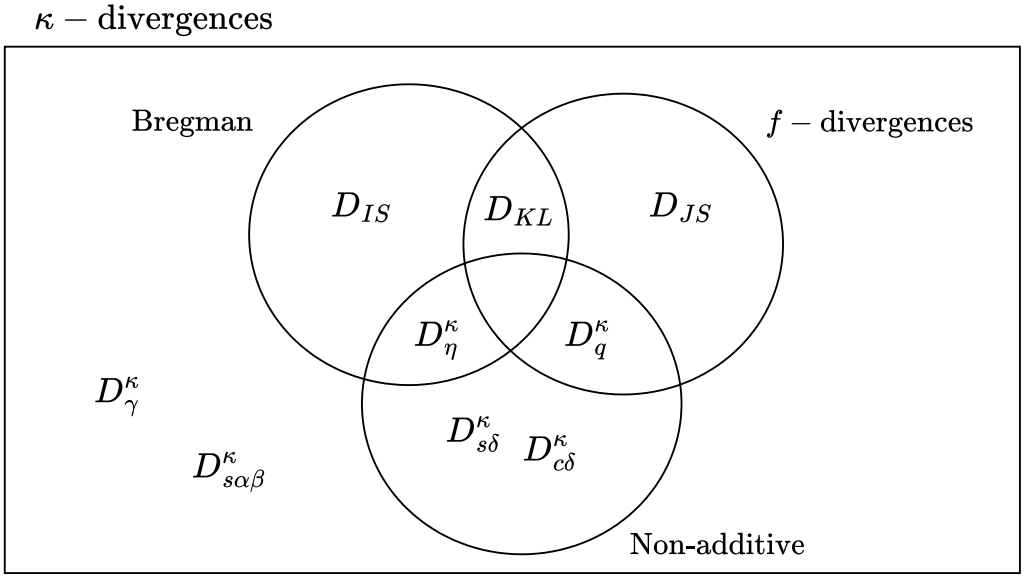

Figure 3.1: A simplified representation of κ-divergence space

Theorem 3.2 (Non-negativity) *For any probability measures P and Q and any $\kappa \in [1, \infty)$, $D_\kappa(P \parallel Q) \geq 0$ with equality if and only if $P = Q$ almost everywhere.*

Proof. Since $f(\cdot; \kappa)$ is convex for a given κ and $f(1; \kappa) = 0$, by Jensen's inequality:

$$D_\kappa(P \parallel Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}; \kappa\right) dx \tag{3-2}$$

$$\geq f\left(\int_{\mathcal{X}} q(x) \cdot \frac{p(x)}{q(x)} dx; \kappa\right) \tag{3-3}$$

$$= f\left(\int_{\mathcal{X}} \frac{p(x)q(x)}{q(x)} dx; \kappa\right) \tag{3-4}$$

$$= f\left(\int_{\mathcal{X}} p(x) dx; \kappa\right) \tag{3-5}$$

$$= f\left(1; \kappa\right) \tag{3-6}$$

Since f is convex with $f(1; \kappa) = 0$, we have $D_\kappa(P \parallel Q) \geq 0$ with equality if and only if $P = Q$ almost everywhere. ■

The literature presents some divergences that follow our definition of a κ-divergence. Table 3.1 summarizes a subset of these functions for a more detailed evaluation of their convex behavior. The divergences choose are: q-divergence (Tsallis et al., 1998), Rényi divergence (Rényi, 1961), α-divergence (Amari, 1985), β-divergence (Basu et al., 1998), γ-divergence (Fujisawa and Eguchi, 2008), and sAB-divergence (Cichocki et al., 2011). The last three divergences - $c\delta$, $s\delta$, and η divergences - are proposed in this work and described in Section 3.3.

First, we write the generator function of these divergences using a standard κ parameter.

Table 3.1: Overview of κ generalized divergences generator functions. For each divergence, the generator function $f(\phi, \kappa)$ and the first derivative of the generator function $f'(\phi, \kappa)$

Divergence	$f(\phi, \kappa)$	$f'(\phi, \kappa)$
Tsallis (q-divergence)	$\frac{\phi^\kappa - \phi}{\kappa - 1}$	$\frac{\frac{\kappa\phi^\kappa}{\phi} - 1}{\kappa - 1}$
Rényi	$\frac{\phi^\kappa - 1}{\kappa - 1}$	$\frac{\kappa\phi^\kappa}{\phi(\kappa - 1)}$
α-divergence	$\frac{\phi^\kappa - \kappa\phi + \kappa - 1}{\kappa(\kappa - 1)}$	$\frac{-\kappa + \frac{\kappa\phi^\kappa}{\phi}}{\kappa(\kappa - 1)}$
β-divergence	$\phi^{(1-\kappa)} + (\kappa - 1)\phi$	$\kappa - 1 + \frac{\phi^{(1-\kappa)}(1-\kappa)}{\phi}$
γ-divergence	$\frac{\phi^\kappa - \phi}{\kappa(\kappa - 1)}$	$\frac{\kappa\phi^{(\kappa-1)} - 1}{\kappa(\kappa - 1)}$
sAB-divergence	$\frac{\phi^{\kappa_1+\kappa_2}}{\kappa_2(\kappa_1+\kappa_2)} + \frac{1}{\kappa_1(\kappa_1+\kappa_2)} - \frac{\phi^{\kappa_1}}{\kappa_1\kappa_2}$	$\frac{(\kappa_1+\kappa_2)\phi^{\kappa_1+\kappa_2-1}}{\kappa_2(\kappa_1+\kappa_2)} - \frac{\kappa_1\phi^{\kappa_1-1}}{\kappa_1\kappa_2}$
η-divergence	$\phi^\kappa - \log(\phi) - 1$	$\kappa\phi^{\kappa-1} - \frac{1}{\phi}$
$c\delta$-divergence	$\phi[\log(c + \phi)]^\kappa$	$\frac{\kappa\phi\log(c+\phi)^{\kappa-1}}{c+\phi} + \log(c + \phi)^\kappa$
$s\delta$-divergence	$\phi([\log(\phi)]^2)^\kappa$	$\Big(\log\phi\Big)^{2\kappa-1}\Big(\log\phi + 2\kappa\Big)$

Figure 3.2 shows how the generator function value varies with ϕ for different κ values (ranging from 1.1 to 3.0, indicated by the color gradient). The figure illustrates the characteristic behaviors of each divergence family, including minima positions, growth rates, and asymptotic properties. Lighter blue curves represent smaller κ values, while darker blue curves indicate larger values, as shown in the color bar. All divergences exhibit unique functional forms while maintaining the common property of minimum value near $\phi = 1$.

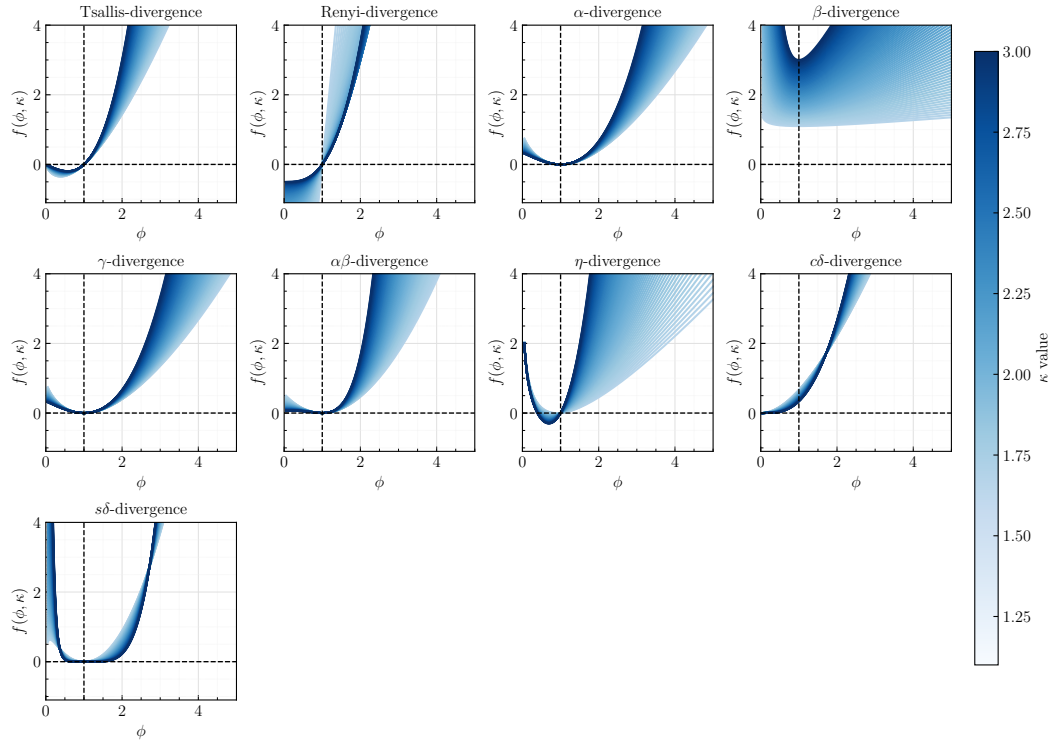

Figure 3.2: Visualization of κ-divergence generator functions across varying parameter values

3.3 Introducing the δ and η Divergences

While many divergence measures, like the D_{KL}, have been extensively studied, they often suffer from limitations such as sensitivity to outliers, lack of scale invariance, or poor behavior when distributions have limited overlap. This section introduces two novel families of divergence measures - δ and η divergences - each designed to address specific analytical challenges while maintaining desirable mathematical properties.

These divergence families offer varying degrees of flexibility through their parameterization, allowing practitioners to tailor the measures to specific applications. The $c\delta$-divergence provides robustness through its adjustable constant parameter, making it particularly suitable for comparing distributions with regions of low density. The $s\delta$-divergence, focusing on squared logarithmic differences, offers enhanced discrimination for distributions with similar means but different variances. Meanwhile, the η-divergence generalizes the scale-invariant properties of the Itakura-Saito distance Itakura (1968).

Following this, we formally define each divergence measure and establish its key mathematical properties.

3.3.1
δ-entropy and a possible δ-divergence

A non-additive entropic functional specialized for black holes is the δ-entropy (Tsallis, 2009; Tsallis and Cirto, 2013). Below, we briefly describe δ-entropy and propose a direct corresponding δ-divergence function.

Definition 3.3 (δ-entropy) *The δ-entropy $h_\delta(X)$ of a continuous random variable X with probability density function $p(x)$ is defined as*

$$h_\delta(X) = -\int_{\mathcal{X}} p(x) \left[\log p(x)\right]^\delta, \quad \delta > 0 \tag{3-7}$$

where $\mathcal{X}$ is the support set of X.

Definition 3.4 (δ-divergence) *Let P and Q be two distributions with probability density functions $p(x)$ and $q(x)$. For $\ln_\delta(x) \equiv [\log(x)]^\delta$, the δ-divergence with $\delta \in \mathbb{N} \setminus \{0\}$ is given by*

$$\begin{aligned} D_\delta(P \parallel Q) = &\int_{\mathcal{X}} p(x) \left[\log \tfrac{p(x)}{q(x)}\right]^\delta dx \\ = &\int_{\mathcal{X}} p(x) \ln_\delta \left(\tfrac{p(x)}{q(x)}\right) dx \end{aligned} \tag{3-8}$$

where the integral is taken over regions where $\frac{p(x)}{q(x)} > 0$.

Lemma 3.5 *δ-divergence satisfies the following properties:*

- ***Non-negativity:*** *For $\delta \geq 1$, $D_\delta(P \parallel Q) \geq 0$ with equality if and only if $P = Q$ almost everywhere.*
- ***Generality:*** *As $\delta \to 1$, δ-divergence reduces to the D_{KL}:*

$$\lim_{\delta \to 1} D_\delta(P \parallel Q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx = D_{KL}(P \parallel Q). \tag{3-9}$$

Proof.

1. **Non-negativity:** For $\delta \geq 1$, the function $f(t) = [\log(t)]^\delta$ is convex for $t > 0$. By Jensen's inequality, we have

$$\mathbb{E}_P \left[\ln_\delta \left(\frac{p(X)}{q(X)}\right)\right] \geq \ln_\delta \left(\mathbb{E}_P \left[\frac{p(X)}{q(X)}\right]\right) \tag{3-10}$$

   Since $\mathbb{E}_P \left[\frac{p(X)}{q(X)}\right] = \int_{\mathcal{X}} p(x) \cdot \frac{p(x)}{q(x)} dx = \int_{\mathcal{X}} \frac{p^2(x)}{q(x)} dx \geq \left(\int_{\mathcal{X}} p(x) dx\right)^2 = 1$, we have $\ln_\delta(1) = 0$.

   Therefore, $D_\delta(P \parallel Q) = \mathbb{E}_P \left[\ln_\delta \left(\frac{p(X)}{q(X)}\right)\right] \geq 0$ with equality if and only if $\frac{p(x)}{q(x)} = 1$ almost everywhere, which implies $P = Q$ almost everywhere.

2. **Generality (Limit to D_{KL}):** Taking the limit as $\delta \to 1$:

$$\lim_{\delta \to 1} D_\delta(P \parallel Q) = \lim_{\delta \to 1} \int_{\mathcal{X}} p(x) \left[\log \frac{p(x)}{q(x)}\right]^\delta dx \tag{3-11}$$

$$= \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx \tag{3-12}$$

$$= D_{KL}(P \parallel Q) \tag{3-13}$$

■

3.3.1.1 Limitations of δ-divergence

The δ-divergence exhibits several fundamental limitations that constrain its practical applicability. First, when the likelihood ratio $\frac{p(x)}{q(x)} < 1$, the term $[\log(\frac{p(x)}{q(x)})]^\delta$ produces complex-valued results for any non-integer real values of δ (where $\delta \in \mathbb{R} \setminus \mathbb{N}$). These complex outputs significantly compromise the interpretability of δ-divergence as a statistical distance measure.

Second, for δ-divergence to be properly classified within the class of κ-divergences, the generating function $f(t, \kappa) = [\log(t)]^\kappa$ must satisfy necessary convexity conditions.

Theorem 3.6 (Convexity of δ-divergence) *The δ-divergence with generating function $f(t, \kappa) = [\log(t)]^\kappa$ satisfies the convexity requirements of a proper divergence if and only if $\delta \in \mathbb{N}$.*

Proof. For $D_\delta(P \parallel Q)$ to qualify as a proper κ-divergence, its generating function $f(t, \kappa) = [\log(t)]^\kappa$ must be convex on $\mathbb{R}_+$. This constraint necessitates the second derivative $f''(t) \geq 0$ for all $t > 0$.

We compute the first derivative:

$$f'(t) = \kappa[\log(t)]^{\kappa-1} \cdot \frac{1}{t} \tag{3-14}$$

The second derivative is given by:

$$f''(t) = \kappa(\kappa - 1)[\log(t)]^{\kappa-2} \cdot \frac{1}{t^2} - \kappa[\log(t)]^{\kappa-1} \cdot \frac{1}{t^2} \tag{3-15}$$

$$= \frac{\kappa}{t^2} \left[(\kappa - 1)[\log(t)]^{\kappa-2} - [\log(t)]^{\kappa-1}\right] \tag{3-16}$$

We analyze the convexity in two regions:

Case 1: For $t > 1$, we have $\log(t) > 0$. The convexity condition requires:

$$(\kappa - 1)[\log(t)]^{\kappa-2} - [\log(t)]^{\kappa-1} \geq 0 \tag{3-17}$$

$$(\kappa - 1) - \log(t) \geq 0 \tag{3-18}$$

$$\log(t) \leq \kappa - 1 \tag{3-19}$$

This inequality cannot be satisfied for all $t > 1$ unless $\kappa - 1 = \infty$, which is impossible for finite κ.

Case 2: For $0 < t < 1$, we have $\log(t) < 0$. The analysis differs based on the value of κ:

- When $\kappa \in \mathbb{N}$ (positive integers):
  - The function $[\log(t)]^{\kappa}$ is well-defined for all $t > 0$
  - For even κ, $f''(t) > 0$ for all $t \neq 1$
  - For odd $\kappa \geq 3$, convexity holds on specific intervals
- When $\kappa \in \mathbb{R}_+ \setminus \mathbb{N}$ (positive non-integers):
  - The function $[\log(t)]^{\kappa}$ yields complex values when $t < 1$
  - Even when restricted to domains where $t > 1$, the function fails to satisfy global convexity conditions

Therefore, the δ-divergence satisfies the convexity requirements of a proper κ-divergence if and only if $\kappa \in \mathbb{N}$. ■

Corollary 3.7 *The parameter space of the δ-divergence as a proper statistical distance measure is restricted to the discrete set of positive integers rather than the continuous domain of positive real numbers.*

Third, the numerical computation of δ-divergence encounters significant instability in regions where $p(x)$ or $q(x)$ approaches zero. This computational instability is exacerbated with increasing values of δ, as higher powers amplify the effect of near-zero probabilities in the likelihood ratio.

To address these limitations, we propose two variations of the δ-divergence: (1) the $c\delta$-divergence, which incorporates a stability parameter c to regularize the computation, and (2) the $s\delta$-divergence, which employs a squared likelihood ratio to mitigate the analytical and computational challenges.

3.3.2 $c\delta$-divergence

Definition 3.8 ($c\delta$-divergence) *The adjusted δ-divergence with parameter $\delta \in \mathbb{R}$ and constant $c \geq 1$ is defined as*

$$D_\delta^{c\delta}(P \parallel Q) = \int_{\mathcal{X}} p(x) \left[\log \left(c + \frac{p(x)}{q(x)} \right) \right]^\delta dx \tag{3-20}$$

where p and q are the density functions of non-negative measures P and Q, respectively, defined on a measurable space $\mathcal{X}$.

Theorem 3.9 (Non-negativity) *For any non-negative measures P and Q defined on $\mathcal{X}$, any $\delta > 0$, and constant $c > 0$, the adjusted δ-divergence satisfies $D_\delta^{c\delta}(P \parallel Q) \geq 0$, with equality if and only if $P = Q$ almost everywhere.*

Property 1 (Asymmetry) *In general,*

$$D_\delta^{c\delta}(P \parallel Q) \neq D_\delta^{c\delta}(Q \parallel P) \tag{3-21}$$

Property 2 (Relation to D_{KL}) *When $\delta = 1$ and $c = 0$, the adjusted δ-divergence reduces to the D_{KL}:*

$$D_1^{c\delta}(P \parallel Q)|_{c=0} = \int_{\mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right) dx = D_{KL}(P \parallel Q) \tag{3-22}$$

Lemma 3.10 (Convexity in First Argument) *For $\delta \geq 1$ and $c \in \mathbb{N} \setminus \{0\}$, the function $P \mapsto D_\delta^{c\delta}(P \parallel Q)$ is convex for any fixed Q.*

Property 3 (Robustness with $c > 0$) *The presence of the constant $c > 0$ in the logarithm makes the $c\delta$-divergence more robust to situations where $\frac{p(x)}{q(x)}$ approaches zero, as compared to standard δ-divergence.*

3.3.3 $s\delta$-divergence

Definition 3.11 ($s\delta$-divergence) *The squared δ-divergence with parameter $\delta \in \mathbb{R}$ is defined as*

$$D_\delta^{s\delta}(P \parallel Q) = \int_{\mathcal{X}} p(x) \left(\left[\log \left(\frac{p(x)}{q(x)} \right) \right]^2 \right)^\delta dx \tag{3-23}$$

where p and q are the density functions of non-negative measures P and Q, respectively, defined on a measurable space $\mathcal{X}$.

Theorem 3.12 (Non-negativity) *For any non-negative measures P and Q defined on $\mathcal{X}$ and any $\delta > 0$, the squared δ-divergence satisfies $D_\delta^{s\delta}(P \parallel Q) \geq 0$, with equality if and only if $P = Q$ almost everywhere.*

Proof. For any x where $p(x) > 0$ and $q(x) > 0$, the term $\log\left(\frac{p(x)}{q(x)}\right)$ equals zero if and only if $\frac{p(x)}{q(x)} = 1$, which means $p(x) = q(x)$.

Since the square of a non-zero real number is always positive, $\left[\log\left(\frac{p(x)}{q(x)}\right)\right]^2 > 0$ whenever $p(x) \neq q(x)$.

For $\delta > 0$, raising a positive number to the power of δ preserves positivity. Moreover, multiplying by $p(x) > 0$ maintains the sign.

Therefore, the integrand is non-negative everywhere and strictly positive where $p(x) \neq q(x)$. This implies $D_\delta^{s\delta}(P \parallel Q) \geq 0$, with equality if and only if $P = Q$ almost everywhere. ■

Property 4 (Symmetry) *For all values of δ, the squared δ-divergence is symmetric:*

$$D_\delta^{s\delta}(P \parallel Q) = D_\delta^{s\delta}(Q \parallel P) \tag{3-24}$$

when defined with appropriate reference measures.

Property 5 (Relation to Rényi's Divergence) *When $\delta = 1$, the squared δ-divergence relates to the second-order entropy of the log-likelihood ratio:*

$$D_1^{s\delta}(P \parallel Q) = \int_{\mathcal{X}} p(x) \left[\log\left(\frac{p(x)}{q(x)}\right)\right]^2 dx \tag{3-25}$$

which corresponds to the variance of the log-likelihood ratio under distribution P.

Lemma 3.13 (Convexity in First Argument) *For $\delta \geq 0$, the function $P \mapsto D_\delta^{s\delta}(P \parallel Q)$ is convex for any fixed Q.*

Property 6 (Strong Discrimination) *Due to the squared logarithm, the $s\delta$-divergence can provide more substantial discrimination between distributions with similar means but different variances compared to standard divergences based on first-order log-likelihood ratios.*

3.3.4 η-divergence

Definition 3.14 (η-divergence) *Let P and Q be non-negative measures defined on a measurable space $\mathcal{X}$. The η-divergence (or generalized Itakura-Saito distance) with parameter $\eta \in \mathbb{R}$ is defined as:*

$$D_\eta(P \parallel Q) = \int_{\mathcal{X}} \left(\frac{p(x)^\eta}{q(x)^\eta} - \log\left(\frac{p(x)}{q(x)}\right) - 1\right) dx \tag{3-26}$$

where p and q are the density functions of P and Q, respectively. For convention, we will consider $\eta = \kappa$.

Theorem 3.15 (Non-negativity) *For any non-negative measures P and Q defined on $\mathcal{X}$ and any $\eta \in \mathbb{R}$, the generalized Itakura-Saito divergence satisfies $D_\eta(P \parallel Q) \geq 0$, with equality if and only if $P = Q$ almost everywhere.*

Proof. Let $f(t) = t^\kappa - \log(t) - 1$ for $t > 0$. Taking the derivative, we get $f'(t) = \kappa t^{\kappa-1} - \frac{1}{t}$. Setting $f'(t) = 0$ yields $\kappa t^{\kappa-1} = \frac{1}{t}$, thus $\kappa t^\kappa = 1$, which gives $t = \left(\frac{1}{\kappa}\right)^{\frac{1}{\kappa}}$.

For the second derivative, $f''(t) = \kappa(\kappa-1)t^{\kappa-2} + \frac{1}{t^2}$. For $\kappa \geq 1$ or $\kappa < 0$, we have $f''(t) > 0$ for all $t > 0$. For $0 < \kappa < 1$, we have $\kappa(\kappa-1)t^{\kappa-2} < 0$, but $\frac{1}{t^2}$ dominates for sufficiently small or large values of t, ensuring $f(t)$ remains strictly convex.

Since $f(t)$ is strictly convex and attains its minimum at $t = \left(\frac{1}{\kappa}\right)^{\frac{1}{\kappa}}$, we have $f(t) \geq f\left(\left(\frac{1}{\kappa}\right)^{\frac{1}{\kappa}}\right)$ for all $t > 0$, with equality if and only if $t = \left(\frac{1}{\kappa}\right)^{\frac{1}{\kappa}}$.

Setting $t = \frac{p(x)}{q(x)}$ and integrating, we get:

$$D_\kappa^\eta(P \parallel Q) = \int_{\mathcal{X}} f\left(\frac{p(x)}{q(x)}\right) q(x)dx \tag{3-27}$$

$$\geq \int_{\mathcal{X}} f\left(\left(\frac{1}{\kappa}\right)^{\frac{1}{\kappa}}\right) q(x)dx \tag{3-28}$$

$$= f\left(\left(\frac{1}{\kappa}\right)^{\frac{1}{\kappa}}\right) \int_{\mathcal{X}} q(x)dx \tag{3-29}$$

For normalized distributions (i.e., probability measures), $\int_{\mathcal{X}} q(x)dx = 1$. Now, observe that $f(1) = 1^\kappa - \log(1) - 1 = 1 - 0 - 1 = 0$. Since f attains its minimum at $t = \left(\frac{1}{\kappa}\right)^{\frac{1}{\kappa}}$ and not at $t = 1$ (except when $\kappa = 1$), we have $f\left(\left(\frac{1}{\kappa}\right)^{\frac{1}{\kappa}}\right) < 0$.

However, equality in $D_\kappa(P \parallel Q) = 0$ is achieved if and only if $\frac{p(x)}{q(x)} = 1$ for almost all x, which means $P = Q$ almost everywhere. ■

Property 7 (Asymmetry) *In general, for $\kappa \neq 1$,*

$$D_\kappa(P \parallel Q) \neq D_\kappa(Q \parallel P) \tag{3-30}$$

Property 8 (Special Case) *When $\kappa = 1$, the generalized divergence reduces to the standard Itakura-Saito divergence:*

$$D_1^\eta(P \parallel Q) = \int_{\mathcal{X}} \left(\frac{p(x)}{q(x)} - \log\left(\frac{p(x)}{q(x)}\right) - 1\right) dx \tag{3-31}$$

Lemma 3.16 (Convexity in First Argument) *For $\kappa \geq 1$, the function $P \mapsto D_\kappa(P \parallel Q)$ is convex for any fixed Q.*

Proof. For $\kappa \geq 1$, the function $g(x) = \frac{x^\kappa}{q^\kappa}$ is convex in x for any fixed $q > 0$. This can be verified by taking the second derivative:

$$\frac{d^2}{dx^2} g(x) = \kappa(\kappa - 1)\frac{x^{\kappa-2}}{q^\kappa} \geq 0 \text{ for } \kappa \geq 1 \tag{3-32}$$

The function $h(x) = -\log\left(\frac{x}{q}\right) - 1$ is also convex in x, as its second derivative is $\frac{d^2}{dx^2} h(x) = \frac{1}{x^2} > 0$.

Since the sum of convex functions is convex, $g(x) + h(x)$ is convex, which implies that $D_\kappa(P \parallel Q)$ is convex in P. ■

Lemma 3.17 (Convexity in Second Argument) *For $0 \leq \kappa \leq 1$, the function $Q \mapsto D_\kappa(P \parallel Q)$ is convex for any fixed P.*

Proof. We need to analyze the convexity of the function $j(q) = \frac{p^\kappa}{q^\kappa} - \log\left(\frac{p}{q}\right) - 1$ with respect to q for fixed p.

Taking the first derivative:

$$\frac{d}{dq} j(q) = -\kappa\frac{p^\kappa}{q^{\kappa+1}} + \frac{1}{q} \tag{3-33}$$

And the second derivative:

$$\frac{d^2}{dq^2} j(q) = \kappa(\kappa + 1)\frac{p^\kappa}{q^{\kappa+2}} - \frac{1}{q^2} \tag{3-34}$$

For this to be non-negative (ensuring convexity), we need:

$$\kappa(\kappa + 1)\frac{p^\kappa}{q^{\kappa+2}} \geq \frac{1}{q^2} \tag{3-35}$$

Simplifying:

$$\kappa(\kappa + 1)\frac{p^\kappa}{q^\kappa} \geq 1 \tag{3-36}$$

This condition is satisfied for all $p, q > 0$ when $0 \leq \kappa \leq 1$, which establishes the convexity of $D_\kappa(P \parallel Q)$ in the second argument Q for this range of κ. ■

Theorem 3.18 (Generalized Scale Relation) *For any scalar $\alpha > 0$ and non-negative measures P and Q:*

$$D_\kappa(\alpha P \parallel \alpha Q) = D_\kappa(P \parallel Q) \tag{3-37}$$

Proof.

$$D_\kappa(\alpha P \parallel \alpha Q) = \int_{\mathcal{X}} \left(\frac{(\alpha p(x))^\kappa}{(\alpha q(x))^\kappa} - \log\left(\frac{\alpha p(x)}{\alpha q(x)}\right) - 1 \right) dx \tag{3-38}$$

$$= \int_{\mathcal{X}} \left(\frac{\alpha^\kappa p(x)^\kappa}{\alpha^\kappa q(x)^\kappa} - \log\left(\frac{p(x)}{q(x)}\right) - 1 \right) dx \tag{3-39}$$

$$= \int_{\mathcal{X}} \left(\frac{p(x)^\kappa}{q(x)^\kappa} - \log\left(\frac{p(x)}{q(x)}\right) - 1 \right) dx \tag{3-40}$$

$$= D_\kappa(P \parallel Q) \tag{3-41}$$

This computation shows that the generalized Itakura-Saito divergence is scale-invariant for all values of κ. ■

Property 9 (Relation to Bregman Distance) *The generalized Itakura-Saito divergence is a Bregman distance generated by the function $f(x) = x^\kappa - \log(x) - 1$, satisfying:*

$$D_\kappa(P \parallel Q) = f(P) - f(Q) - \langle \nabla f(Q), P - Q \rangle \tag{3-42}$$

where ∇f denotes the gradient of f and $\langle \cdot, \cdot \rangle$ represents the appropriate inner product.

Corollary 3.19 (Scale Invariance) *The generalized Itakura-Saito divergence is scale-invariant for all values of κ:*

$$D_\kappa(\alpha P \parallel \alpha Q) = D_\kappa(P \parallel Q) \tag{3-43}$$

for any scalar $\alpha > 0$.

Proof. This follows directly from the Generalized Scale Relation theorem. ■

3.4 Experiment I: Exploring the Behavior of κ-divergences

3.4.1 Problem setup

We explore our three proposed divergences - $c\delta$, $s\delta$, and η - within a variational inference framework to approximate a target probability distribution. The goal is to find the optimal parameters $\theta = (\mu, \sigma)$ for a distribution $q_\theta(x)$ from a prior[1], such that $q_\theta(x)$ closely approximates a target distribution $p(x)$

[1]For all experiments, the approximating distribution $q_\theta(x)$ is chosen from the family of Gaussian distributions $q_\theta(x) = \mathcal{N}(x|\mu, \sigma^2)$

by minimizing a divergence $D_\kappa(p||q_\theta)$[2]. This experimental setup focuses on a one-dimensional problem with a known target distribution $p(x)$ defined as a mixture of two Gaussian distributions with probability density function as,

$$p(x) = w_1\mathcal{N}(x|\mu_1, \sigma_1^2) + w_2\mathcal{N}(x|\mu_2, \sigma_2^2) \tag{3-44}$$

where $\mathcal{N}(x|\mu, \sigma^2)$ denotes the probability density function of a Gaussian distribution with mean μ and variance σ^2. The specific parameters used are mixture weights $w_1 = 0.7$ and $w_2 = 0.3$, component means $\mu_1 = -1.0$ and $\mu_2 = 1.5$, and component standard deviations $\sigma_1 = 0.5$ and $\sigma_2 = 0.5$.

> Minka (2005) defines that "*If two identical Gaussians are separated enough, an exclusive divergence prefers to represent only one of them, while an inclusive divergence prefers to stretch across both.*".

We use these two concepts to define the behavior of divergence measures considering an approximate inference problem (as described in Chapter 2). The first concept is the mode-seeking (or exclusive divergence). A specific divergence measure is defined as mode-seeking if it tends to represent only the mode with the highest mass. The second concept is the mass-covering (or inclusive divergence). In this case, the divergence tends to cover as much of the distribution p as possible.

3.4.2 Mixture of Gaussians approximation with δ and η divergences

The Figures 3.3, 3.4, and 3.5 present the results of the approximation of a mixture of Gaussians p using a simple distribution q for $s\delta$, $c\delta$, and η, respectively. The results show that all three divergences present a mass-covering behavior. As empirically observed by Poole et al. (2016), mass-covering divergences can increase sample diversity without losing sample quality. This behavior can prevent "mode collapse" and potentially introduce lower-probability samples.

[2]The optimization objective minimizes a selected κ-divergence measure between $p(x)$ and $q_\theta(x)$. The distributions are evaluated over a discrete grid of 1000 points uniformly spaced between -4 and 4. Let the resulting density vectors be $\mathbf{p}$ and $\mathbf{q}_\theta$. Before calculating the divergence, both vectors are normalized to sum to one, effectively treating them as probability mass functions over the discrete grid. We use Adam optimizer with a learning rate of 0.01 over 2000 epochs.

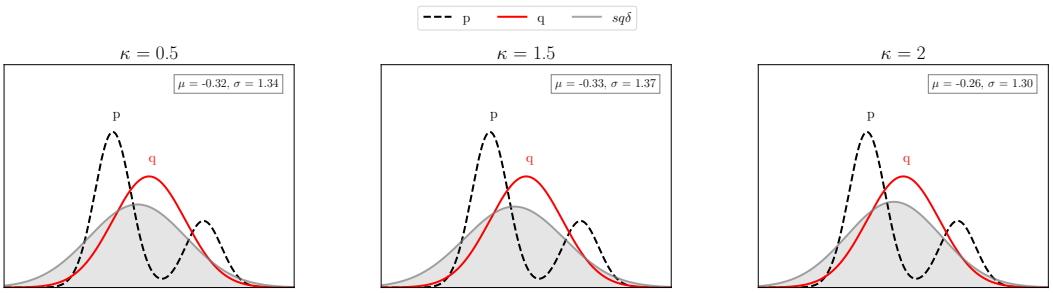

Figure 3.3: The Gaussian (gray area) which minimizes $s\delta$-divergence to p (a mixture of two Gaussians) considering a prior q, for varying $\kappa = \{0.5, 1.5, 2\}$

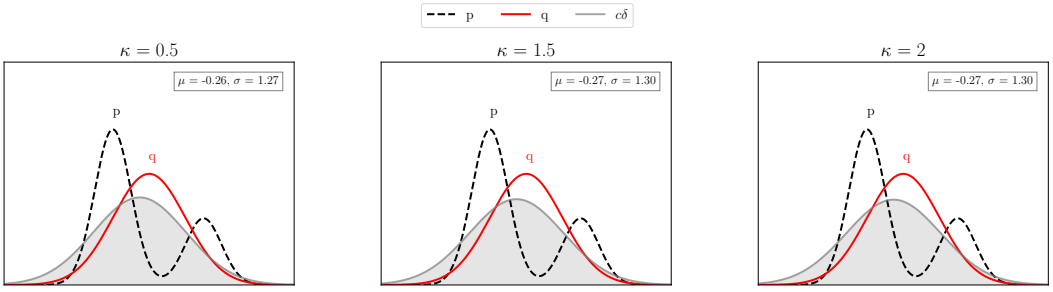

Figure 3.4: The Gaussian (gray area) which minimizes $c\delta$-divergence to p (a mixture of two Gaussians) considering a prior q, for varying $\kappa = \{0.5, 1.5, 2\}$

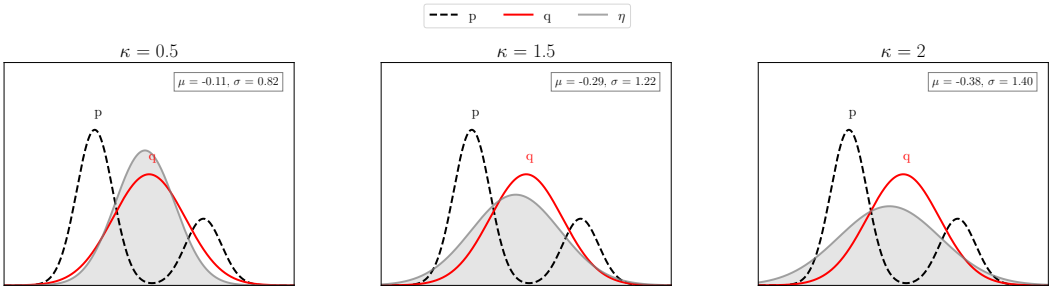

Figure 3.5: The Gaussian (gray area) which minimizes η-divergence to p (a mixture of two Gaussians) considering a prior q, for varying $\kappa = \{0.5, 1.5, 2\}$

3.5 Summary

This chapter presented a new class of divergences - the κ-divergences. This class can be observed as a generalization of f-divergence with more flexible convex functions that can be effectively used in learning scenarios. We summarize some of the parameterized divergences found in the literature that are cases of κ-divergence and propose two novel families of divergences: δ and η. Analyzing the behavior of these novel divergences, we observe a dominance of mass-covering behavior.

4
Generalized Variational Graph Auto-Encoders

This chapter presents:

1. A generalization of the variational graph autoencoder (VGAE) model. In this generalization, we employ the three components of a VGAE (inference model, generative model, and learning objective), but our objective is a generalized approach based on the κ-divergences presented in Chapter 3;

2. A sensitive analysis based on the parameter κ for link prediction and node classification tasks. We look for the optimal κ parameter value for each κ-divergence measure applied in GGVA;

3. Experiments: in Experiment II, we evaluate the GGVA framework in link prediction and node classification in five academic citation networks; in Experiment III, we propose a link prediction application for GGVA framework in a novel real-world dataset (the Brazilian power grid system);

4.1
Introduction

We propose a generalized graph variational autoencoders (GGVA) that extends the traditional VGAE framework by incorporating κ-divergences as a learning objective. Our model employs a GCN encoder architecture to learn latent representations of graph-structured data while utilizing various divergence measures to capture the information loss between the approximate posterior and prior distributions.

4.2
A *Generalized* Latent Variable Model for Graph-structured Data

Graphs present unique challenges for representation learning due to the complex interdependencies between nodes and the heterogeneity of connectivity patterns. The GGVA framework addresses these challenges by formulating a generalized latent variable model that captures local and global structural properties while maintaining computational tractability.

Given an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with node features $X \in \mathbb{R}^{|V| \times d}$, where d is the feature dimension, we aim to learn a low-dimensional latent representation $Z \in \mathbb{R}^{|V| \times h}$. The generative process assumes that the observed graph structure arises from interactions between latent node representations. Specifically, the probability of an edge existing between nodes i and j is modeled as $p(A_{ij} = 1 | z_i, z_j) = f_\theta(z_i, z_j)$, where f_θ is a parameterized similarity function and A is the adjacency matrix.

The GGVA model differs from standard graph VAEs by employing a more flexible regularization scheme based on a family of κ-divergences parameterized by $\kappa > 1$ (see details in Chapter 3). We argue this generalization allows for more nuanced control over the trade-off between reconstruction accuracy and latent space regularity, adapting to the specific characteristics of the input graph.

The overall structure of our approach is similar to a traditional VGAE implementation with three learning components: an inference model, a generative model, and a learning objective based on an evidence lower bound. Next, we briefly describe these components in the context of GGVA.

4.2.1 Three learning components

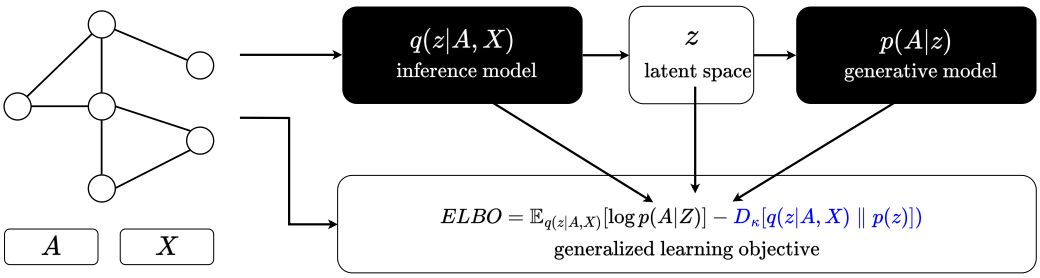

Figure 4.1: The Generalized Graph Variational Autoencoder (GGVA) framework.

In summary, our method consists of three key components: (1) an inference model $q(Z|A, X)$ encoding graph structure and node features, (2) a generative model $p(A|Z)$ that reconstructs graph adjacency from latent representations through a decoder, and (3) a learning objective $\mathcal{L}(p, q)$ combining positive and negative reconstruction terms with a divergence regularize.

The inference model in GGVA, denoted as $q(Z|X, A)$, approximates the true posterior distribution over latent variables, given the observed graph structure and node features. This model is implemented as an encoder architecture that leverages GCN architecture (Kipf and Welling, 2016a). Considering a two-layer GCN that parametrizes the inference model in GGVA is,

$$q(Z|A,X) = \prod_{i=1}^{N} q(z_i|A,X) \tag{4-1}$$

where $q(z_i|A,X) = \mathcal{N}(z_i|\mu_i, diag(\sigma_i^2))$, $\mu = GCN_\mu(A,X)$ is the matrix of mean vector μ_i, and $\log\sigma = GCN_\sigma(A,X)$. The two layer GCN is defined as $GCN(X,A) = \tilde{A}\text{ReLU}(\tilde{A}XW_0W_1)$ with weight matrices W_i.

The generative model in GGVA, denoted as $p(A|Z)$, defines the probability distribution over graph structures given the latent representations. As in traditional VGAE architecture, we employ a simple inner product decoder.

$$p(A|Z = \prod_{i=1}^{N}\prod_{i=1}^{N} p(A_{ij}|z_i,z_j), \tag{4-2}$$

where $p(A_{ij} = 1|z_i,z_j) = \sigma(z_i^\top z_j)$, A_{ij} are the elements of A and $\sigma(\cdot)$ is the logistic sigmoid function.

The learning objective of GGVA extends the traditional variational lower bound by replacing the D_{KL} term with a more general κ-divergence regularization. This modification provides greater flexibility in balancing reconstruction accuracy and distribution matching, allowing the model to adapt to the specific characteristics of the input graph.

The overall loss function can be defined as follows:

$$\mathcal{L}(p,q) = \mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{neg}} + \lambda\mathcal{L}_{\text{div}}. \tag{4-3}$$

The first component, $\mathcal{L}_{\text{pos}}$, is the negative log-likelihood for positive edges (edges that exist in the graph):

$$\mathcal{L}_{\text{pos}} = -\mathbb{E}_{q(Z|X,A)}\left[\sum_{(i,j)\in E}\log p(A_{ij}=1|z_i,z_j)\right]. \tag{4-4}$$

The second component, $\mathcal{L}_{\text{neg}}$, is the negative log-likelihood for negative edges (edges that do not exist in the graph), sampled using a structure-aware negative sampling strategy:

$$\mathcal{L}_{\text{neg}} = -\mathbb{E}_{q(Z|X,A)}\left[\sum_{(i,j)\in\mathcal{N}}\log(1-p_\theta(A_{ij}=1|z_i,z_j))\right], \tag{4-5}$$

where $\mathcal{N}$ is the set of sampled negative edges.

The third component, $\mathcal{L}_{\text{div}}$, is the divergence regularization term, which depends on the chosen κ-divergence:

$$\mathcal{L}_{\text{div}} = D_\kappa(q(Z|X,A)\|p(Z)). \tag{4-6}$$

The parameter λ controls the strength of the regularization, balancing the trade-off between reconstruction accuracy and distribution matching. The choice of κ-divergence and its parameter κ significantly influences the model's behavior. Lower values of κ (closer to 1) tend to produce more mass-covering

behavior, capturing a broader range of the data distribution. Higher values of κ lead to more mode-seeking behavior, focusing on high-density regions of the data distribution.

The reparametrization trick ($Z = \mu + \sigma \odot \epsilon, \epsilon \sim \mathcal{N}(0, 1)$) is an alternative method for generating samples from the encoder $q(z|x)$ (Kingma and Welling, 2013) for training via stochastic gradient descent. Let z be a continuous random variable, and $z \sim q(z|A, X)$ be some conditional distribution. It is possible to express the z as a determinist variable $z = g(\epsilon, x)$, where ϵ is an auxiliary variable with independent marginal $p(\epsilon)$ and $g_\theta(\cdot)$ is a vector-valued function parameterized by θ.

4.3 Experiment II: Benchmark Learning on Graphs

4.3.1 Problem setup

We evaluate the GGVA framework[1] based on five benchmark datasets for link prediction and node classification (Table 4.1) (Yang et al., 2016). Each data set consists of academic citation networks as nodes and citation relationships as edges, and the features are a bag of words in each document.

Name	**Classes**	**Nodes**	**Edges**	**Features**	Task
Cora	7	2708	10.556	1.433	LP/NC
CoraML	7	2.995	16.316	2.879	LP/NC
Citeseer	6	3.327	9.104	3.703	LP/NC
PubMed	3	19.717	88.648	500	LP/NC
DBLP	4	17.716	105.734	1.639	LP/NC

Table 4.1: Summary of the datasets used for link prediction and node classification tasks. Note: LP - link prediction, NC - node classification, AD - anomaly detection

Figure 4.2 summarizes information about the datasets. For better visualizations, we only consider 2000 nodes per graph. The bottom row presents the degree distributions in linear scale (histogram bars) and logarithmic scale (upper right dotted line), showing that most nodes have relatively few connections. In contrast, a small number of nodes have many connections, suggesting scale-free properties typical in citation networks and a power-law distribution.

[1]For all models, we consider a similar value of hyperparameters with 32 hidden channels, 16 output channels, 0.01 learning rate, and training with Adam optimizer.

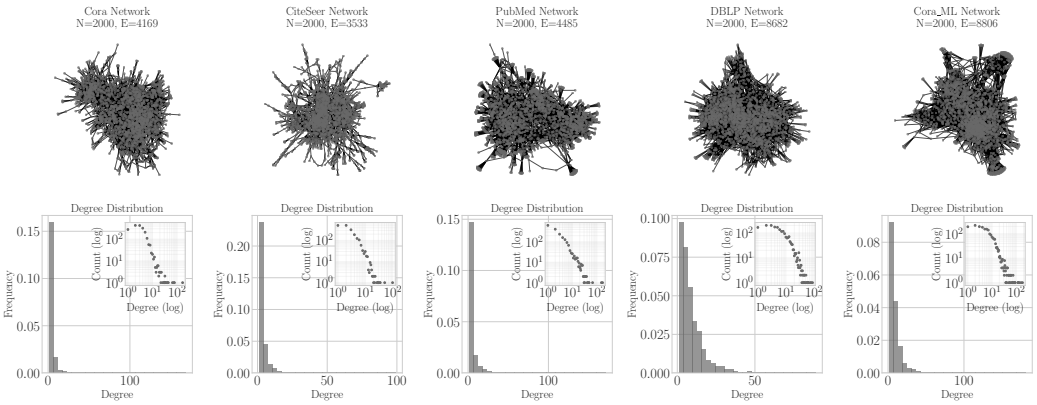

Figure 4.2: Visualization and analysis of five academic citation networks. The figure shows the network topology (top) and the degree distribution (bottom).

We compare our results with three baseline models - GAT (Veličković et al., 2018), GraphSAGE (Hamilton et al., 2017), and the VGAE - using the same environment for a fair evaluation. The metric chosen for comparison was the average precision (AP) metrics[2].

4.3.2 Node classification and link prediction

First, we employ a brute force sensitivity analysis for each dataset to select a good guess value for κ, considering node classification and link prediction tasks. Table 4.2 summarizes the results showing the best-performing κ values identified through this comparative analysis. Each value was determined by conducting two independent training runs with a small number of 50 epochs, each using the GGVA framework and considering a discrete set of 19 candidate values $\kappa = \{1.1, 1.15, 1.2, \ldots, 2\}$.

[2]AP summarizes a precision-recall curve as the weighted mean of precisions achieved at each threshold: $AP = \sum_n (R_n - R_{n-1}) P_n$. We use the implementation available in Pedregosa et al. (2011).

Table 4.2: Optimal κ parameters for node classification and link prediction across divergence types and datasets.

Task/Model	**Cora**		**CiteSeer**		**PubMed**		**DBLP**		**CoraML**	
	κ	AP	κ	AP	κ	AP	κ	AP	κ	AP
Node classification										
GGVA-Tsallis	1.60	0.8486	1.10	0.6274	1.45	0.7891	1.15	0.8666	1.95	0.9414
GGVA-Renyi	1.65	0.8333	1.65	0.6303	1.70	0.6669	1.55	0.8578	1.90	0.9282
GGVA-sAB	1.65	0.8332	1.45	0.6610	1.10	0.7994	1.50	0.8651	1.65	0.9202
GGVA-γ	1.40	0.8252	1.20	0.6325	1.75	0.7686	1.70	0.8679	1.85	0.9396
GGVA-β	1.10	0.7393	1.10	0.6133	1.50	0.7736	1.10	0.8402	1.15	0.8638
GGVA-η	1.15	0.7425	1.30	0.5190	1.50	0.7533	1.75	0.6953	1.55	0.6266
GGVA-$c\delta$	1.65	0.8425	1.75	0.6488	1.90	0.7748	1.75	0.8690	2.00	0.9414
GGVA-$s\delta$	2.00	0.8261	1.85	0.6399	1.10	0.7696	1.35	0.8633	1.65	0.9423
GGVA-α	1.90	0.8346	1.65	0.6051	1.85	0.7379	2.00	0.8625	1.80	0.9292
Link prediction										
GGVA-Tsallis	1.25	0.7315	1.50	0.7537	1.85	0.8865	1.90	0.8263	1.85	0.8000
GGVA-Renyi	1.15	0.7404	2.00	0.7279	1.45	0.8851	1.45	0.8326	1.15	0.7904
GGVA-sAB	1.75	0.7331	1.90	0.7723	1.95	0.8940	1.45	0.8217	1.30	0.8112
GGVA-γ	1.70	0.7386	1.60	0.7585	1.40	0.8887	1.90	0.8296	1.50	0.8133
GGVA-β	1.60	0.7422	1.30	0.7488	1.45	0.8826	1.10	0.8138	1.50	0.8282
GGVA-η	1.40	0.7368	1.15	0.7401	1.85	0.8805	1.15	0.8045	1.10	0.8169
GGVA-$c\delta$	1.45	0.7344	1.45	0.7502	1.75	0.8866	1.60	0.8209	1.80	0.7911
GGVA-$s\delta$	1.90	0.7374	1.90	0.7469	1.45	0.8860	1.90	0.8318	1.10	0.7843
GGVA-α	1.30	0.7362	1.40	0.7624	1.55	0.8871	1.75	0.8189	1.75	0.7925

Figure 4.3 presents the efficiency-accuracy trade-off for node classification. The results demonstrate that the GGVA framework consistently achieves competitive AP while maintaining reasonable computational requirements across different network structures and sizes. GGVA framework consistently outperforms (or at least has similar results) its backbone model (VGAE) across all tested datasets, demonstrating the effectiveness of incorporating generalized information measures. The figure also shows that η and β have the worst results compared with the other κ-divergences.

Analysing the same efficiency-accuracy trade-off for link prediction (Figure 4.4, we observe that GGVA consistently outperforms GAT and GraphSAGE. However, the average results match those of VGAE in most cases. As observed in the node classification task, the divergence with the worst results in GGVA is η.

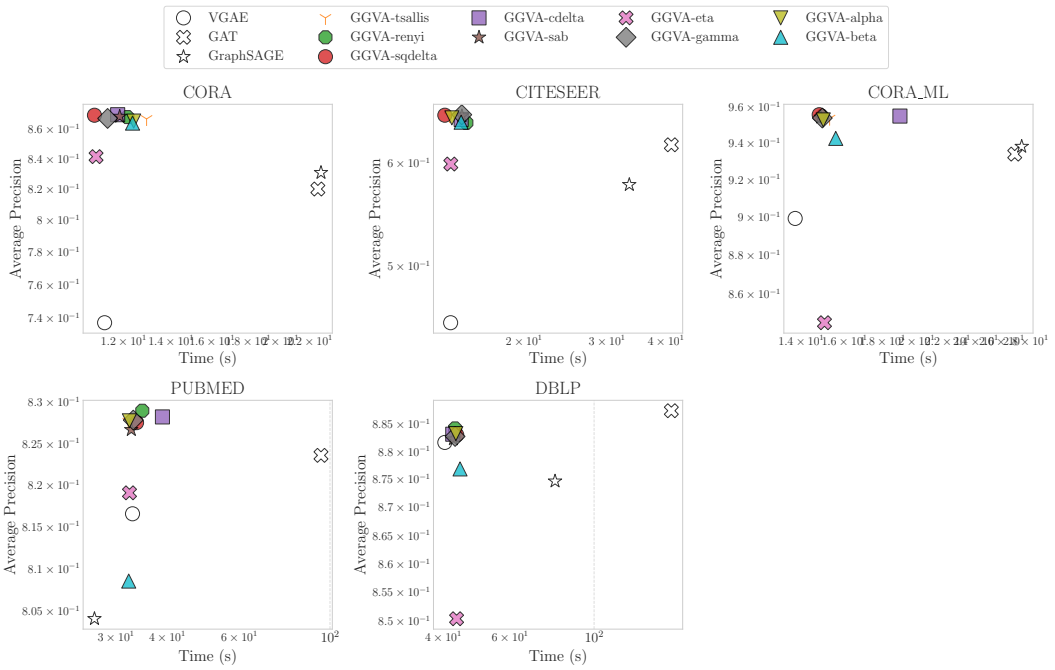

Figure 4.3: Efficiency-accuracy trade-off for node classification. Accuracy is measured as average precision (AP) and efficiency in seconds. Numbers show mean results for 10 runs with random initializations on fixed dataset splits.

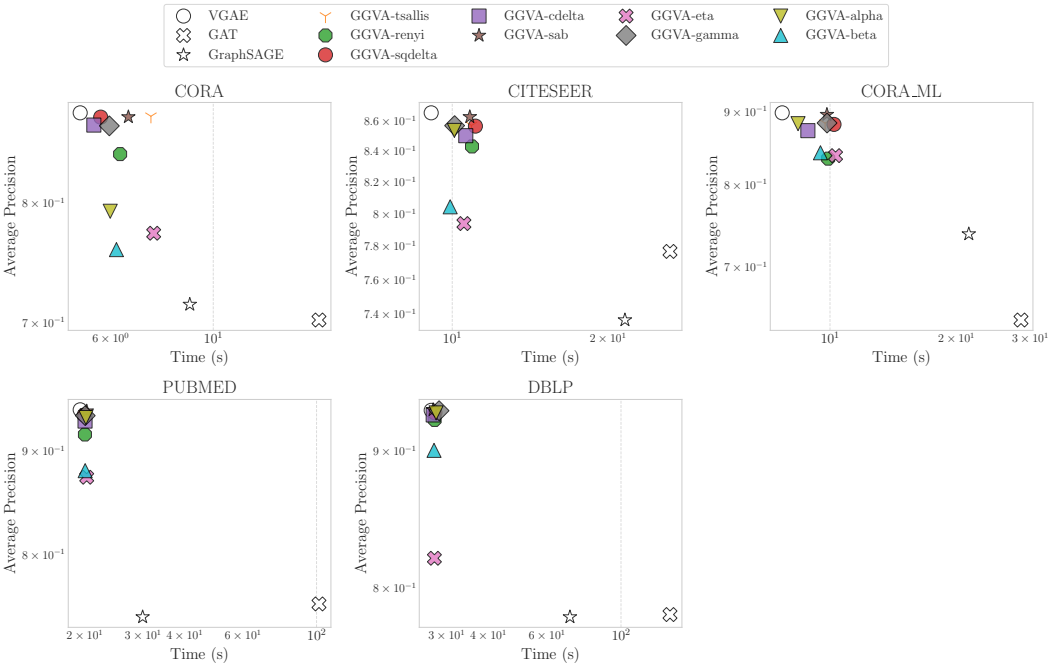

Figure 4.4: Efficiency-accuracy trade-off for link prediction. Accuracy is measured as average precision (AP) and efficiency in seconds. Numbers show mean results for 10 runs with random initializations on fixed dataset splits.

4.3.3
Qualitative analysis for node embeddings

For clarity, we focus our analysis on the Cora dataset. Consistent with the quantitative results, 2D projections of the learned embeddings exhibit distinct clusters that align with Cora's seven topic classes. This projection achieves a Silhouette score[3] comparable to VGAE's, although the resulting visual clustering patterns differ.

As quantitative results show, the η-divergence doesn't have good average precision w.r.t the other generalized divergences in GGVA. This behavior is more explicit when we observe the embeddings of η, which show imprecise clusters and a negative silhouette score.

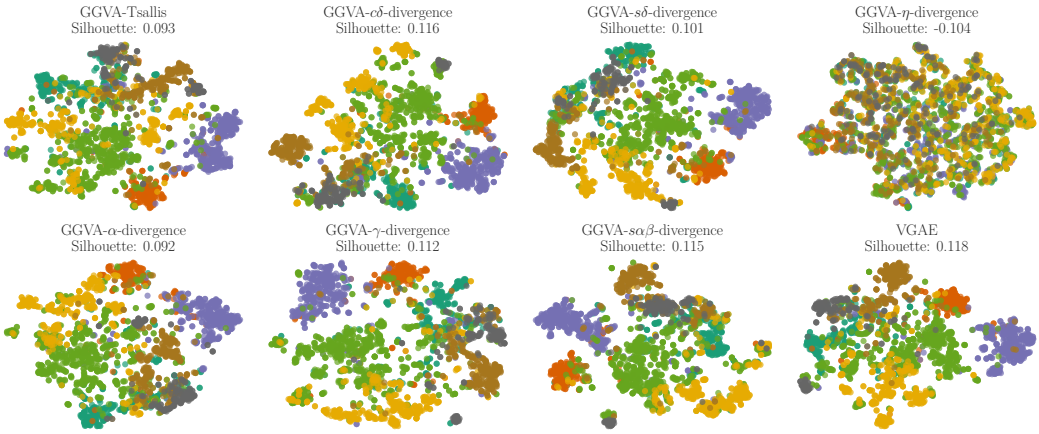

Figure 4.5: t-SNE embeddings of the nodes in the BPGN dataset from trained GGVA and VGAE

4.4
Experiment III: Learning in Power Grid Networks

4.4.1
Brazilian power grid network

The applications of LoG in electrical power grid networks have potential in monitoring and planning (Liao et al., 2021; Ringsquandl et al., 2021; Hansen et al., 2022). In this experiment, we investigate the link prediction capacities of GGVA and baseline models applied in the Brazilian power grid network (BPGN) (ONS, 2024). Some motivations for this problem are: (i) link prediction can assist in designing new transmission lines by suggesting

[3]The silhouette score (Rousseeuw, 1987) measures how similar a point is to its cluster (cohesion) compared to other clusters (separation). The metric range is $(1, -1)$. 1 represents the best result, and -1 represents the worst result. Values near 0 indicate overlapping clusters. Negative values generally indicate that a sample has been assigned to the wrong cluster, as a different cluster is more similar.

optimal connections based on historical data and power demand patterns; (ii) link prediction can help detect anomalies by identifying missing, unexpected, or spurious connections. If a model predicts a high probability for a link that does not exist, it may indicate an unauthorized or malfunctioning connection.

We define the BPGN as a graph where nodes represent electricity substations. The distance weights the connection between two substations[4]. Figure 4.6 shows the topology of this network considering communities detected using the Clauset-Newman-Moore greedy modularity maximization algorithm (Clauset et al., 2004).

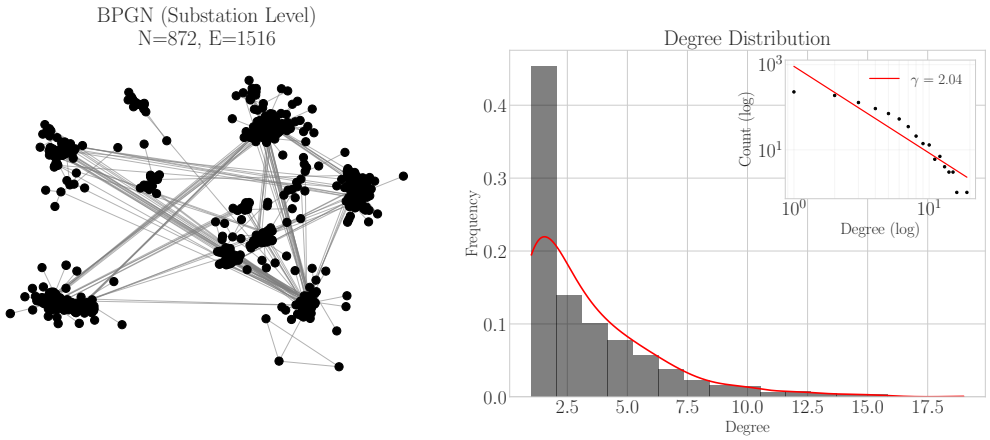

Figure 4.6: Brazilian power grid network by substations. (Left) The node colors represent the communities detected. (Right) The degree distribution and the logarithmic scale.

Table 4.3 shows the optimal κ values. As we do in the citation networks experiment, each value was determined by conducting two independent training runs with a small number of 50 epochs, each using the GGVA framework and considering a discrete set of 19 candidate values $\kappa = \{1.1, 1.15, 1.2, \ldots, 2\}$. We observe that the optimal value differs for each divergence, with the unique exception of γ and η with $\kappa^* = 1.65$.

[4]We compute 53 features for the BPGN - 5 centrality measures: degree, squared degree, betweenness, closeness, and eigenvector; and 48 edge attribute features like length, resistance, reactance, and operational capacity for different contexts.

Table 4.3: Optimal κ values by divergence type for BPGN

Model	κ^*	**AP**
GGVA-Tsallis	1.85	0.6877 ± 0.0001
GGVA-Renyi	2.00	0.6864 ± 0.0007
GGVA-sAB	1.60	0.6868 ± 0.0006
GGVA-γ	1.65	0.6873 ± 0.0018
GGVA-β	1.10	0.6841 ± 0.0013
GGVA-η	1.65	0.6722 ± 0.0063
GGVA-$c\delta$	1.90	0.6876 ± 0.0009
GGVA-$s\delta$	1.25	0.6865 ± 0.0003
GGVA-α	1.75	0.6877 ± 0.0006

Results show the optimal κ value that maximizes average precision (AP) for each divergence type.

Table 4.4 summarizes the results for link prediction in BPGN. We observe that VGAE has the best AP (0.8728), closely followed by several GGVA variants (e.g., GGVA-cdelta, GGVA-gamma, GGVA-sab), which achieve APs around 0.86-0.87. These models significantly outperform GAT (0.5682) and GraphSAGE (0.6538) in accuracy. GGVA-η is an exception with lower performance (0.7455).

Table 4.4: Link prediction performance on BPGN

Model	**AP**	**Time (s)**
GGVA-Tsallis	0.8650 ± 0.0056	4.23 ± 0.42
GGVA-Renyi	0.8659 ± 0.0046	5.50 ± 0.55
GGVA-sAB	0.8669 ± 0.0050	4.26 ± 0.43
GGVA-γ	0.8709 ± 0.0047	5.89 ± 0.59
GGVA-β	0.8679 ± 0.0057	4.85 ± 0.49
GGVA-η	0.7455 ± 0.0088	5.30 ± 0.53
GGVA-$c\delta$	0.8707 ± 0.0057	4.29 ± 0.43
GGVA-$s\delta$	0.8667 ± 0.0033	4.87 ± 0.49
GGVA-α	0.8611 ± 0.0057	4.23 ± 0.42
GAT	0.5682 ± 0.0100	17.48 ± 1.75
GraphSAGE	0.6538 ± 0.0111	2.48 ± 0.25
VGAE	**0.8728 ± 0.0057**	5.18 ± 0.52

Results show mean ± standard deviation across runs. The best AP result is in **bold**.

4.4.2
Qualitative analysis for node embeddings

Due to the absence of ground-truth labels for BPGN, we employ K-Means (Macqueen, 1967) clustering ($k = 10$) on the learned embedding space to generate proxy labels. Subsequently, we evaluate the quality of these embeddings by calculating the silhouette score based on the resulting clusters. Figure 4.7 shows that the learned embeddings using GGVA - emphasis in $s\delta$ and γ - have reasonably distinguishable clusters.

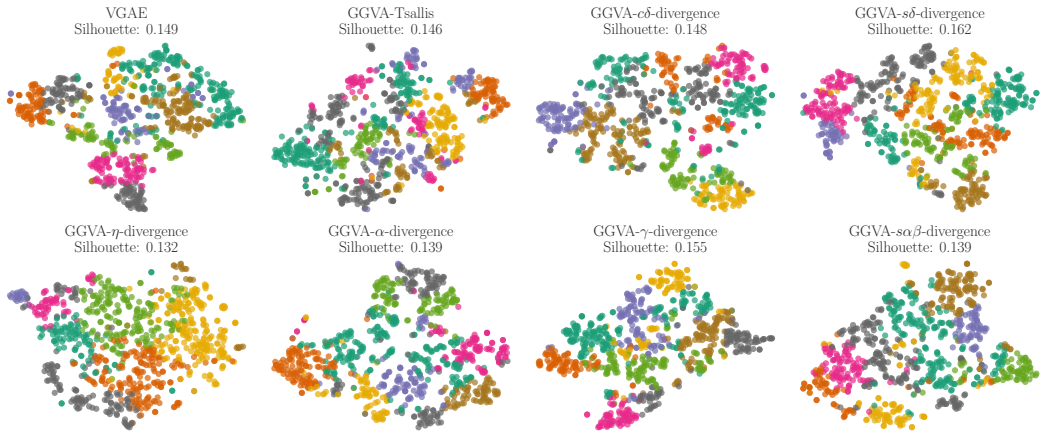

Figure 4.7: t-SNE embeddings of the nodes in the BPGN dataset from trained GGVA and VGAE

4.5
Discussion

The results from Experiment II demonstrate the efficacy of the proposed GGVA framework on benchmark citation networks. Across the five datasets (Cora, CoraML, Citeseer, PubMed, DBLP), the GGVA based on different κ-divergences consistently achieved competitive performance in node classification, as measured by average precision (AP). Figures 4.3 and 4.4 illustrate the efficiency-accuracy trade-off, revealing that GGVA models often surpass or match the performance of the baseline VGAE in link prediction, upon which they are built, while maintaining reasonable computational times, generally outperforming the baseline models GAT and GraphSAGE in efficiency. Furthermore, the sensitivity analysis presented in Table 4.2 highlights a crucial aspect: the optimal hyperparameter κ varies considerably depending on the specific dataset, the task (node classification vs. link prediction), and the chosen κ-divergence measure. This variability underscores the flexibility of the GGVA framework. It suggests that different generalized divergences, with appropriately tuned parameters, can capture distinct structural or feature-based information important for different learning objectives on graphs.

Considering Experiment III, the link prediction task was evaluated on the BPGN. The results, summarized in Table 4.4, present a nuanced picture. While several GGVA variants (notably GGVA-$c\delta$, GGVA-sAB, and GGVA-γ) achieved high AP scores, very close to the top performer, the baseline VGAE model registered the highest AP (0.8818 ± 0.0047). This indicates that for this specific infrastructure network and feature set, the standard VGAE was effective, slightly outperforming the generalized divergences explored within the GGVA framework on average. However, the differences were marginal for the best GGVA variants. All GGVA variants and VGAE significantly outperformed GAT and GraphSAGE for AP. The optimal κ values determined for BPGN (Table 4.3) also differed from those found optimal for citation networks, indicating the problem-dependent nature of hyperparameter tuning. Although GGVA did not distinctly outperform VGAE on BPGN as consistently observed in Experiment II for link prediction, its strong performance confirms its viability for real-world graph learning tasks beyond citation networks, offering flexible divergence options that perform competitively with the established baseline.

The selection of the hyperparameter κ requires careful attention. Our findings demonstrate that the optimal κ is highly sensitive, varying substantially across datasets, learning tasks, and the specific generalized divergence employed. This sensitivity offers a key advantage: it endows the framework with flexibility, allowing it to potentially capture diverse structural or feature-based nuances by appropriately tuning κ alongside the divergence choice. Otherwise, this dependency introduces a practical challenge, as identifying the optimal κ typically necessitates an additional hyperparameter search (e.g., grid search), consequently increasing the computational overhead required for model optimization and deployment, which is suggestive for future research directions. An efficient method for choosing optimal κ values for robust learning objectives becomes a desirable goal to achieve.

4.6 Summary

Variational methods are a powerful tool for learning on graph-structured data. This chapter investigated using flexible divergences as learning objectives in setting variational graph autoencoders. This framework is called generalized variational graph autoencoders (GGVA). We employ nine different κ-divergences for node classification and link prediction tasks in citation networks and link prediction in the Brazilian power grid network.

The results indicate that GGVA outperforms the baseline models —

GAT, GraphSAGE, and VGAE — in all scenarios, with emphasis on node classification tasks. A qualitative analysis employed in learned embeddings, where the GGVA variants show a reasonable silhouette score, was used to help explain this performance.

5
Conclusion and Future Work

5.1
Conclusion

The classical variational methods in graph-structured data apply an evidence lower bound based on Kullback-Leibler divergence to minimize the information loss in an autoencoder setup. In this work, we investigate the use of generalized divergence measures for unsupervised learning on graphs via variational autoencoders. We propose a comprehensive collection of parametrized divergences, which we unify in the class of κ-divergences and investigate their behavior for mode-seeking or mass-covering in a mixture of Gaussians approximation problems.

Taking these generalized measures, we build the generalized graph variational autoencoders (GGVA) framework and evaluate their results for link prediction and node classification tasks considering a benchmark with five academic citation networks (Cora, CiteSeer, PubMed, DBLP, CoraML) and three baseline models (VGAE, GAT, GraphSAGE). Using the Brazilian power grid network, we also evaluate our framework in a novel dataset for link prediction.

Main findings

- Based on a variational inference problem for approximating a mixture of Gaussians using a simple prior, our three novel divergence measures - $c\delta$, $s\delta$, and η - show a mass-covering behavior;
- For node classification, GGVA outperforms the baseline models in the five academic citation networks considering time efficiency and average precision;
- For link prediction in academic citation networks, GGVA outperforms GAT and GraphSAGE considering time efficiency and average precision. And the results are similar with VGAE;
- For link prediction in the Brazilian power grid network, we observe that GGVA outperforms GAT and GraphSAGE and has competitive results aligned with VGAE.

5.2 Future Work

Some research directions for our current work are:

- A key methodological challenge involves establishing a rigorous framework for optimal parameter selection κ. Developing an approach for tuning κ ensures model robustness and maximizes performance across diverse datasets and tasks.
- Extending the theoretical underpinnings of our approach presents a significant opportunity. We want to explore the integration of generalized measures of information and principles from non-extensive statistical mechanics, particularly within the context of Temporal Graph Neural Networks (TGNNs).
- Finally, improving the explainability of the learned representations is essential for building trust and extracting actionable insights. We aim to develop methods specifically designed to enhance the explainability of the latent space generated by graph neural networks, exploring how a generalized mutual information in GNNExplainer-inspired methods could improve explainability.

Bibliography

Abe, S. and Okamoto, Y. (2001). *Nonextensive statistical mechanics and its applications*, volume 560. Springer Science & Business Media.

Amari, S.-i. (1985). *Differential-geometrical methods in statistics*. Springer Science & Business Media.

Barber, D. and Wiegerinck, W. (1998). Tractable variational structures for approximating graphical models. *Advances in Neural Information Processing Systems*, 11.

Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559.

Battaglia, P., Pascanu, R., Lai, M., Jimenez Rezende, D., et al. (2016). Interaction networks for learning about objects, relations and physics. *Advances in neural information processing systems*, 29.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. (2018). Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR.

Benettin, G., Galgani, L., Giorgilli, A., and Strelcyn, J.-M. (1980). Lyapunov characteristic exponents for smooth dynamical systems and for Hamiltonian systems; a method for computing all of them. *Meccanica*, 15(1):9–20.

Bishop, C., Lawrence, N., Jaakkola, T., and Jordan, M. (1997). Approximating posterior distributions in belief networks using mixtures. *Advances in neural information processing systems*, 10.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Boltzmann, L. (1872). *Weitere studien über das wärmegleichgewicht unter gasmolekülen*, volume 66. Aus der kk Hot-und Staatsdruckerei.

Brody, S., Alon, U., and Yahav, E. (2021). How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*.

Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges.

Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience*, 10(3):186–198.

Cichocki, A. and Amari, S.-i. (2010). Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568.

Cichocki, A., Cruces, S., and Amari, S.-i. (2011). Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy*, 13(1):134–170.

Cichocki, A., Lee, H., Kim, Y.-D., and Choi, S. (2008). Non-negative matrix factorization with α-divergence. *Pattern Recognition Letters*, 29(9):1433–1440.

Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 70(6):066111.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

Csiszár, I. (1967). On information-type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318.

de Oliveira, R., Brito, S., da Silva, L., and Tsallis, C. (2021). Connecting complex networks to nonadditive entropies. *Scientific Reports*, 11(1):1130.

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29.

Derrow-Pinion, A., She, J., Wong, D., Lange, O., Hester, T., Perez, L., Nunkesser, M., Lee, S., Guo, X., Wiltshire, B., et al. (2021). Eta prediction with graph neural networks in google maps. In *Proceedings of the 30th*

ACM international conference on information & knowledge management, pages 3767–3776.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Easley, D., Kleinberg, J., et al. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*, volume 1. Cambridge university press Cambridge.

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., and Larson, J. (2024). From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Euler, L. (1741). Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, pages 128–140.

Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202.

Ganguly, A. and Earp, S. W. (2021). An introduction to variational inference. *arXiv preprint arXiv:2108.13083*.

Gell-Mann, M. and Tsallis, C. (2004). *Nonextensive entropy: interdisciplinary applications*. Oxford University Press.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions*, 6:721–741.

Geyer, C. J. (1992). Practical markov chain monte carlo. *Statistical science*, pages 473–483.

Gibbs, J. W. (1902). *Elementary Principles in Statistical Mechanics*. Scribner's Sons.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR.

Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Hamilton, W. L. (2020). *Graph representation learning*. Morgan & Claypool Publishers.

Hansen, J. B., Anfinsen, S. N., and Bianchi, F. M. (2022). Power flow balancing with decentralized graph neural networks. *IEEE Transactions on Power Systems*, 38(3):2423–2433.

Hernández-Lobato, J. M., Li, Y., Rowland, M., Hernández-Lobato, D., Bui, T., and Turner, R. E. (2016). Black-box α-divergence minimization. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1511–1520.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Itakura, F. (1968). Analysis synthesis telephony based on the maximum likelihood method. *Reports of the 6ˆ< th> Int. Cong. Acoust., 1968*.

Järvenpää, M. and Corander, J. (2023). On predictive inference for intractable models via approximate bayesian computation. *Statistics and Computing*, 33(2).

Jebara, T. (2004). *Machine learning: discriminative and generative*, volume 755. Springer Science & Business Media.

Johnson, R., Li, M. M., Noori, A., Queen, O., and Zitnik, M. (2024). Graph artificial intelligence in medicine. *Annual Review of Biomedical Data Science*, 7(Volume 7, 2024):345–368.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37:183–233.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes.

Kingma, D. P. and Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392.

Kipf, T. N. and Welling, M. (2016a). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kipf, T. N. and Welling, M. (2016b). Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*.

Knoblauch, J., Jewson, J., and Damoulas, T. (2019). Generalized variational inference: Three arguments for deriving new posteriors. *arXiv preprint arXiv:1904.02063*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al. (2023). Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Li, Y. and Turner, R. E. (2016). Rényi divergence variational inference. *Advances in neural information processing systems*, 29.

Liang, H., Borde, H. S. d. O., Sripathmanathan, B., Bronstein, M., and Dong, X. (2025). Towards quantifying long-range interactions in graph machine learning: a large graph dataset and a measurement. *arXiv preprint arXiv:2503.09008*.

Liao, W., Bak-Jensen, B., Pillai, J. R., Wang, Y., and Wang, Y. (2021). A review of graph neural networks and their applications in power systems. *Journal of Modern Power Systems and Clean Energy*, 10(2):345–360.

Liu, G., Catacutan, D. B., Rathod, K., Swanson, K., Jin, W., Mohammed, J. C., Chiappino-Pepe, A., Syed, S. A., Fragis, M., Rachwalski, K., Magolan, J., Surette, M. G., Coombes, B. K., Jaakkola, T., Barzilay, R., Collins, J. J., and Stokes, J. M. (2023). Deep learning-guided discovery of an antibiotic targeting acinetobacter baumannii. *Nature Chemical Biology*, 19(11):1342–1350.

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–298. University of California press.

Mavromatis, C. and Karypis, G. (2024). Gnn-rag: Graph neural retrieval for large language model reasoning. *arXiv preprint arXiv:2405.20139*.

Mihoko, M. and Eguchi, S. (2002). Robust blind source separation by beta divergence. *Neural computation*, 14(8):1859–1886.

Minka, T. P. (2005). Divergence measures and message passing. In *Microsoft Research Technical Report*.

Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., and Bronstein, M. M. (2017). Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5115–5124.

Morimoto, T. (1963). Markov processes and the h-theorem. *Journal of the Physical Society of Japan*, 18(3):328–331.

Newman, M. (2018). *Networks*. Oxford university press.

Nowozin, S., Cseke, B., and Tomioka, R. (2016). f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, 29.

ONS (2024). Linhas de transmissão da rede em operação. Accessed in 12/10/2024.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Plastino, A. and Plastino, A. (1995). Non-extensive statistical mechanics and generalized fokker-planck equation. *Physica A: Statistical Mechanics and its Applications*, 222(1-4):347–354.

Poole, B., Alemi, A. A., Sohl-Dickstein, J., and Angelova, A. (2016). Improved generator objectives for gans. *arXiv preprint arXiv:1612.02780*.

Regli, J.-B. and Silva, R. (2018). Alpha-beta divergence for variational inference. *arXiv preprint arXiv:1805.01045*.

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pages 547–562. University of California Press.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *International conference on machine learning*, pages 1278–1286.

Ringsquandl, M., Sellami, H., Hildebrandt, M., Beyer, D., Henselmeyer, S., Weber, S., and Joblin, M. (2021). Power to the relational inductive bias: Graph neural networks in electrical power grids. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, page 1538–1547. ACM.

Robledo, A. and Velarde, C. (2022). How, why and when tsallis statistical mechanics provides precise descriptions of natural phenomena. *Entropy*, 24(12).

Roudbari, N. S., Punekar, S. R., Patterson, Z., Eicker, U., and Poullis, C. (2024). From data to action in flood forecasting leveraging graph neural networks and digital twin visualization. *Scientific Reports*, 14(1):18571.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Saul, L. and Jordan, M. (1995). Exploiting tractable substructures in intractable networks. *Advances in neural information processing systems*, 8.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Silva, T., de Souza da Silva, E., and Mesquita, D. (2024). On divergence measures for training GFlownets. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., Tran, V. M., Chiappino-Pepe, A., Badran, A. H., Andrews, I. W., Chory, E. J., Church, G. M., Brown, E. D., Jaakkola, T. S., Barzilay, R., and Collins, J. J. (2020). A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702.e13.

Tirnakli, U., Beck, C., and Tsallis, C. (2007). Central limit behavior of deterministic dynamical systems. *Physical Review E*, 75(4):040106.

Tsallis, C. (1988). Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52:479–487.

Tsallis, C. (2009). *Introduction to nonextensive statistical mechanics: approaching a complex world*, volume 1. Springer.

Tsallis, C. (2023). Non-additive entropies and statistical mechanics at the edge of chaos: a bridge between natural and social sciences. *Philosophical Transactions of the Royal Society A*, 381(2256):20220293.

Tsallis, C. and Cirto, L. J. (2013). Black hole thermodynamical entropy. *The European Physical Journal C*, 73:1–7.

Tsallis, C., Gell-Mann, M., and Sato, Y. (2005). Asymptotically scale-invariant occupancy of phase space makes the entropy S_q extensive. *Proceedings of the National Academy of Sciences*, 102(43):15377–15382.

Tsallis, C., Mendes, R., and Plastino, A. R. (1998). The role of constraints within generalized nonextensive statistics. *Physica A: Statistical Mechanics and its Applications*, 261(3-4):534–554.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph attention networks. In *International Conference on Learning Representations*.

Wan, N., Li, D., and Hovakimyan, N. (2020). F-divergence variational inference. *Advances in neural information processing systems*, 33:17370–17379.

Wang, Z., So, O., Gibson, J., Vlahov, B., Gandhi, M. S., Liu, G.-H., and Theodorou, E. A. (2021). Variational inference mpc using tsallis divergence. *arXiv preprint arXiv:2104.00241*.

Wang, Z., Veličković, P., Hennes, D., Tomašev, N., Prince, L., Kaisers, M., Bachrach, Y., Elie, R., Wenliang, L. K., Piccinini, F., Spearman, W., Graham, I., Connor, J., Yang, Y., Recasens, A., Khan, M., Beauguerlange, N., Sprechmann, P., Moreno, P., Heess, N., Bowling, M., Hassabis, D., and Tuyls, K. (2024). Tacticai: an ai assistant for football tactics. *Nature Communications*, 15(1):1906.

Wen, T. and Jiang, W. (2019). Measuring the complexity of complex network by tsallis entropy. *Physica A: Statistical Mechanics and its Applications*, 526:121054.

Wieder, O., Kohlbacher, S., Kuenemann, M., Garon, A., Ducrot, P., Seidel, T., and Langer, T. (2020). A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies*, 37:1–12.

Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., and Weinberger, K. (2019). Simplifying graph convolutional networks. In *International conference on machine learning*, pages 6861–6871. Pmlr.

Yang, Z., Cohen, W., and Salakhudinov, R. (2016). Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, pages 40–48. PMLR.

Zhang, Z., Chen, L., Zhong, F., Wang, D., Jiang, J., Zhang, S., Jiang, H., Zheng, M., and Li, X. (2022). Graph neural network approaches for drug-target interactions. *Current Opinion in Structural Biology*, 73:102327.

Zhu, L., Chen, Z., Schlegel, M., and White, M. (2023). General munchausen reinforcement learning with tsallis kullback-leibler divergence. *Advances in Neural Information Processing Systems*, 36:57639–57659.

Zimmert, J. and Seldin, Y. (2021). Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49.

Zitnik, M., Agrawal, M., and Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466.

A
An Overview of Learning on Graphs

Following, we present an overview of graph neural networks (GNN), the main architectures, and theoretical aspects. In summary, a GNN is a general framework for deep learning on graph-structured data. The main idea behind this architecture is to employ *deep encoder* that generates representations of nodes based on the graph structure and any relevant node feature information.

A.1
Graph Neural Networks

Graph Neural Networks (GNNs) are an effective framework for LoG and the most general class of deep learning architectures, considering that most other deep learning architectures can be viewed as a special case of GNNs with additional geometric structure Bronstein et al. (2021). GNNs use a messaging passing in which vector messages are exchanged between nodes and updated using neural networks - for example, multilayer perceptron (Gilmer et al., 2017). In a simple form, the GNNs update a hidden embedding h_v^k by the following message-passing rule with neighbourhood aggregation for each iteration,

$$h_v^{k+1} = \phi(h_v^k, \bigoplus_{u \in \mathcal{N}_v} \psi(h_u^k)) \tag{A-1}$$

where $\bigoplus$ is a permutation invariant function aggregator, ψ is a neighbour aggregation function, and ϕ is a propagation function. With this basic form, we can observe that the message depends only on the current state h_u and it depends on the entire state. Any notion of local information has to be implemented by ψ. Another aspect of this architecture is that nodes never stop transmitting their state, so there is no easy way to define termination conditions. In this case, not only does ψ need to learn to return zero, but ϕ also has to learn the identity function on those zero messages.

The intuition of Equation A-1 is that for each iteration, every node aggregates information from its local neighbourhood, and as these iterations progress, each node embedding contains more information from further reaches of the graph. This information can be described as (i) structure and (ii) feature information. For *structure information*, after k iterations of GNN message-

passing, the h_v^k might encode information about the degrees of all nodes in $\mathcal{N}_v^k$. And after k iterations, the h_v^k encode feature information of all nodes in $\mathcal{N}_v^k$.

A.1.1 Three possible variations of GNN Layers

There are different forms to compute how the information is transmitted from $\mathcal{N}_v^k$ to v. According to Bronstein et al. (2021), the GNNs layers design observed in the literature can be divided into three variations - *convolutional*, *attentional*, and *message-passing* - that guide the extent to which ϕ change $\mathcal{N}_v$ features. Following we describe these three variations and present a simplified representation (Figure A.1).

A **message-passing layer** involves computing arbitrary vectors - the *messages* - across the edges of a graph using a neighborhood aggregation function ψ. Some examples in literature are interaction networks (Battaglia et al., 2016), Message-passing Neural Networks (MPNN) (Gilmer et al., 2017), and relational inductive biases (Battaglia et al., 2018). We can define these variations following Equation A-1 as

$$h_v^k = \phi(\mathbf{x}_v, \bigoplus_{u \in \mathcal{N}_v} \psi(\mathbf{x}_u)) \tag{A-2}$$

A **convolutional layer** involves aggregating the features of the neighborhood nodes based on fixed weights. Some examples in literature are the Graph Convolutional Network (GCN) (**?**), ChebyNet (Defferrard et al., 2016), and simple-GCN (Wu et al., 2019). We can define this variation as

$$h_v^k = \phi(\mathbf{x}_v, \bigoplus_{u \in \mathcal{N}_v}^k c_{vu}\psi(\mathbf{x}_u)) \tag{A-3}$$

where c_{vu} represents the importance of u to node $v's$ representation.

The **attentional layer** involves weighing the neighborhood influence during the aggregation. Some examples in literature are the Graph Attention Network (GAT) (Veličković et al., 2018), MoNet (Monti et al., 2017), and the GATv2 (Brody et al., 2021). We can define this variation as

$$h_v^k = \phi(\mathbf{x}_v, \bigoplus_{u \in \mathcal{N}_v} \delta(\mathbf{x}_v, \mathbf{x}_u)\psi(\mathbf{x}_u)) \tag{A-4}$$

where δ is a learnable self-attention mechanism used to compute the importance coefficients $\alpha_{v,u}$. We can achieve the importance coefficients via

$$\alpha_{v,u} = \frac{\exp(\delta^\top [\mathbf{x}_v \bigoplus \mathbf{x}_u])}{\sum_{\tilde{u} \in \mathcal{N}_v} \exp(\delta^\top \mathbf{x}_v \bigoplus \mathbf{x}_{\tilde{u}})} \tag{A-5}$$

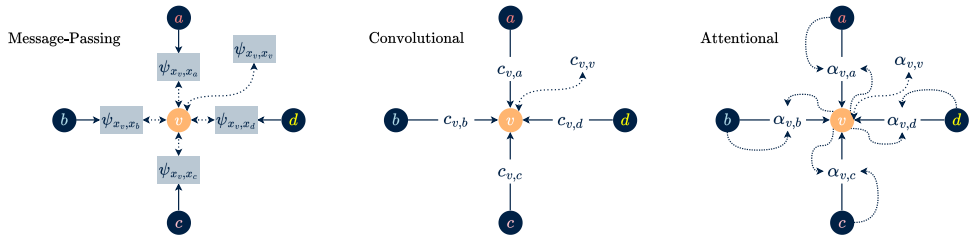

Figure A.1: Comparison of three neural network architectures for graph processing. Message-Passing (left), Convolutional (center), and Attentional (right). Each diagram illustrates how information flows between nodes in the network with their respective mathematical formulations.

A.1.2
Traditional tasks in learning on graphs

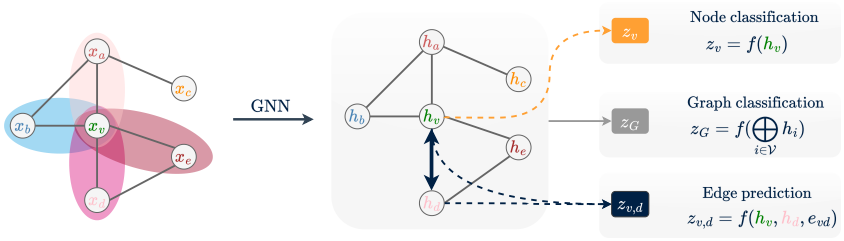

Figure A.2: Illustration of tasks in learning on graphs. The figure shows three fundamental graph learning tasks: node classification ($z_v = f(h_v)$), graph classification ($z_G = f(\bigoplus_{v \in V} h_v)$), and edge prediction ($z_{u,v} = f(h_u, h_v)$).

Figure A.2 illustrates a GNN framework for learning on graph-structured data. The left side depicts an input graph with nodes containing features (x_1, x_2, etc.). After processing through the GNN, the right side shows the transformed graph with learned node representations (h_a, h_b, etc.) that capture both node features and structural information. These learned embeddings are then utilized for three fundamental graph learning tasks, as indicated by the different output pathways.

Each task leverages the GNN embeddings in a distinct manner: node classification predicts labels for individual nodes using their respective embeddings; graph classification aggregates all node embeddings (using some pooling operation $\bigoplus$) to make predictions about the entire graph; and edge prediction evaluates potential connections between node pairs by combining their representations. This unified architecture demonstrates how GNNs provide a versatile framework for various LoG applications, from molecular property prediction to social network analysis and recommender systems.

A.1.3 Theoretical aspects of GNNs

Given any graph $\mathcal{G}$ with an adjacency matrix A, a deep encoder that generates embeddings based on A should ideally satisfy *permutation invariance* or *permutation equivariance*.

Definition A.1 (Permutation invariance and equivariance) *Let $A \in \mathbb{R}^{n\times n}$ be the adjacency matrix of a graph $\mathcal{G}$, and let $P \in \{0,1\}^{n\times n}$ be a permutation matrix. We say that a function $f : \mathbb{R}^{n\times n} \to \mathbb{R}^d$ is permutation invariant if:*

$$f(PAP^\top) = f(A) \quad \forall P \tag{A-6}$$

Similarly, a function $g : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times d}$ is permutation equivariant if:

$$g(PAP^\top) = Pg(A) \quad \forall P \tag{A-7}$$

Intuitively, permutation invariance means that the output of f does not depend on the specific ordering of the entries in A. In contrast, permutation equivariance means that the output of f is permuted consistently when we permute the A.

These properties are crucial for GNNs as they ensure that the learned representations depend only on the graph structure and not an arbitrary node ordering.

B
An Overview of Divergence Measures

We present an overview of the literature's main divergence measures and families of divergences. We focus on the f-divergences, the α, β, γ families, and the non-additive divergences families. By convention, we consider two probability distributions P and Q with probability density functions $p(x)$ and $q(x)$, respectively.

B.1
Foundational f-divergences

The generator function $f(\phi)$ of an f-Divergence is a convex function that defines the divergence between two probability distributions P and Q through the following general equation,

$$D_f(P \parallel Q) = \int_{\mathcal{X}} f\left(\frac{p(x)}{q(x)}\right) q(x)\, dx \tag{B-1}$$

where $p(x)$ and $q(x)$ are the probability densities (or mass functions) of the distributions P and Q, respectively; $\mathcal{X}$ is the support set of X; $f(\phi) : \mathbb{R}_+ \to \mathbb{R}$ is a convex function that satisfies $f(1) = 0$, and $\frac{p(x)}{q(x)}$ is the likelihood ratio or the Radon-Nikodym derivative of P w.r.t Q.

We can derive the corresponding $f(t)$ function for an f-divergence using the integral expression of a known divergence formula. A common formulation that satisfies f is the Kullback-Leibler divergence (D_{KL}).

Definition B.1 (Kullback-Leibler divergence) *Given two distributions P and Q, $D_{KL}(P \parallel Q)$ is given by*

$$D_{KL}(P \parallel Q) = \int_{\mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \tag{B-2}$$

B.2
α, β, and γ divergence families

The families of α, β, and γ divergences are well-established in literature considering their flexibility and robust characteristics (Cichocki and Amari, 2010).

Definition B.2 (Rényi divergence) *The Rényi α-divergence (Rényi, 1961) with $\alpha > 0$ and $\alpha \neq 1$ is defined as*

$$D_\alpha^R(P \parallel Q) = \frac{1}{1-\alpha} \log \int_{\mathcal{X}} p^\alpha(x) q^{1-\alpha}(x) dx \tag{B-3}$$

Definition B.3 (α-divergence) *The α-divergence (Amari, 1985; Cichocki et al., 2008) with $\alpha \in \mathbb{R}[0,1]$ is defined as*

$$D_\alpha(P \parallel Q) = \frac{1}{\alpha(1-\alpha)} \left[1 - \int_{\mathcal{X}} p^\alpha(x) q^{1-\alpha}(x) dx\right] \tag{B-4}$$

Definition B.4 (β-divergence) *The β-divergence (Basu et al., 1998; Mihoko and Eguchi, 2002) with $\beta \in \mathbb{R}[0,1]$ is defined as*

$$D_\beta(P \parallel Q) = \frac{1}{\beta(\beta-1)} \int_{\mathcal{X}} p(x)^\beta dx + \frac{1}{\beta} \int_{\mathcal{X}} q(x)^\beta dx - \frac{1}{\beta-1} \int_{\mathcal{X}} p(x) q(x)^{\beta-1} dx \tag{B-5}$$

Definition B.5 (γ-divergence) *The γ-divergence (Fujisawa and Eguchi, 2008) with $\gamma \in \mathbb{R}[0,1]$ is defined as*

$$D_\gamma(P \parallel Q) = \frac{1}{\gamma(\gamma-1)} \log \frac{(\int_{\mathcal{X}} p(x)^\gamma dx)(\int_{\mathcal{X}} q(x)^\gamma dx)^{\gamma-1}}{(\int_{\mathcal{X}} p(x) q(x)^\gamma dx)^\gamma} \tag{B-6}$$

Definition B.6 (sAB-divergence) *The scale invariant sAB-divergence introduced by Cichocki et al. (2011) is defined as*

$$\begin{aligned} D_{sAB}^{\alpha,\beta}(P \parallel Q) \equiv \; & \tfrac{1}{\beta(\alpha+\beta)} \log \textstyle\int_{\mathcal{X}} p(x)^{\alpha+\beta} dx \\ & + \tfrac{1}{\alpha(\alpha+\beta)} \log \textstyle\int_{\mathcal{X}} q(x)^{\alpha+\beta} dx \\ & - \tfrac{1}{\alpha\beta} \log \textstyle\int_{\mathcal{X}} p(x)^\alpha q(x)^\beta dx, \end{aligned} \tag{B-7}$$

for $(\alpha, \beta) \in \mathbb{R}^2$ such that $\alpha \neq 0, \beta \neq 0$ and $\alpha + \beta \neq 0$.

B.3
Non-additive divergences

Non-extensive statistical mechanics introduce powerful measures and novel perspectives for modeling complex systems, systems at the edge of chaos. The Tsallis q-entropy represents a seminal approach in this field, providing a well-established entropic functional that generalizes the conventional Boltzmann-Gibbs-Shannon entropy (Tsallis, 1988) with many desirable properties. This generalization is defined as:

$$S_q = k\frac{1-\sum_{i=1}^{W} p_i^q}{q-1} \tag{B-8}$$

where q is the entropic index, capture a broader spectrum of correlations within systems exhibiting non-additive properties (Gell-Mann and Tsallis, 2004). Unlike traditional statistical mechanics that assume short-range interactions and ergodicity, the Tsallis framework effectively addresses systems where long-range interactions dominate the dynamic, long-term memory effects influence system behavior, and multifractal or hierarchical structures emerge naturally.

The parameter q serves as a measure of non-extensivity, with $q = 1$ recovering the standard Boltzmann-Gibbs entropy as a special case (Tsallis, 2009). For $q \neq 1$, the formalism captures complex correlation structures manifesting in numerous natural and artificial complex systems, from plasma physics and turbulent flows to financial markets and biological networks (Abe and Okamoto, 2001). This approach has proven particularly valuable in characterizing systems at the edge of chaos, where traditional statistical methods often fail to provide adequate descriptions of emergent behaviors and critical phenomena (Plastino and Plastino, 1995; Tsallis, 2023).

B.3.1
Tsallis q-divergence

Definition B.7 (Tsallis divergence) *The q-divergence (Tsallis et al., 1998) with $q \in \mathbb{R}$ is defined as*

$$D_q^T(P \parallel Q) = \frac{1}{1-q}\left(\int_{\mathcal{X}} p(x)^q q(x)^{1-q} dx - 1\right) \tag{B-9}$$

C
Detailed experiments results for citation networks

Table C.1: Average precision (AP) by model and dataset for node classification

Model	Cora	Citeseer	CoraML	PubMed	DBLP
GAT	0.8201 ± 0.0075	0.6195 ± 0.0084	0.9338 ± 0.0027	0.8236 ± 0.0046	**0.8874 ± 0.0016**
GraphSAGE	0.8308 ± 0.0116	0.5773 ± 0.0073	0.9380 ± 0.0014	0.8041 ± 0.0070	0.8747 ± 0.0027
VGAE	0.7373 ± 0.0098	0.4525 ± 0.0203	0.8992 ± 0.0063	0.8165 ± 0.0070	0.8816 ± 0.0013
GGVA-alpha	0.8656 ± 0.0091	0.6497 ± 0.0133	0.9528 ± 0.0034	0.8277 ± 0.0047	0.8832 ± 0.0016
GGVA-beta	0.8642 ± 0.0103	0.6447 ± 0.0182	0.9425 ± 0.0036	0.8085 ± 0.0243	0.8769 ± 0.0017
GGVA-cdelta	**0.8700 ± 0.0083**	0.6479 ± 0.0099	0.9549 ± 0.0022	0.8282 ± 0.0054	0.8831 ± 0.0011
GGVA-eta	0.8415 ± 0.0237	0.5988 ± 0.0479	0.8457 ± 0.0596	0.8190 ± 0.0168	0.8504 ± 0.0208
GGVA-gamma	0.8674 ± 0.0085	**0.6533 ± 0.0080**	0.9538 ± 0.0039	0.8278 ± 0.0063	0.8827 ± 0.0009
GGVA-renyi	0.8684 ± 0.0096	0.6441 ± 0.0121	0.9548 ± 0.0035	**0.8290 ± 0.0056**	0.8842 ± 0.0015
GGVA-sab	0.8688 ± 0.0066	0.6520 ± 0.0103	0.9542 ± 0.0033	0.8267 ± 0.0063	0.8822 ± 0.0014
GGVA-sqdelta	0.8696 ± 0.0065	0.6527 ± 0.0085	**0.9557 ± 0.0046**	0.8275 ± 0.0047	0.8830 ± 0.0011
GGVA-tsallis	0.8671 ± 0.0086	0.6519 ± 0.0100	0.9535 ± 0.0041	0.8276 ± 0.0054	0.8834 ± 0.0012

Results show mean ± standard deviation across runs. The best results in each column are in **bold**.

Table C.2: Average Precision (AP) by model and dataset for link prediction

Model	Cora	CiteSeet	CoraML	PubMed	DBLP
GAT	0.7019 ± 0.0175	0.7769 ± 0.0265	0.6430 ± 0.0319	0.7561 ± 0.0045	0.7819 ± 0.0016
GraphSAGE	0.7142 ± 0.0268	0.7365 ± 0.0214	0.7387 ± 0.0134	0.7450 ± 0.0235	0.7803 ± 0.0085
VGAE	**0.8850 ± 0.0150**	**0.8653 ± 0.0170**	**0.8981 ± 0.0153**	**0.9430 ± 0.0036**	0.9320 ± 0.0056
GGVA-alpha	0.7927 ± 0.0451	0.8536 ± 0.0185	0.8826 ± 0.0141	0.9348 ± 0.0097	0.9299 ± 0.0094
GGVA-beta	0.7594 ± 0.0159	0.8043 ± 0.0090	0.8421 ± 0.0110	0.8799 ± 0.0085	0.9005 ± 0.0059
GGVA-cdelta	0.8729 ± 0.0121	0.8501 ± 0.0260	0.8727 ± 0.0276	0.9310 ± 0.0112	0.9283 ± 0.0081
GGVA-eta	0.7735 ± 0.0058	0.7939 ± 0.0104	0.8382 ± 0.0115	0.8734 ± 0.0070	0.8207 ± 0.0050
GGVA-gamma	0.8721 ± 0.0183	0.8568 ± 0.0197	0.8835 ± 0.0127	0.9368 ± 0.0080	0.9317 ± 0.0047
GGVA-renyi	0.8451 ± 0.0101	0.8430 ± 0.0358	0.8342 ± 0.0360	0.9169 ± 0.0249	0.9243 ± 0.0104
GGVA-sab	0.8809 ± 0.0109	0.8626 ± 0.0137	0.8956 ± 0.0151	0.9416 ± 0.0085	**0.9324 ± 0.0047**
GGVA-sqdelta	0.8807 ± 0.0122	0.8564 ± 0.0189	0.8815 ± 0.0191	0.9374 ± 0.0104	0.9300 ± 0.0063
GGVA-tsallis	0.8816 ± 0.0174	0.8569 ± 0.0175	0.8833 ± 0.0128	0.9353 ± 0.0071	0.9302 ± 0.0078

Results show mean ± standard deviation across runs. The best results in each column are in **bold**.