Statistical Guarantees of Deep Generative Models Involving Diverse Spaces: Generation Consistency and Robustness

Anish Chakrabarty

Theoretical Statistics and Mathematics Unit

Indian Statistical Institute, Kolkata

A thesis submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
in Statistics

July, 2025

To the ones who could've

Acknowledgements

This thesis owes its existence to the extraordinary support and guidance of my supervisors, Prof. Swagatam Das and Prof. Probal Chaudhuri. While their formal supervision laid the academic foundation, it was their candid advice and unwavering encouragement that truly shaped this journey. I am profoundly thankful for the freedom, direction, and trust they offered me in equal measure.

I've had the unique experience of a dual existence between the Theoretical Statistics and Mathematics Unit (SMU) and the Electronics and Communication Sciences Unit (ECSU), whose mutual pull didn't take me apart, but rather nourished me twofold. It gave me the privilege of working alongside a diverse group of colleagues. To my coauthors—Sankha Da, who offered pivotal guidance throughout, and Basu, whose computational expertise kept me afloat—I am sincerely grateful. I am much indebted to Prof. Arindam Chatterjee, whose introduction to learning theory and inference in high dimensions sparked a lasting interest in these areas.

The Machine Learning Research Group at ECSU provided an environment rich in intellectual energy and collaboration, and I am thankful to the many people who worked behind the scenes to maintain this culture. The lab, thanks to Kushal and Faizan, quickly became home. To my fellow journeymen from SMU—Deepak, Maity, Sayan, Prabhakar—thank you for making the long days of Delhi and Kolkata feel short and the difficult moments much lighter.

Beyond research, the years in ISI have been made unforgettable by a broader set of experiences and communities. My time here was enriched by the people I met at 205. I'm especially grateful to my fellow boarders at the RS hostel for keeping me grounded and for reminding me, always, what really matters.

To the many friends who helped carry me through—thank you. There are too many of you to name, but certain debts of gratitude are impossible not to mention: to Subha & Co., for all the laughter and joy; to Dipranjan, for the long walks and discussions that never went anywhere but always left a mark; to Jayanta Da, for being the person I could never pigeonhole. Finally, to my family—Ma, Baba, and Didu—thank you for simply being there, always. I could not have done this without you.

Abstract

Generative modeling focuses on the task of producing new data samples that closely resemble those drawn from an original, unknown distribution. Despite being well-known in statistical estimation theory, the approach has gained substantial traction in recent years, driven by groundbreaking results in areas such as image synthesis, natural language generation, and network modeling. The complexity of modern-era data domains and the ensuing adaptations that suitable models must undergo have presented new challenges. These advances raise several fundamental questions, the first of which is: When do generative models accurately approximate the true data distribution? One may also ask: How well do these models perform under contaminated data? This work explores these questions through the lens of generative modeling frameworks that, by design, involve distinct data spaces.

We focus on two major classes of such models that blend optimal transport and representation learning in their objectives: Wasserstein autoencoders (WAE) and Cycle-consistent cross-domain translators. WAE, on its way to regeneration, learns a latent code, which in turn aids the simulation of newer pseudo-random replicates. By providing statistical characterizations of the latent distribution and the transforms inducing a dimensionality reduction in the process, we present a detailed error analysis underlying WAEs. From a non-parametric density estimation perspective, we establish deterministic bounds on the latent and reconstruction errors that adapt to the intrinsic dimensions of input data. We also study the extent of distortion that WAE-generated samples suffer when learned using contaminated data. Key takeaways for practitioners from our analysis include specific architectural suggestions that foster near-perfect sampling.

The framework developed thus far fittingly extends to unpaired cycle-consistent cross-domain models. We show that the sufficient conditions for successful data translation under Sobolev and Hölder-smooth distributions resemble those in the case of WAEs. Our analysis also suggests error upper bounds due to ill-posed transformations and validates the choice of divergences used in objectives.

Finally, in search of a consolidated solution to the robustification problem, we present parallel formulations based on the Gromov-Wasserstein (GW) distance. Due to the equivalence of Gromov-Monge samplers (GW), following GW, and cross-domain translation models, including WAE and GWAE, this answers the second question. We study the robust recovery guarantees, concentration, and tractable computational properties of the newly introduced distance measures under diverse contamination scenarios. We substantiate all our findings based on real-world data in varying generative tasks.

Contents

List of Figures

List of Tables

List of Notations and Acronyms

$\mathcal{X}$	Data space (typically a Borel subset of $\mathbb{R}^d$).
$\mathcal{Y}$	Target data space for cross-domain models (subset of $\mathbb{R}^{d'}$).
$\mathcal{Z}$	Latent space, typically $\mathbb{R}^k$ with $k \leq d, d'$.
x, y, z	Generic elements of $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ respectively.
d, d', k	Dimensions of $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$.
$\mathcal{P}(\mathcal{X})$	Space of Borel probability measures on $\mathcal{X}$.
$\widehat{\mu}_n$	Empirical measure based on n i.i.d. samples from μ.
p_μ, p_ρ	Densities of μ and ρ with respect to Lebesgue measure.
$\mathbb{E}_\mu[\cdot]$	Expectation with respect to measure μ.
$\tilde{\mu}_n$	Smooth estimates of p_μ.
E	Encoder map from $\mathcal{X}$ to $\mathcal{Z}$.
D	Decoder map from $\mathcal{Z}$ to $\mathcal{X}$.
G, T	Generator or transport map between probability spaces.
$f \circ g$	Composition of functions f and g.
$\mathscr{F}(\mathcal{X}, \mathcal{Z})$	Class of measurable functions mapping $\mathcal{X}$ to $\mathcal{Z}$.
$T_{\#}\mu$	Push-forward of measure μ under map T.
$\Pi(\mu, \nu)$	Set of all couplings (transport plans) with marginals μ and ν.
π	A coupling belonging to $\Pi(\mu, \nu)$.
$\|\cdot\|$	Euclidean (ℓ_2) norm unless stated otherwise.
$\|\cdot\|_p$	ℓ_p norm for $p \geq 1$.
$\|\cdot\|_\infty$	ℓ_∞ (supremum) norm.
$\|f\|_{L^p}$	L^p norm of function f.
$L^p(\mathcal{X})$	Space of p-fold Lebesgue-integrable functions on $\mathcal{X}$.

$\sup, \inf$	Supremum and infimum of a set or function.
$\vee, \wedge$	Maximum and minimum over a finite set.
$\mathrm{W}_c^p(\mu, \nu)$	p-Wasserstein distance between μ and ν under metric c.
$\mathrm{GW}(\mu, \nu)$	Gromov–Wasserstein distance between metric-measure spaces.
TGW	Tukey-type robust Gromov–Wasserstein distance.
HGW	Huber-type robust Gromov–Wasserstein distance.
LRGW	Locally Robust Gromov–Wasserstein distance.
$\mathrm{TV}(\mu, \nu)$	Total variation distance.
$\mathrm{JS}(\mu, \nu)$	Jensen–Shannon divergence.
f-divergence	Divergence of the form $\int f\left(\frac{d\mu}{d\nu}\right) d\nu$.
$\mathrm{MMD}_k(\mu, \nu)$	Maximum Mean Discrepancy induced by kernel k.
$k(\cdot, \cdot)$	Positive definite kernel.
$\mathcal{H}_k$	Reproducing Kernel Hilbert Space (RKHS).
$\|\cdot\|_{\mathcal{H}_k}$	Norm in RKHS $\mathcal{H}_k$.
$C_u(\mathcal{X})$	Space of uniformly continuous functions on $\mathcal{X}$.
$\mathcal{C}^s(\mathcal{X})$	Hölder space of order s.
$\mathcal{W}_R^{m,p}(\mathcal{X})$	Sobolev space of order m with integrability p of radius R.
$\mathcal{W}^{s,\infty}(\mathcal{X})$	Sobolev space with bounded weak derivatives.
$\mathcal{B}_{p,q}^s$	Besov space with smoothness s.
$\mathcal{N}(\mathcal{F}, \|\cdot\|, \varepsilon)$	ε-covering number of function class $\mathcal{F}$.
$\mathrm{VC}(\mathcal{F})$	VC-dimension of function class $\mathcal{F}$.
α	Contamination proportion.
ε	Robustification or truncation parameter.
η	Contaminating distribution.
n	Sample size.
$\mathcal{O}(\cdot)$, $o(\cdot)$	Big-O and Little-o asymptotic order notation.
$\lesssim$ ($\gtrsim$)	Inequality up to a universal constant.

Chapter 1

Introduction

1.1 Generative Modeling as Learning a Distribution

The task of generating pseudo-random data from a distribution has been a time-hallowed problem in Statistics. Traditional approaches to generative modeling, such as those based on graphical models (Koller and Friedman, 2009), are grounded in explicitly defined probabilistic assumptions. While these models are easy to interpret and are widely utilized in classical statistics, in time, necessities have evolved, even to the extent of aiming to develop algorithms capable of synthesizing realistic audio, generating coherent text, or producing images aligned with a given description. These tasks are inherently difficult due to the complex and often intractable nature of the underlying data distributions, rendering once-capable techniques (Diggle and Gratton, 1984) ineffective. The primary task of characterizing such data, e.g., a convincing image, from a statistical perspective poses a challenge itself. Fortunately, for many domains of interest, real instances of data lie in abundance. Coupled with the rapid increase in computational capacity, this opens the door to a data-driven approach: instead of specifying the target distribution explicitly, we can aim to learn it from the observed data. Once such a model is trained, it can guide the synthesis of new data samples that reflect the learned structure of the original distribution.

To contextualize, let us revisit what it means to 'learn' a family $\mathcal{P}$ of distributions, given $n \in \mathbb{N}_+$ independent and identically distributed samples $\sim \mu \in \mathcal{P}$. If the family is exactly characterized by a set of parameters, i.e., $\mathcal{P} = \{\mu_\theta : \theta \in \Theta\}$, it suffices to output an estimate $\hat{\theta}$. This approach is much akin to *Explicit* generative modeling and evidently falls apart in the presence of complex, high-dimensional data, such as images. From a non-parametric viewpoint (Tsybakov, 2008; Wasserman, 2006), the goal is recast to output instead an *evaluator*, i.e., a function $\tilde{\mu} : \mathcal{X} \to \mathbb{R}$ such that $\tilde{\mu}(\cdot)$ is an efficient point-wise estimator of $\mu(\cdot)$. The convergence of such estimators can only be made to avoid the curse of dimensionality if the underlying data is supported on an *intrinsic* space, whose dimensionality is much smaller compared to $\dim(\mathcal{X})$ (Kim et al., 2019). Otherwise, one is rather constricted to search for

deterministic pathways to model the sampling process of the target distribution. In other words, given a random seed ω, the goal becomes to find a *generator* $\mathcal{G} : \mathcal{Z} \to \mathcal{X}$ such that the law of $\mathcal{G}(\omega)$, conditioned on the input observations, is approximately μ. Typically, ω is drawn following a distribution supported on $\mathcal{Z}$, which is easy to sample from. In introducing one of the earliest algorithms to generate Gaussian replicates this way, Box and Muller (1958) remark, *"When an electronic computer is used it is desirable . . . rather than to rely on tables"*. An immediate example of such an *Implicit* generative modeling is the Probability Integral Transform (PIT), where $\omega \sim \mathcal{U}(0, 1)$ and $\mathcal{G}$ is defined as the inverse of the distribution function corresponding to a univariate μ. This approach favors modern-day generative tasks since instead of estimating μ, the focus is shifted towards learning a transformation, inducing a *sampler*.

From an information-theoretic perspective, the last two methods are equivalent. Since all the information at one's disposal remains encapsulated in the observed samples from μ, learning an accurate density estimate $\tilde{\mu}$ will allow further sampling. On the other hand, modeling an optimal generator will enable simulating newer observations, and in turn, reducing errors due to subsequent estimations. The difference lies in their computational complexity. Given prior knowledge on the regularity of μ, finding an optimal density estimator boils down to adaptability in terms of the bandwidth (Goldenshluger and Lepski, 2011; Kerkyacharian et al., 1996). Otherwise, the basis functions during the construction of energy estimates need careful adaptation (Cleanthous et al., 2025; Donoho and Johnstone, 1995; Efroimovich, 1986). All such techniques incur little computational cost; however, their poor scalability onto high-dimensional, sparse data domains leads to limited applicability in modern generative modeling. In contrast, generators parametrized by deep neural networks (DNN) approximate ideal transformations onto such data domains with high precision. While the search for frugal training algorithms to curb incurred costs continues, currently, all state-of-the-art samplers originate from this approach.

The empirical success of these models drives theoretical scrutiny of their machinery. This work is primarily motivated to answer the following questions:

- Under which sufficient conditions do deep implicit generative models accurately sample from the target distribution?
- To what extent are these models capable of withstanding contamination in input data, and is there an actionable way to make them robust?

To answer the first question, we investigate whether a suitably characterized $\mathcal{G}$ enables accurate estimation of μ. In the process, we provide detailed prescriptions on how to build an efficient model, given that the input distributions are 'smooth'. The second question is even more crucial as implicit models only have access to a set of samples, and any corruption of the same, in the form of outliers, quickly derails the estimation. To that end, we find cost-

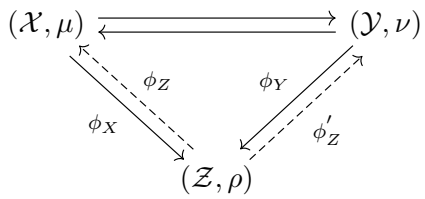


Figure 1.1: Generative models involving diverse spaces.

effective solutions that robustify with minimum alteration to the architecture. As the list of deep generative models (DGM) keeps growing, each being unique in its design, we focus on a special class, namely those involving unlike data spaces.

1.2 The Landscape of Deep Generative Models

Among the plethora of DGMs conceived to date, the ones gaining widespread usage include Generative Adversarial Network (GAN) and its variants (Arjovsky et al., 2017; Goodfellow et al., 2014; Li et al., 2017), GAN-based cross-domain samplers (Kim et al., 2017; Yi et al., 2017; Zhu et al., 2017), Variational Autoencoder (VAE) and its progenies (Higgins et al., 2017; Kingma and Welling, 2014b; Tolstikhin et al., 2018), diffusion models (Ho et al., 2020; Sohl-Dickstein et al., 2015; Song et al., 2021a,b), normalizing flow-based and flow matching techniques (Gat et al., 2024; Lipman et al., 2023; Papamakarios et al., 2021; Rezende and Mohamed, 2015), rectified flow-based and consistency models (Liu et al., 2023; Song et al., 2023). Aiming to cater to diverse tasks, ranging from graph generation (Guo and Zhao, 2022), protein sequence modeling (Tang et al., 2024), speech enhancement (Richter et al., 2023), and text-to-image translation (Zhou et al., 2021), categorization of DGMs is highly subjective and needs to be contextualized. To suit our narrative, we classify them based on the role and nature of $\mathcal{Z}$, namely the *latent space*.

The *first* characterization emerges due to the informativeness of ρ, i.e., the distribution supported on $\mathcal{Z}$ that supplies the random seeds. This is commonly termed the *latent distribution*. In GANs, ρ is typically chosen as a standard Gaussian or Uniform and matches the target distribution μ in terms of the ambient dimensionality (Goodfellow et al., 2014). As a result, the objective of GANs turns out as minimizing the discrepancy $d_{\mathcal{D}}(\mathcal{G}_{\#}\rho, \mu)$, where $d_{\mathcal{D}}(\cdot, \cdot)$ denotes a divergence measure endowed with discriminators $\mathcal{D}$. Evidently, ρ is oblivious to the samples originating from the unknown μ, and can be treated as an uninformative prior. Diffusion (Song et al., 2021b) and consistency (Song et al., 2023) models, built around two interconnected stochastic processes, share a similar trait in their design. The first is a predefined forward (or noising) process $\{\overrightarrow{X}_t\}_{t\in[0,T]}$, which gradually corrupts the sampled data from μ over time, with its marginals denoted by q_t. The goal of this propagation is to reach ρ in distribution. In practice, however, it suffices to approximate a stationary sur-

rogate. The second process is a learnable reverse (or generative) process $\{\overleftarrow{Y_t}\}_{t\in[0,T]}$, with marginals p_t, trained to approximate the time-reversed dynamics of $\{\overrightarrow{X_t}\}$. To generate new data, one samples from the latent distribution ρ and runs the reverse process to obtain Y_T, which is intended to closely resemble samples from μ. Generative models inspired by the VAE dynamic, e.g., Wasserstein AE (WAE) (Tolstikhin et al., 2018), Gromov-Wasserstein AE (Nakagawa et al., 2023), differ in this aspect. Generally, samples from μ are first embedded into $\mathcal{Z}$ $(\phi_{X\#}\mu)$, trying to closely approximate the distribution ρ. The goal of such an encoding lies in preserving the input information to the greatest extent. The latter generative part samples from this encoded law to eventually transform them $((\phi_Z \circ \phi_X)_{\#}\mu)$ into new replicates following μ. Several unpaired cross-domain models also assume the existence of a shared latent space between participating input laws (μ, ν) (Liu et al., 2017a). Instead of direct transform sampling, the problem is reformulated as finding an optimal alignment between $\phi_{X\#}\mu$ and $\phi_{Y\#}\nu$. Clearly, the encoded latent distributions remain informative about the ambient laws.

The *second* characterization, already hinted at in the previous discussion, is based on the dimensionality of the participating spaces. GAN variants and diffusion-based architectures only deal with latent spaces that have the same ambient dimensionality as $\mathcal{X}$. For most applications, e.g., images, word embeddings, $\mathcal{X}$ is a subset of $\mathbb{R}^d$, and by ambient dimension, we refer to d. This implies that the underlying generator transform $\mathcal{G}$ maps onto the same space. In WAEs, however, the latent representation typically demands a dimensionality reduction $(\dim(\mathcal{Z}) < d)$. Such encoded distributions have multiple applications in visualization and clustering. In cross-domain models, the existence of multiple data spaces ($\mathcal{Y} \subseteq \mathbb{R}^{d'}$, given $d \neq d'$), e.g., image-to-image, text-to-image, makes the distinction even more prominent. In the thesis, our focus lies in this latter class of generative models, which involve distinct spaces and informative priors. We observe that a GAN-like framework can be extended to accommodate latent spaces of varying dimensionalities (Bunne et al., 2019), thereby leading to the models of interest in this work.

1.3 Thesis Overview

Most of the earlier theoretical scrutiny into DGMs has been directed towards GANs. Furthermore, it is only recently that similar studies on diffusion models have gained momentum (Benton et al., 2024; Bortoli, 2022; Bruno et al., 2025; Chen et al., 2023a,b; Li et al., 2024; Silveri and Ocello, 2025). This thesis, based on a selection of my works, focuses on finding statistical guarantees for the less-studied models developed between the two.

Regeneration and Latent Space Consistency of Wasserstein Autoencoders

In Chapter 2, based on Chakrabarty and Das (2021); Chakrabarty et al. (2025a), we recast the WAE (WAE-GAN and WAE-MMD) objective as concurrent density estimation tasks using neural-network induced transforms. By introducing the notion of *information preservation* (IP), a simple probabilistic characterization of what an ideal encoding should entail, we prescribe model architectures to foster consistency of estimators in the latent space. The guarantee comes with deterministic upper bounds on the latent loss incurred by WAEs in a non-parametric regime, which adapt to the intrinsic dimensions (e.g., Minkowski, upper Wasserstein) of the input distribution (μ) and allow for ρ to be invariant to group actions. We also show that in case the encoded density is smooth enough (à la Besov), finding an optimal encoder boils down to searching for the minimum distance estimate in the latent space. Moreover, the complexity of the associated Scheffé tournament becomes as large as solving an optimal transport. While decoding using WAE-MMDs, we show that reconstruction is achieved as a consequence of latent consistency under carefully chosen kernels. All of our theoretical findings, from IP to the error rates of convergence, are reflected exactly in numerical experiments based on real and simulated data sets. Finally, we find the extent of additional error one incurs while estimating the target density based on WAE-generated samples in case the input observations suffer Huber contamination.

Translation and Cycle Consistency of Cross-domain Generative Models

In Chapter 3, following Chakrabarty and Das (2022), we extend the non-parametric estimation framework developed in Chapter 2 to unpaired cross-domain generative models imposing cycle-consistency. We analyze the statistical errors involved in translations based on IP networks and their margin due to ill-posedness. Under Sobolev-smooth input laws, we find that using L^1 norm and 1-Wasserstein distance in the cyclic loss tends to be equivalent, attesting to (Zhu et al., 2017)'s observation that the latter does not improve performance. We also demonstrate that it is sufficient to ensure translation consistency to achieve cycle-consistency in total variation, if the translations preserve smoothness. This is significant as the result may not hold in general. Our earlier prescriptions of IP networks are also shown to be adept in cross-domain models.

Robustifying Cross-Domain Generative Models

In Chapter 4, searching for robust solutions to DGMs from the previous chapters, we provide principled methods to robustify Gromov-Wasserstein (GW) distances under a diverse landscape of contamination models (Huber, Wasserstein, etc.). This presents a unique way of metrizing the equivalence class of isomorphic metric-measure (mm) spaces under corruption. Our discussion in this chapter is based on Chakrabarty et al. (2024, 2025b). Drawing

from classical truncation techniques in robust statistics, we propose Tukey and Huber's GW 'distances', which offer outlier-robustness alongside preserving topological (metric) properties of the original metric. They tend to be asymptotically unbiased towards estimating the unperturbed GW value if the number of outliers from both spaces increases at a slower rate compared to the inliers. The algorithmic computations follow a similar complexity to the original and easily adapt to entropic regularization. Provably, Tukey's GW (TGW) becomes a lower bound to existing robust OT metrics (ROBOT (Mukherjee et al., 2021)), given that the distributions are supported on the same ambient metric space. We also propose a lower bound to TGW, which essentially offers stricter robustification at the cost of the usual triangle inequality. The benefits of such a proxy (termed Locally Robust GW) include the relation to a dual formulation that connects the problem to solving a robust OT dual. It extends the framework to probabilistic mm spaces, proposing a way to robustify Sturm's formulation, also leading to a robust cross-domain sampler. As a consequence, we find a way to robustify cycle-consistent cross-domain DGMs. Empirical evidence also shows that our prescribed models surpass existing benchmarks on image synthesis, shape-matching, and barycentric interpolations.

Conclusion

Finally, we summarize our findings from the thesis in Chapter 5, and discuss the impact of our approach in the literature. We also point towards several open problems for future pursuit.

Chapter 2

Regeneration and Latent Space Consistency of Wasserstein Autoencoders

Summary

Amongst the numerous variants Variational Autoencoder (VAE) has inspired, the Wasserstein Autoencoder (WAE) stands out due to its heightened generative quality and intriguing theoretical properties. WAEs consist of an encoding and a decoding network— forming a bottleneck— with the prime objective of generating new samples resembling the ones it was catered to. In the process, they aim to achieve a target latent representation of the encoded data. This chapter offers a comprehensive theoretical understanding of the machinery behind WAEs. From a statistical viewpoint, we pose the problem as concurrent density estimation tasks based on neural network-induced transformations. This allows us to establish deterministic upper bounds on the realized errors WAEs commit, supported by simulations on real and synthetic data sets. We also analyze the propagation of these stochastic errors in the presence of adversaries. As a result, both the large sample properties of the reconstructed distribution and the resilience of WAE models are explored.

2.1 Introduction

Variational Autoencoder (VAE) (Kingma and Welling, 2014a) is one of the earlier agents of modern-day deep generative modeling. Vanilla autoencoders, a precursor to VAEs and conceived primarily to address representation learning, lacked the ability to add stochastic variation in the reconstructed signal. As a result, they could not 'generate' new observations resembling the target. VAEs came into being with the promise of overcoming this limitation, inspiring numerous variants in the process (Wei et al., 2020). Perhaps the one that stirs up a statistician's intrigue the most is the Wasserstein autoencoder (Tolstikhin et al., 2018). Ap-

proaching the problem from an optimal transport (OT) point of view, it achieved significant improvement in generated image quality.

As motivated in the introduction, the discussion on WAEs starts with an unknown target probability distribution μ. The goal lies in simulating new observations from the same by learning it gradually based on samples. Statistically, input observations such as images are often deemed residents of a high-dimensional non-Euclidean space, perhaps manifolds. In our discussion, we surmise that μ, in general, is defined on a Borel subset $\mathcal{X}$ of $\mathbb{R}^d$. This also conforms to the well-known fact that the information necessary to 'represent' an image typically possesses a low-dimensional structure compared to its ambient dimension d (Bengio et al., 2013). There lie two constituents in a typical WAE model: an 'encoder' (E), and a 'decoder' (D). Sampled observations from μ are fed into the encoder, which produces replicates of a low-dimensional representation of the same. As such, it may be viewed as a parametric class of Borel functions from $\mathcal{X}$ to the 'latent space' $\mathcal{Z} \subseteq \mathbb{R}^k$, $d > k$. In practice, both encoders and decoders are parameterized by neural networks (NNs). The goal of encoding is to reach a desired distribution ρ defined on this space, fittingly called the 'latent law'. Evidently, there must remain some discrepancy between the encoded and the desired latent distributions. Tolstikhin et al. (2018) prescribes the usage of Jensen-Shannon divergence (JS) and Maximum Mean Discrepancy (MMD) to encapsulate this 'latent loss'. This quantity makes a major contribution toward the overall objective that drives WAEs. It is also the target latent law that inspires smooth interpolation between modes of μ while generating new observations.

Once the encoding is over, reconstruction must take place. Decoders can be similarly described as the class of functions (mapping $\mathcal{Z} \to \mathcal{X}$) that aim to induce inverse maps to those fostered by the encoders. Encoded observations go through such a transformation in an attempt to get back to where they originally came from, μ. The deviation of the regenerated distribution from the input law makes for the reconstruction error. In a WAE model, this loss is represented by the Wasserstein distance (WD).

2.2 Background

The first instance of a VAE-variant achieving comparable generative performance to that of Generative Adversarial Networks (GAN) came in the form of WAE. Husain et al. (2019) supported this empirical similitude theoretically by showing a primal-dual relationship between the two objectives. However, while statistical scrutiny of VAEs has come a long way, WAEs remain underappreciated in this regard. For example, it is well-known that a VAE model with Gaussian decoders behaves similarly to Robust PCA (Candès et al., 2011) under mild assumptions. As a result, such VAEs are capable of recovering uncorrupted observations hailing from input data manifolds, fending off outliers (Dai et al., 2018). However, the

ability of WAE architectures towards robust reconstruction lies unchecked. The Gaussian assumptions on both the encoder and decoder networks also have a profound impact on the VAE's capability to reconstruct the input law. Dai and Wipf (2019) shows that in case the data manifold has the full ambient dimension, reaching the global minima of the VAE loss is equivalent to ensuring a successful recovery. However, for image data, where the observations typically have a lower-dimensional true representation, non-unique solutions may exist. Similar avenues for WAEs awaited exploration.

In our works (Chakrabarty and Das, 2021; Chakrabarty et al., 2025a), we set out to answer some of these questions. We reformulate the WAE-GAN objective as a minimum distance estimation. Under regularity conditions on networks deployed, we show that f-WAEs (Husain et al., 2019) can recover approximately both latent and reconstruction targets, given that they are smooth. Recently, Chakraborty and Bartlett (2024), in a similar approach, modified the error bounds based on the upper Minkowski dimension of the input distribution. However, no prior work corroborates error convergence rates and regeneration guarantees with simulations based on real datasets.

Contributions. Key highlights of this chapter are as follows:

- We introduce a probabilistic characterization of *information preservation* (IP), which becomes the cornerstone of our depiction of ideal encoders in a WAE model [Section 2.4.1]. We explore divergence metrics that allow IP, which in turn, enables us to prescribe ideal model architectures that foster consistency of estimators in the latent space.
- We establish deterministic upper bounds on the latent loss incurred by WAE-GAN and WAE-MMDs in a non-parametric regime [Theorem 2.4]. The bounds are adaptable to intrinsic dimensions of μ and allow for the target latent law (ρ) to be invariant to group actions. In the process, we explore the desirable properties of underlying kernels in a WAE-MMD setup that promote latent space consistency.
- The reconstruction guarantees we propose [Theorem 2.5, Remark 2.10] assume no regularity of the decoder network and come with accompanying prescriptions of the model architecture. All of our theoretical results are empirically substantiated by numerical experiments based on real and simulated data sets [Section 2.4.3, 2.5.1].
- We additionally examine the effects of *contamination* in input data on reconstructions using WAEs [Section 2.6]. The discussion explores desirable properties of kernel estimates that limit the corruption in translated data and WAEs' inherent capability to offer robustness against distribution shifts.

Organization. Section 2.3 is devoted to basic definitions and the statistical formulation of the WAE problem. It outlines the assumptions on which the forthcoming analysis is based.

Similar characterizations also extend to Chapter 3 and are contextualized topically. Section 2.4 builds toward a statistical guarantee regarding consistent estimation while encoding in WAEs. In the process, we explore a sufficient condition that encoders need to satisfy to promote latent space consistency [Section 2.4.1]. We identify real architectures that preserve information [Example 1, 2, and 3], which we eventually test in simulations [Section 2.4.3]. Our approach addresses the issue of lossy encoding [Remark 2.6] and ties the search for a minimum distance estimate to an OT optimization [Remark 2.7]. The following discussion provides deterministic upper bounds on the reconstruction error incurred by WAEs [Section 2.5], which we also validate based on experiments [Section 2.5.1]. Finally, we check the extent of inherent robustness WAEs possess, given contamination in input data, in Section 2.6. Prioritizing the organization, we place additional figures in the Appendix, along with all proofs of theorems and additional lemmas.

2.3 Preliminaries

We consider the input data space $\mathcal{X}$, equipped with the metric c_x to be Polish. For most real scenarios, a typical characterization of the same is $\mathbb{R}^d$, $d \geq 1$. We refer to the space of probability measures defined on $\mathcal{X}$ as $\mathcal{P}(\mathcal{X})$. We denote the associated sigma-algebra by $\Sigma_{\mathcal{X}}$. The same conventions follow for the latent space $\mathcal{Z} \subseteq \mathbb{R}^k$ $(k < d)$, equipped with the metric c_z. The class of measurable functions mapping $\mathcal{X}$ to $\mathcal{Z}$ is denoted by $\mathscr{F}(\mathcal{X}, \mathcal{Z})$. For ease of understanding, we abbreviate the 'encoder' and 'decoder' transforms as E and D, respectively. Given non-negative real sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$, the suppression of the universal constant $C > 0$, such that $\limsup_{n\to\infty} \frac{a_n}{b_n} \leq C$, is represented as $a_n \lesssim b_n$, or equivalently $a_n = \mathcal{O}(b_n)$. We also denote $x \vee y := \max\{x, y\}$ and $x \wedge y := \min\{x, y\}$.

Definition 2.1 (Push-forward)**.** *Given $f \in \mathscr{F}(\mathcal{X}, \mathcal{Z})$, the push-forward of $\mu \in \mathcal{P}(\mathcal{X})$ is defined as $f_{\#}\mu(\omega) = \mu(f^{-1}(\omega))$, where $\omega \in \Sigma_{\mathcal{Z}}$.*

Definition 2.2 (Integral Probability Metric (Müller, 1997))**.** *For a class of bounded, measurable evaluation functions $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$, the integral probability metric (IPM) measuring the discrepancy between $\mu, \nu \in \mathcal{P}(\mathcal{X})$ is given by*

$$d_{\mathcal{F}}(\mu, \nu) = \sup_{f\in\mathcal{F}} \Big\{ \int_{\mathcal{X}} f(x)d\mu(x) - \int_{\mathcal{X}} f(x)d\nu(x) \Big\}.$$

Remark 2.1. *In our discussion, we frequent a particular variant of this measure, namely, Maximum Mean Discrepancy (MMD). It is obtained by taking $\mathcal{F}$ as the unit ball in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, i.e. $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$. In case the kernel $\kappa(\cdot, \cdot)$ based on a compact metric space that results in $\mathcal{H}$ is continuous— also, dense in the space of bounded continuous functions— the associated MMD becomes a metric (Gretton et al., 2012).*

On the other hand, given that the underlying class of critics $\mathcal{F} := \mathcal{L}^1_{c_x}$, *i.e.* 1*-Lipschitz functions with respect to* c_x, *the* 1*-Wasserstein metric also boils down to an IPM (Villani (2009), Remark 6.5). This duality may not hold in general, which is evident from the definition of the* r*-Wasserstein distance:*

$$W^r_{c_x}(\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \Big\{ \int_{\mathcal{X}\times\mathcal{X}} [c_x(x,y)]^r d\gamma(x,y) \Big\}^{\frac{1}{r}},$$

where $\Gamma(\mu,\nu) = \big\{\gamma \in \mathcal{P}(\mathcal{X}\times\mathcal{X}) : \int_{\mathcal{X}} \gamma(x,y)dy = \mu,\ \int_{\mathcal{X}} \gamma(x,y)dx = \nu\big\}$ *is the set of all couplings between measures* μ *and* ν*;* $r \in [1,\infty)$.

Definition 2.3 (Probability space automorphism)**.** *Let us denote* $X = (\mathcal{X}, \Sigma_{\mathcal{X}}, \mu)$, *where* $\mu \in \mathcal{P}(\mathcal{X})$. *We call* $f : X \to X$ *an automorphism if it admits a measure preserving, essential inverse* f' *such that* $f \circ f' = f' \circ f = id_X$, μ *almost everywhere.*

2.3.1 Problem Setup and Background

Throughout the forthcoming discussion, we denote the input data distribution by μ and that corresponding to the latent space by ρ. Typically, the Lagrangian formulation of the WAE loss is given as

$$\inf_{E\in\mathscr{F}(\mathcal{X},\mathcal{Z})} \big\{W^1_{c_x}(\mu, (D\circ E)_{\#}\mu) + \lambda \cdot \Omega(E_{\#}\mu, \rho)\big\}, \tag{2.1}$$

where $\lambda > 0$ and $\Omega : \mathcal{P}(\mathcal{Z}) \times \mathcal{P}(\mathcal{Z}) \to \mathbb{R}_{\geq 0}$. In general, the resultant encoder E (as in (2.1)) maps each $x \in \mathcal{X}$ to a (conditional) posterior measure in the latent space and only upon rescaling becomes a *probabilistic encoder* (Husain et al., 2019). To establish consistency of plug-in estimates under the empirical WAE-GAN loss (Tolstikhin et al., 2018), it becomes sufficient to consider $\Omega(\cdot,\cdot)$ as the total variation (TV) metric (Chakrabarty and Das, 2021). This is based on the fact that TV acts as an upper bound to the Jensen-Shannon divergence (JS), classically deployed as a regularizer. This modified framework has attracted theoretical intrigue due to its equivalence with the original one under the invertibility of decoders. It is often called the f-WAE (Husain et al., 2019). Building on this density-matching regime, in the current article, we also analyze the WAE-MMD architecture, i.e., when $\Omega \equiv d_{\mathcal{H}}$.

First, let us focus on the set of solutions that bring about zero loss. This is crucial since, during training, practitioners frequently achieve such near-perfect results. However, the solution maps thus obtained may result in noisy reconstructions. It stems from the fact that WAEs essentially try to solve an 'inverse' problem. Our first result suggests that if the latent space admits nontrivial automorphisms, non-unique solutions may exist that achieve zero loss.

Lemma 2.1 (Invariance of zero solutions (Moriakov et al., 2020))**.** *Let* $Z = (\mathcal{Z}, \Sigma_{\mathcal{Z}}, \rho)$. *Also, the encoder-decoder pair* (E, D) *satisfies* $\mathcal{L}_{c_x,\lambda} = 0$ *for a probability divergence* $\Omega(\cdot,\cdot)$ *that*

metrize $\mathcal{P}(\mathcal{Z})$*. Then, given a non-trivial probability space automorphism* $\varphi : Z \to Z$*, the pair* $(\varphi^{-1} \circ E, D \circ \varphi)$ *also brings about zero loss.*

Observe that simply applying an automorphism can result in a multitude of distinct zero solutions. Coupled with the intractability of φ, prescribing an unambiguous solution to a practitioner solving the WAE problem based on neural network-based maps becomes difficult. Moreover, the disentangled representation is sensitive to rotations of the latent embedding (Rolinek et al., 2019). Thus, one needs to do more than just point out solutions that achieve zero loss. Also, when seen from an OT point of view, the existence and consequently, approximation of such non-unique target maps becomes questionable. We elaborate on the same in Section 2.4.1. This brings us to adopting the constrained formulation instead:

$$\inf_{E \in \mathscr{F}(\mathcal{X},\mathcal{Z})} W_{c_x}^1 \left(\mu, (D \circ E)_{\#}\mu\right) \text{ subject to } \Omega(E_{\#}\mu, \rho) \leq t, \tag{2.2}$$

where $t \geq 0$ signifies the tolerable error margin. Although the Lagrangian form (2.1) is dual to (2.2), it is more intuitive to establish non-asymptotic deviation bounds (and hence, consistency) of estimators in the latent space under the latter formulation. For instance, given $t > 0$, it essentially narrows down our search for an ideal E among candidates $\in \mathscr{F}(\mathcal{X}, \mathcal{Z})$ that limit the error incurred during encoding to t. On the other hand, such a deterministic upper bound readily enables checking the sample complexity of an existing encoder architecture to incur an error t. Our proofs of sufficient conditions to achieve regeneration consistency and accompanying prescriptions of networks that capacitate the same also stem from (2.2).

2.3.2 Data Distributions

Typically, WAE-based architectures deal with image data. The statistical construct we follow favors such cases without being restricted to them only. For example, pixel values of raw image data tend to lie in bounded intervals. As such, considering the support of the probability distribution from which they may originate to be bounded seems plausible. Feature-extracted image data also attests to this assumption. A key aspect of input distributions that we are interested in is their regularity. Earlier, we (Chakrabarty and Das, 2021) tested WAEs' ability to reconstruct Hölder densities based on compact supports. Here, let us extend our setup to cater to more diverse distributions.

Let us denote the space of p-fold Lebesgue-integrable functions by $L_p(\mathbb{R}^d)$, endowed with the norm $\|\cdot\|_p$, $p \in [1, \infty)$. The uniform norm is denoted by $\|\cdot\|_\infty$.

Definition 2.4 (Sobolev Space)**.** *Given* $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$*,* $\alpha_i \in \mathbb{N}^+ \cup \{0\}$*, a multi-index such that* $|\alpha| = \sum_{i=1}^d \alpha_i$ *stands for the length, the mixed partial weak differential operator of order* $|\alpha|$ *is given by* $D^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$*. Here,* $x^\alpha = x_1^{\alpha_1} \ldots x_d^{\alpha_d}$ *whenever* $x \in \mathbb{R}^d$*. Under this*

setup, the L^p-Sobolev Space of order m with radius $R \in \mathbb{R}_{\geq 0}$ is defined as

$$\mathcal{W}_R^{m,p}(\mathbb{R}^d) = \Big\{ f \in L_p(\mathbb{R}^d) : D^\alpha f \in L_p(\mathbb{R}^d)\, \forall |\alpha| \leq m : \|f\|_{\mathcal{W}^{m,p}} \equiv \|f\|_p + \sum_{|\alpha|=m} \|D^\alpha f\|_p < R \Big\}.$$

Remark 2.2. *In case $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable at x, set $D^\alpha f = f^{(\alpha)}$, i.e., the classical mixed partial derivative. Also, denote by $C_u(\mathbb{R}^d)$ the space of uniformly continuous functions. This allows us to extend the earlier class for $p = \infty$, namely*

$$\mathcal{W}_R^{m,\infty}(\mathbb{R}^d) = \Big\{ f \in C_u(\mathbb{R}^d) : f^{(\alpha)} \in C_u(\mathbb{R}^d)\, \forall |\alpha| \leq m : \|f\|_{\mathcal{W}^m} \equiv \|f\|_\infty + \sum_{|\alpha|=m} ||f^{(\alpha)}||_\infty < R \Big\}.$$

We find the extension of this class for non-integer $s \in \mathbb{R}_{>0}$, with its integer part $\lfloor s \rfloor$, particularly helpful in the analysis. The following definition formalizes the same.

Definition 2.5 (Hölder-Zygmund Space)**.** *Under the setup described in Remark (2.2), define*

$$\mathcal{C}_R^s(\mathbb{R}^d) = \left\{ f \in C_u(\mathbb{R}^d) : \|f\|_{\mathcal{C}^s} \equiv \|f\|_{\mathcal{W}^{\lfloor s \rfloor}} + \sum_{|\alpha|=\lfloor s \rfloor} \sup_{\substack{x \neq y \\ x,y \in \mathbb{R}^d}} \frac{|D^\alpha f(x) - D^\alpha f(y)|}{|x-y|^{s-\lfloor s \rfloor}} < R \right\}.$$

Let us denote the input density corresponding to μ by p_μ, with respect to the Lebesgue measure.

Assumption 2.1 (Regularity of Input Law)**.** *There exists $m_x \in \mathbb{N}^+$ such that $p_\mu \in \mathcal{W}_R^{m_x,p}(\Omega_x)$, where the support $\Omega_x \subseteq \mathcal{X}$ is compact, $p \in [1, \infty)$.*

This assumption is put in place to give coherence to the discussion so far. We will focus on the case of unbounded domains under varying regularity as generalizations of the initial results. The more challenging of tasks is perhaps characterizing the latent distribution. In our density matching paradigm, it should ideally be a distribution that embodies the dimensionality-reduced representation of p_μ. Let us similarly assume that ρ also has the corresponding density p_ρ.

Assumption 2.2. *p_ρ is based on a compact and convex support $\Omega_z \subseteq \mathcal{Z}$, such that there exists a positive constant c satisfying $\inf_z p_\rho(z) \geq c$.*

The generative aspect of WAEs comes from their capability to simulate novel samples that resemble input observations. The generated set includes smooth interpolations between modal values of μ. As such, the latent law— encapsulating the geometric input information— must distribute positive mass between encoded modes. Tolstikhin et al. (2018) demonstrates the same fact with facial image data. This stems from the idea that the meld between two faces in the Wasserstein geodesic might result in another one, even if unrealistic. The convexity of the support of p_ρ, coupled with its departure from nullity, is a mathematical

representation of the same philosophy. Asatryan et al. (2023) argues that an explicit lower bound to the density can always be found for a slightly modified measure (Remark 3.3). As such, we assume ρ to have a *strong density*. Also, to conform to disentanglement, ρ should ideally have a diagonal or block-diagonal dispersion matrix. In our non-parametric depiction, we keep ample room for such specifications without being restricted to these only.

2.4 Latent Space Consistency

With the foundations laid, let us concentrate on the encoding. In an empirical WAE problem, we only have access to a set of samples $\{X_i\}_{i=1}^n$ drawn i.i.d. from μ. Thus, the sample version of the optimization task (2.2) needs to satisfy the corresponding constraint: $\Omega(E_{\#}\hat{\mu}_n, \rho) \leq t$, given $t \geq 0$. Here, $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is the classical plug-in estimate. We use the same notation to signify the empirical distribution throughout the article. Observe that the resultant set of encoder transforms that fulfill this criterion are indeed functions of n, i.e., $E = E(n)$. In the absence of uniqueness, our suggestions of a *capable* encoder begin with a definition of its chassis: neural networks.

Definition 2.6 (Feed-Forward Neural Networks)**.** *Given $L \in \mathbb{N}^+$, a Neural Network (NN) with L hidden layers is defined as the collection of maps $\phi : \mathbb{R}^{N_0} \longrightarrow \mathbb{R}^{N_{L+1}}$, $\{N_i\}_{i=0}^{L+1} \in \mathbb{N}^+$ given by*

$$\phi(x) := A_L \circ \sigma \circ A_{L-1} \circ \cdots \circ \sigma \circ A_0(x),$$

where $A_i(y) = M_i y + b_i$; $M_i \in \mathbb{R}^{N_{i+1} \times N_i}$ and $b_i \in \mathbb{R}^{N_{i+1}}$ is an affinity, $i = 0, ..., L$. Here, σ signifies the activation function, applied componentwise. Under this setup, we call $W = \vee_{i=1}^L N_i$ the width of the network and L its depth. Denote this collection by $\Phi(W, L)_{N_0}^{N_{L+1}}$. Additionally, the quantity $S = \sum_{i=1}^L N_i$ is said to be the size. A reparameterized version of the same, given as $N(\Phi) = d + S$, denotes the number of neurons in the network.

Remark 2.3. *Based on its simplicity and resilience to vanishing gradients, ReLU ($\sigma(x) = x \vee 0$, also called ramp or first order spline) has emerged as one of the most desired activations to practitioners. However, we highlight the remarkable capability of ReLU-based NNs to approximate regular functions (Petersen and Voigtlaender, 2018; Yarotsky, 2017, 2018). Another activation function that is critical to our analysis is GroupSort (Anil et al., 2019). Given a vector $x = (x_1, \cdots, x_{rn}) \in \mathbb{R}^{rn}$, where $r \geq 2$, it splits the components into n groups each of size r, followed by sorting them in decreasing order, i.e.,*

$$\sigma_r(\underbrace{x_1, \ldots, x_r}_{\text{Group } 1}, \cdots, \underbrace{x_{rn-(r-1)}, \ldots, x_{rn}}_{\text{Group } n}) = \big(x^1_{(r)}, \ldots, x^1_{(1)}, \cdots, x^n_{(r)}, \ldots, x^n_{(1)}\big),$$

where $x^i_{(j)}$ denotes the jth order statistic from the ith group. Preserving all the goodness offered by ReLU, it additionally provides adversarial robustness (Huster et al., 2019). We

highlight that NNs deploying GroupSort (equivalently OPLU, when grouping size is 2 (Chernodub and Nowicki, 2016)) are better suited to universally approximating non-linear Lipschitz maps.

2.4.1 Encoder maps

Encoders are transformations that enforce dimensionality reduction, preserving key properties of μ. Though not obvious, typically, such maps enforce a non-linear reduction due to the non-linearities (activations, e.g., tanh) present in the underlying NN. The process it undergoes is significantly different from the classically known DR techniques. However, in case the maps are assumed to be linear embeddings (decoder as well), latent factors obtained by a VAE tend to align along the Principal Component (PC) directions (Rolinek et al., 2019). Regularized VAEs can also be related to the DR carried out by non-linear ICA (Hyvärinen and Pajunen, 1999) under a parametric regime. The similarity stems from the achievement of identifiability of the parameters characterizing the latent law in both methods (Khemakhem et al., 2020). This departure from traditional techniques forces us to change our viewpoint on DR as we know it. Besides, the encoding in the posterior density-matching setup of WAEs differs even further. Instead of looking at the encoder's capacity to conserve local and broader geometry of the spaces in terms of distances, we focus on its trait to limit *distortions* of estimators. Let us provide a probabilistic definition of the same.

Definition 2.7 (Information Preserving Transform)**.** *Given an estimate $\hat{\mu}_n$ based on n observations from the distribution μ and $\epsilon > 0$, a map $I \in \mathscr{F}(\mathcal{X}, \mathcal{Z})$ is said to be Information Preserving under the distance metric $d(\cdot,\cdot)$ if there exist constants $k_1, k_2 > 0$, such that*

$$\mathbb{P}\Big(d\big(I_{\#}\hat{\mu}_n, \widehat{(I_{\#}\mu)}_m\big) \leq \epsilon\Big) \geq 1 - k_1 e^{-k_2 (n \wedge m)\epsilon^2},$$

where $\widehat{(I_{\#}\mu)}_m$ denotes an estimate of the translated distribution based on $m \in \mathbb{N}^+$ i.i.d. samples.

The immediate question that comes to mind may be, *what are the transformations that behave as IPTs?* Precisely, the answer lies in the regularity of the functions. Though not apparent, the notion of IPTs is intrinsically related to Bourgain's discretization theorem and Lipschitz embeddings. To that end, we first explore the capability of Lipschitz continuous maps— between the input and latent space— to pose as IPTs. Let us denote by $\mathscr{F}_U(\mathcal{X}, \mathcal{Z})$ the class of U-Lipschitz functions mapping $(\mathcal{X}, c_x)$ to $(\mathcal{Z}, c_z)$, $U \geq 0$. So far, we have not imposed any regularity on the class of latent distributions. In such a general setting, the role of the underlying divergence, metrizing the measure space, becomes crucial. In this context, we recall the caution sounded by Sriperumbudur et al. (2009) that the total variation metric ($d_{\mathcal{F}} \equiv d_{\mathrm{TV}}$, where $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$) is often unable to ensure strong consistency of

estimators under them. The issue is rooted in the class of critics $\mathcal{F}$ being 'too large'. The *first* method to circumvent this problem is to look at IPMs employing more precise critics.

Theorem 2.1 (Information Preservation of Lipschitz Encoders)**.** *Let $\mathcal{H}$ denote a class of bounded real-valued functions on $\mathcal{Z}$, such that the associated entropy has at most polynomial discrimination. In other words, there exists $q, A \in \mathbb{R}_{>0}$ such that $\forall \epsilon > 0$, $\log \mathcal{N}\left(\mathcal{H}, \|\cdot\|_\infty, \epsilon\right) \leq A\epsilon^{-q}$. Then for any $g \in \mathscr{F}_U(\mathcal{X}, \mathcal{Z})$ there exists constants l, E_1, E_2 and $E_3 > 0$ such that given $0 < t \leq \frac{2}{3}$,*

$$d_{\mathcal{H}}\left(g_{\#}\tilde{\mu}_n, \widehat{(g_{\#}\mu)}_m\right) \leq t + \frac{lUh^{m_x}}{2} + \mathcal{O}(m^{-\frac{1}{q\vee 2}})$$

holds with probability $\geq 1 - \left(E_1 + E_2(\frac{\sqrt{d}UB_x}{h^{d+1}t})^d\right)\exp\left\{-E_3(m \wedge nh^d)t^2\right\}$, where $\tilde{\mu}_n$ is a density estimate of μ based on the Regularly invariant kernel κ (see Definition 2.8), satisfying $\sup_\kappa \sup_{x\in\Omega_x} \kappa(\cdot,\cdot) \leq 1$, and with bandwidth $h \equiv h_n \searrow 0$.

The theorem implies that Lipschitz transforms can approximately pose as IPTs. By choosing h judiciously, one may show that the approximation error turns $o(1)$ in the asymptotic regime. We deliberately use the smoother kernel density estimate instead of the plug-in to appreciate Assumption 2.1. The choice of the kernel function— as regularly invariant— is of high significance, which the proof (see Appendix) demonstrates. We elaborate on the same while discussing MMDs (Definition 2.8). Note that here, our goal is to demonstrate IP, which eventually leads to guarantees in latent approximation. To search for optimality in terms of kernel estimators, the theorem can be extended to adaptive bandwidth and kernel selection à la Goldenshluger and Lepski (2011). The adaptability improves for added boundary correction (Bertin et al., 2019) on the kernel estimates to remove the so-called boundary bias. Now, the classes $\mathcal{H}$ whose entropy increases polynomially lie in abundance (Nickl and Pötscher, 2007). A particular case that we emphasize on is $\mathcal{L}^1_{c_z}$, i.e. 1-Lipschitz functions with respect to c_z. It is known that $\log \mathcal{N}\left(\mathcal{L}^1_{c_z}, \|\cdot\|_\infty, \epsilon\right) \lesssim \zeta(\mathcal{Z}^1)\epsilon^{-k}$, where $\zeta(\mathcal{Z}^1)$ is the Lebesgue measure of the set $\{z : c_z(z, \mathcal{Z}) < 1\}$ (van der Vaart and Wellner (1996), Theorem 2.7.1). The choice of critics as Lipschitz, also provides a generalization over most Besov functions.

Corollary 2.1. *Given $g \in \mathscr{F}_U(\mathcal{X}, \mathcal{Z})$ and the empirical distribution $\hat{\mu}_n$, there exists a positive constant $E_1^{'}$, such that*

$$d_{\mathcal{L}^1_{c_z}}\left(g_{\#}\hat{\mu}_n, \widehat{(g_{\#}\mu)}_m\right) \leq t + \mathcal{O}(m^{-\frac{1}{k\vee 2}}) + \mathcal{O}(n^{-\frac{1}{d}})$$

holds with probability at least $1 - e^{-E_1^{'}(n\wedge m)t^2}$.

Remark 2.4 (Extension for b-Lipschitz critics)**.** *In case of the divergence $d_{\mathcal{L}^b_{c_z}}(\cdot,\cdot)$, a result similar to that of the earlier theorem can be established based on the fact that $\log \mathcal{N}\left(\mathcal{L}^b_{c_z}, \|\cdot\|_\infty, \epsilon\right) \leq \mathcal{N}\left(\mathcal{Z}, c_z, \frac{\epsilon}{8b}\right)\log\left(\frac{8}{\epsilon}\right)$ (Gottlieb et al. (2017), Lemma 6), given $b > 0$. Observe that it also enables one to remove the boundedness of the support latent class of measures lie on. Instead,*

we may impose milder restrictions, such as having sub-exponential tails (essentially bounded). Specifically, if $\mathbb{E}_q\{\|Z\|\mathbb{1}_{\{\|Z\|>\log(m)\}}\} = \mathcal{O}(m^{-\frac{(\log m)^\delta}{k}})$, *where* $q \in \mathcal{P}(\mathcal{Z})$ *and* $\delta > 0$*; we may recover Corollary 2.1 with only an altered approximation error* $\mathcal{O}(m^{-\frac{1}{k}}(\log m)^{1+\frac{1}{k}})$.

Now, let us focus on the *second* remedy, that being more regulated classes of translated laws. This is crucial since otherwise, the convergence of empirical measures under TV might become arbitrarily slow (Devroye and Gyorfi, 1990). To that end, let us first recall the notion of Yatracos classes (YC) (Devroye and Lugosi (2001), Chapter 6). Given $\mathcal{F} : \mathcal{Z} \to \mathbb{R}$, the Yatracos class associated to it is the set system $\{z \in \mathcal{Z} : f(z) \geq g(z);\ f, g \in \mathcal{F}\}$, denoted by $\mathcal{Y}(\mathcal{F})$. In other words, it characterizes the domain in terms of candidates in a Scheffé tournament defined on it. Our next result suggests that if the Vapnik-Chervonenkis (VC) dimension of the YC corresponding to the family of latent distributions is finite, we can recover Theorem 2.1. The proposition becomes quite intuitive following the maximal packing argument of Van Handel (2014), Theorem 7.16.

Corollary 2.2. *Let the VC-dim*$[\mathcal{Y}(\mathcal{P}(\mathcal{Z}))] = v_z < \infty$. *Then for any* $g \in \mathscr{F}_U(\mathcal{X}, \mathcal{Z})$ *and* $0 < t \leq \frac{2}{3}$ *the constants* E_1, E_2 *and* $E_3 > 0$ *are retained such that*

$$d_{TV}\left(g_{\#}\tilde{\mu}_n, \widehat{(g_{\#}\mu)}_m\right) \leq t + \frac{lUh^{m_x}}{2} + \mathcal{O}(\sqrt{v_z}m^{-\frac{1}{2}})$$

holds with probability $\geq 1 - \left(E_1 + E_2(\frac{\sqrt{d}UB_x}{h^{d+1}t})^d\right)\exp\left\{-E_3(m \wedge nh^d)t^2\right\}$, *where* $\tilde{\mu}_n$ *is a Regularly Invariant kernel (RIK) density estimate of* μ *(see Definition 2.8).*

There are two key highlights of the latest result that turn out to be indispensable in the upcoming discussion on latent consistency. The first aspect we emphasize is the tail condition of the target law. Sub-exponential is a fairly general notion in the sense that all bounded and sub-Gaussian distributions fall under its umbrella. Moreover, all results obtained under such a characterization can be directly extended to sub-Weibull distributions (Vladimirova et al., 2020). In practice, WAEs are mostly trained with multivariate Gaussian as conjugate prior (and hence, posterior) latent laws (Tolstikhin et al., 2018), which also conforms to our argument. The second facet— arguably the cornerstone of the analysis by Chakrabarty and Das (2021), and responsible for controlling the complexity of the underlying space— is the quantity VC-dim$[\mathcal{Y}(\cdot)]$. The finiteness assumption on the same is frequented in density estimation (Ashtiani et al., 2018; Jain and Orlitsky, 2020) solely based on its plausibility. It is known that the class of k-dimensional Gaussian distributions have VC-dim$[\mathcal{Y}(\cdot)] = \mathcal{O}(k^2)$. The same corresponding to axis-aligned densities hailing from k-variate exponential families turn out to be $\mathcal{O}(k)$ (Devroye and Lugosi (2001), Theorem 8.1).

Remark 2.5. *The IP bounds (Theorem 2.1, Corollary 2.2) suffer from the curse of dimensionality due to the usage of kernels that are oblivious to intrinsic structures of the data*

support. In real high-dimensional datasets, especially images, the manifold hypothesis is often found empirically valid. If μ has a finite volume dimension $< d$, defined as

$$d_{\text{vol}}(\mu) \coloneqq \sup\Big\{\gamma \geq 0 \;:\; \limsup_{r\to 0} \sup_{x\in\mathcal{X}} \frac{\mu\left(B_{\mathbb{R}^d}(x,r)\right)}{r^\gamma} < \infty\Big\},$$

the error rates can be shown to adopt instead d_{vol} under additional integrability assumptions on the underlying kernels (Kim et al., 2019). This particularly occurs if μ has an effectively intrinsic manifold support.

Before moving on to further examples of IPTs, let us examine the worth of NN-based maps in the same context. Observe that, the transformations $A_i(\cdot)$ (see Definition 2.6) can be easily shown to be Lipschitz continuous by limiting $\|M_i\|_p = \sup_{\|y\|_p=1} \|M_i y\|_p \leq t$, given $t > 0$. Anil et al. (2019) gave simple recipes to preserve such norms in case $p = 2$ and ∞. Given this fact, coupled with the Lipschitz continuity of activation functions (e.g., ReLU, leaky ReLU, GroupSort, tanh, sigmoid, etc.) typically applied, it is not difficult to show NN-transforms to be exactly so. However, not all such $\sigma(\cdot)$ preserve gradient norms under composition without additional regularization (e.g., ReLU). Furthermore, recovering the exact Lipschitz constant, and hence the map, often turns out to be NP-hard (Virmaux and Scaman, 2018). So, instead, we harness the approximation capabilities of deep NNs to our aid. ReLU has attracted the most attention in this regard (Chen et al., 2019; Daubechies et al., 2022; Gribonval et al., 2022; Montanelli and Du, 2019; Suzuki, 2019). To motivate our next result, we present a simple observation:

Lemma 2.2. *Given $\mu_1, \mu_2 \in \mathcal{P}(\mathcal{X})$ and $\phi \in \Phi(W, L)_d^k$, under arbitrary IPM $d_{\mathcal{F}}(\cdot,\cdot)$, such that $\mathcal{F} = \{f : \mathcal{Z} \to \mathbb{R}\}$ is symmetric, we obtain*

$$d_{\mathcal{F}}(\phi_{\#}\mu_1, \phi_{\#}\mu_2) - d_{\mathcal{L}^1_{c_z}}(g_{\#}\mu_1, g_{\#}\mu_2) \leq 2\big[\inf_{g\in\mathscr{F}(\mathcal{X},\mathcal{Z})} \|\phi - g\|_\infty + \mathcal{E}(\mathcal{F}, \mathcal{L}^1_{c_z})\big],$$

where $\mathcal{E}(\mathcal{F}, \mathcal{L}^1_{c_z}) = \sup_{f\in\mathcal{F}} \inf_{l\in\mathcal{L}^1_{c_z}} \|f - l\|_\infty$ denotes the essential disagreement between classes of critics and $c_z \equiv L_1$.

Observe that the statement holds true under arbitrary choices of the second class of evaluation functions. We mention $\mathcal{L}^1_{c_z}$, in particular, to continue our discussion in light of Corollary 2.1. The result suggests that it is sufficient for a feed-forward NN-induced function to approximate Lipschitz maps (between the input and latent space) to behave like an IPT. Under the TV metric, the proof becomes much simpler based on the fact that $d_{\text{TV}}(\phi_{\#}\mu_1, \phi_{\#}\mu_2) \leq d_{\text{TV}}(\mu_1, \mu_2)$, for $\phi \in \mathscr{F}(\mathcal{X}, \mathcal{Z})$. However, difficulties may arise in case the underlying distance measure is MMD. Let us first go through some rudiments of kernel methods to facilitate our investigation on the same.

2.4.2 MMD and Kernel Mean Embedding

The well-known Riesz representation theorem ensures the existence of a unique representer $K(x) \in \mathcal{H}$, such that $\forall f \in \mathcal{H}, f(x) = \langle f, K(x)\rangle$ for all $x \in \mathcal{X}$. In this setup, the function $\kappa(x,y) = \langle K(x), K(y)\rangle$ is called the *reproducing kernel* of $\mathcal{H}$. The opposite characterization also holds. By Aronszajn's theorem, given a symmetric, positive definite κ on $\mathcal{X} \times \mathcal{X}$, there exists a unique RKHS $\mathcal{H}_\kappa$. This inspires us to meaningfully narrow down our focus on the distributions $\mathcal{P}_\kappa = \{\eta \in \mathcal{P} : \int \sqrt{\kappa(x,x)}|\eta|(dx) < \infty\}$. The MMD between two of such laws μ_1, μ_2 can be rewritten as $d_{\mathcal{H}_\kappa}(\mu_1,\mu_2) = \|K(\mu_1) - K(\mu_2)\|_{\mathcal{H}_\kappa}$, i.e. the Hilbert space distance between the respective kernel mean embeddings (KME), given by $K(\eta)(x) = \int \kappa(x,y)\eta(dy)$. For a detailed exposition, one may turn to Sriperumbudur et al. (2010). Since it is the kernel function that determines the nature of the RKHS, we introduce some regularities, which in turn aid our cause.

Definition 2.8 (Regularly Invariant Kernels)**.** *A measurable function $\kappa(x,y) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be regular, if for $N \in \mathbb{N}$ it satisfies*

(i) $\int_{\mathcal{X}} \sup_{v\in\mathcal{X}} |\kappa(v, v-u)||u|^N du < \infty$, *and*

(ii) $\int_{\mathcal{X}} \kappa(v, v+u)du = 1; \int_{\mathcal{X}} \kappa(v, v+u)u^\alpha du = 0$ *for every* $v \in \mathcal{X}$ *and* $|\alpha| = 1, ..., N-1$.

If such a kernel additionally satisfies the weaker invariance property:

$$\left\{\int |\kappa(w,v) - \kappa(w,u)|^r dw\right\}^{\frac{1}{r}} = \mathcal{O}\left(\|v-u\|\right),$$

given $r \geq 1$; we say it is regularly invariant.

Observe that most kernel functions based on standard probability distributions with finite and centered moments will tend to be regular. Moreover, an immediate example of our version of invariant would be Energy kernels: $\kappa_\alpha(u,v) = \|u\|^{2\alpha} + \|v\|^{2\alpha} - \|u-v\|^{2\alpha}$, $u, v \in \mathcal{X}$ and $\alpha \in (0,1)$ (Modeste and Dombry, 2024). For further coherence, we introduce the notion of *strong invariance*. A kernel function is called strongly invariant if the following holds: $\|K(u) - K(v)\| = \mathcal{O}\left(\|u-v\|\right)$. This, in turn, implies weak invariance based on the fact that

$$\left\{\int |\kappa(w,v) - \kappa(w,u)|^r dw\right\}^{\frac{1}{r}} \leq \|K(v) - K(u)\| \Big[\int \|K(w)\|^r dw\Big]^{\frac{1}{r}}.$$

For example, in case of Energy kernels, $\|K(u) - K(v)\| = 2\|u-v\|^\alpha$. Based on such functions, MMD indeed promotes exponential decay in information dissipated while encoding. To set the stage for a supporting mathematical argument, let us first notice that given $\mu \in \mathcal{P}_\kappa(\mathcal{X})$,

$$d^2_{\mathcal{H}_\kappa}(\hat{\mu}_n, \mu) = \langle K(\hat{\mu}_n - \mu), K(\hat{\mu}_n - \mu)\rangle_{\mathcal{H}_\kappa} = \int_{\mathcal{X}\times\mathcal{X}} \kappa(x,y)(\hat{\mu}_n - \mu) \otimes (\hat{\mu}_n - \mu)(dxdy).$$

Also, $\mathbb{E}\left[d_{\mathcal{H}_\kappa}(\hat{\mu}_n, \mu)\right] \leq \sqrt{\frac{2}{n}} \sup_{x\in\Omega_x} \kappa(x,x)^{\frac{1}{2}}$.

The next result contextualizes our search for NN-induced maps promoting IP.

Theorem 2.2 (Information preservation under MMD). *Let $\mu \in \mathcal{P}_\kappa(\mathcal{X})$, where $\kappa(\cdot,\cdot)$ is a strongly invariant kernel satisfying $\sup_{z\in\Omega_z} \kappa(z,z) \leq C_\kappa$, such that $C_\kappa > 0$. Given $g \in \mathscr{F}_U(\mathcal{X},\mathcal{Z})$, there exists $\phi \in \Phi(W,L)_d^k$ which implies that*

$$d_{\mathcal{H}_\kappa}\left(\phi_{\#}\hat{\mu}_n, \widehat{(\phi_{\#}\mu)}_m\right) \leq (m\wedge n)^{-\frac{1}{2}}\sqrt{\frac{B\ln(\frac{2}{\delta})}{2}} + \sqrt{\frac{2C_\kappa}{m}} + \sqrt{D_n}\underbrace{\|\phi - g\|_\infty^{\frac{1}{2}}}_{(*)}$$
$$+ \sqrt{\mathcal{O}(c_{g,n}(d^2n)^{-\frac{1}{d}}) + U(m\wedge n)^{-\frac{1}{2}}c_{g,n}\sqrt{\frac{B\ln(\frac{2}{\delta})}{2}}}$$

holds with probability at least $1-\delta$, $\delta > 0$. Here, B is a positive constant dependent on C_κ, and both D_n and $c_{g,n}$ are sequences based on n that $\searrow 0$ almost surely as $n \to \infty$.

Observe that the quantity $(*)$ in Theorem 2.2 acts as an upper bound to the departure of an NN-induced map from its exemplar g. The following examples look for the sharp upper bounds on $(*)$ under different circumstances.

Example 1 (Information Preservation of ReLU Encoders (Shen et al., 2019)). There exists $\phi \in \Phi(W,L)_d^k$ based on ReLU activations with $W = \mathcal{O}(d\lfloor N_1^{\frac{1}{d}}\rfloor \vee N_1 + 1)$ and $L = \mathcal{O}(N_2)$, that satisfy Theorem 2.2 for any $N_1, N_2 \in \mathbb{N}^+$, such that $(*) = \mathcal{O}(\sqrt{d}UB_x N_1^{-\frac{2}{d}} N_2^{-\frac{2}{d}})$. Here, $B_x :=$ diameter of Ω_x with respect to the metric c_x. □

Example 2 (Information Preservation of GroupSort Encoders (Tanielian and Biau, 2021)). Given $\varepsilon > 0$, there exists a GroupSort NN induced map $\phi \in \Phi(W,L)_d^k$ of depth $L = \mathcal{O}(d^2\log_2(\frac{2\sqrt{d}}{\varepsilon}))$ and size $S = (\frac{2\sqrt{d}}{\varepsilon})^{d^2}$, also satisfying $\|M_0\|_{2,\infty} = \sup_{\|x\|=1}\|M_0 x\|_\infty \leq 1$, $\max\{\|M_i\|_\infty; i = 1,\cdots,L\} \leq 1$ and $\max\{\|b_j\|_\infty; j = 0,\cdots,L\} \leq \infty$, such that $(*) = \mathcal{O}(\varepsilon)$. □

Example 3 (Barron Functions as IPT). Based on our previous discussion, it becomes evident that being Lipschitz continuous is a desirable property for encoder transforms to behave as IPTs, approximately at the least. This very fact accentuates the importance of the class of functions known as *Barron functions*. While there exist several characterizations of the same (Wojtowytsch et al., 2022), we keep to the following definition.

Definition 2.9 (Barron Class (Caragea et al., 2020)). *A function $f : \Omega_x(\subset \mathbb{R}^d) \to \mathbb{R}$ is said to belong to Barron class with constant $C > 0$ (say, $\mathcal{B}_C(\Omega_x)$), if the exists $x' \in \Omega_x$ and a measurable function $g : \mathbb{R}^d \to \mathbb{C}$ such that $\forall x \in \Omega_x$ both the conditions*

1. $\int_{\mathbb{R}^d} \sup_{x\in\Omega_x} |\langle \eta, x - x'\rangle||g(\eta)|d\eta \leq C$ *and*
2. $f(x) - c = \int_{\mathbb{R}^d}(e^{i\langle x,\eta\rangle} - e^{i\langle x',\eta\rangle})g(\eta)d\eta$

are satisfied, where $|c| \leq C$.

For vector-valued functions $f : \Omega_x \to \mathbb{R}^k$, in which we are mostly interested, the same criteria need to be satisfied componentwise. Lee et al. (2017) provides an equivalent definition (also based on Fourier inversion) of the Barron class. It is intriguing to observe that $\mathcal{B}_C$ embeds continuously into the class of real-valued Lipschitz maps under the L_1 metric (Wojtowytsch et al. (2022), Theorem 3.3). As such, Barron functions are naturally prone to preserving information while serving as encoders (Theorem 2.1). Now, observe that the real architectures suggested in the context of IPT so far might suffer from the curse of dimensionality. Also, they both tend to be deep, scaling exponentially with the input data dimension. While this serves our purpose in the asymptotic regime, shallow networks ($L = 1$) might be of greater priority to practitioners.

Meanwhile, the Barron class can be shown to accommodate all finite norm-bounded neural networks and their limits (Wojtowytsch et al., 2022). This is crucial since it allows one to demonstrate shallow networks' capability to act as an IPT approximately. Barron (1993), in his seminal paper, first proved that a function $f \in \mathcal{B}_C$ can be approximated up to arbitrary accuracy by a shallow NN deploying sigmoidal activations.

Definition 2.10. *A bounded measurable function* $h : \mathbb{R} \to \mathbb{R}$ *is said to be sigmoidal if* $\lim_{x\to-\infty} h(x) = 0$ *and* $\lim_{x\to\infty} h(x) = 1$. *Common examples of such activation functions include logistic, hyperbolic tangent, and* $h(x) = ReLU(x) - ReLU(x-1)$.

In other words, there exists $\phi \in \Phi(W, 1)_d^1$ with $N(\Phi) = m$, such that $\{\int_{\Omega_x}(f(x) - \phi(x))^2\mu(dx)\}^{\frac{1}{2}} = \|f - \phi\| \lesssim m^{-\frac{1}{2}}$. In the asymptotic regime, however, to achieve an infinitesimally small error, the map ϕ approaches an infinitely wide NN. Wojtowytsch et al. (2022) draws the same conclusion based on μ having a finite second moment (Theorem 3.8). This approximation can be further improved in case μ has bounded support, conforming to Assumption 2.1. Klusowski and Barron (2018) shows the existence of such a ϕ, based on general Lipschitz activations, that ensures $\|f - \phi\|_\infty \lesssim \sqrt{d + \log m}\, m^{-\frac{1}{2}-\frac{1}{d}}$. Despite being the highlight, our discussion on the approximation of Barron functions is not limited to shallow networks. The information-preserving behavior of Barron maps stays intact under compositions as well. This is again rooted in them being Lipschitz continuous. As an immediate consequence, we find an alternative avenue to show that sigmoid-activated deep encoders too act as IPTs.

Theorem 2.3 (Information Preservation of Sigmoidal Encoders (Lee et al., 2017))**.** *Let* $\{N_i\}_{i=0}^{L+1} \in \mathbb{N}^+$ *be a sequence of intermediate dimensions as given in Definition 2.6, where* $N_0 = d$ *and* $N_{L+1} = k$. *Also let* $f_i : \mathbb{R}^{N_{i-1}} \to \mathbb{R}^{N_i}$ *such that* $f_1 \in \mathcal{B}_{C_0}(\Omega_0)$ *and for* $2 \leq i \leq L+1$ *and a given parameter* $s > 0$, $f_i \in \mathcal{B}_{C_{i-1}}(\Omega_{i-1}^{s,N_{i-1}})$. *Here* $\{C_i\}_{i=0}^{L} > 0$ *and* $\Omega_{i-1}^{s,N_{i-1}} := \{y = y_1 + y_2 : y_1 \in \Omega_{i-1}, y_2 \in B^{N_{i-1}}(s)\}$, $B^{N_{i-1}}(s)$ *being the* L_2 *ball of radius* s *in* $\mathbb{R}^{N_{i-1}}$, *that satisfy* $f_i(\Omega_{i-1}) \subseteq \Omega_i$, $1 \leq i \leq L+1$. *In particular, define* $\Omega_0 \equiv \Omega_x$ *and* $\Omega_{L+1} \equiv \Omega_z$. *Under this*

setup, given $f_{1:L+1} := f_1 \circ f_2 \circ \cdots \circ f_{L+1}$ *and* $\varepsilon > 0$*, there exists an* L*-deep sigmoid-activated* ϕ*, with* $\mathcal{O}(N_i \varepsilon^{-2})$ *nodes in the* i*th layer, such that* $(*) = \mathcal{O}(\varepsilon)$. □

With the desired regularity of an ideal encoder specified, we move closer to testing whether the constraint, as in (2.2), is met in a sample problem. While our probabilistic notion based on estimators provides a view into the solution, encoding might also be seen in the light of local geometry. This not only provides further clarity to the definition of *representation* but also opens up a pathway to understanding decoding in greater detail. Observe that the process of encoding seeks to learn embeddings onto the latent space. A meaningful characterization of the same might be a map that aims to preserve pairwise distances between discrete points (the sample set) residing in the input space (Courty et al., 2018). It turns out to be crucial for techniques that rely on the latent space to identify clusters (Jiang et al., 2017). Such a map $E : \Omega_x \to \Omega_z$ satisfying $ac_x(x, y) \leq c_z(E(x), E(y)) \leq Ac_x(x, y)$ is said to be bi-Lipschitz (BL) with distortion $D(E) = \frac{A}{a} < \infty$. These immediately satisfy IP. Moreover, the inverse of such embeddings turns out to be Lipschitz as well. We will later see that this fact can be exploited to aid in efficient decoding. Another feature that stands out is that E restricts the encoded law from being degenerated at a point. Now, the immediate question that arises is whether such embeddings exist. The first affirmative evidence was presented by Johnson and Lindenstrauss (1984), taking both c_x and c_z to be L_2 in their respective spaces.

Lemma 2.3 (Johnson-Lindenstrauss Embedding)**.** *Given a set of size* n *from* Ω_x *and* $0 < \varepsilon < \frac{1}{2}$*, there exists a Bi-Lipschitz map* $E : \mathbb{R}^d \to \mathbb{R}^k$ *with distortion* $(1 + \varepsilon)$*, such that* $k = \mathcal{O}\big(\frac{\log(n)}{\varepsilon^2}\big)$.

This result additionally ties the extent of distortion to the optimal value of the latent dimension k. We include an expository note on this in the Appendix. The bound in the lemma turns out to be optimal up to constant factors (Larsen and Nelson, 2017). Later, Bourgain (1985) also showed the existence of BL transforms that achieve encoding onto $k = \mathcal{O}(\log^2 n)$ with distortion $\lesssim \log(n)$. In our case, however, it is sufficient to show the existence of Lipschitz benchmarks. To that end, we turn to the following result.

Lemma 2.4 (Bartal et al. (2011))**.** *Given* $X \subseteq \Omega_x$*, for every* $\varepsilon > 0$*, there exists a finite* 1*-Lipschitz embedding* $E : X \to \mathbb{R}^k$ *such that* $k = \mathcal{O}\left(\varepsilon^{-2} d^*(X)\left(\frac{\log(d^*(X))}{\varepsilon}\right)\right)$*, where* $d^*(X)$ *is the doubling dimension*[1] *of* X.

This is an exact deterministic answer to our search for an ideal encoder. Such a map can immediately be extended to the whole input space using Kirzbraun's theorem. Now, to show latent space consistency, first observe that given a metric Ω and encoder E, the realized

[1]The doubling dimension is defined as $d^*(X) = \log_2 \lambda$, where $\lambda \geq 1$ (doubling constant) is the smallest number such that at most λ balls of half radius are needed to cover every ball in X.

latent loss turns out to be $\Omega(E_{\#}\hat{\mu}_n, \rho)$. It can be fragmented into the following parts based on the independent sources of variation:

$$\Omega(E_{\#}\hat{\mu}_n, \rho) \leq \underbrace{\Omega(E_{\#}\hat{\mu}_n, \widehat{(E_{\#}\mu)}_m)}_{\text{Information dissipated}} + \underbrace{\Omega(\widehat{(E_{\#}\mu)}_m, \rho)}_{\text{Estimation error}}. \tag{2.3}$$

While a suitably chosen IPT takes care of the first part, the latter embodies the error committed trying to estimate ρ using samples from the encoded law. In a *lossless encoding* ($E_{\#}\mu =_d \rho$), it will boil down to the usual estimation error. Otherwise, one might be left with a surplus error due to the discrepancy between $\widehat{(E_{\#}\mu)}_m$ and a certain $\hat{\rho}_m$. In the light of Corollary 2.2, a non-asymptotic upper bound on the same based on the empirical Yatracos minimizer of $\hat{\rho}_m$ can be immediately obtained (Chakrabarty and Das (2021), Theorem 1). This demands the finiteness of VC-dim$[\mathcal{Y}(\cdot)]$ corresponding to $\mathcal{P}(\mathcal{Z})$. Before stating further results, let us look at the 'lossless' setting itself.

Remark 2.6 (Lossless encoding in WAEs)**.** *As a phenomenon, this occurs only when E is an exact measure-preserving map. Since we allow the distributions to have densities, the idea translates to having a change of variables given as $\int_{\Omega_z} p_\rho(z)dz = \int_{\Omega_x} (p_\rho \circ E)(x)\,[J_E(x)]dx$, where in general, $J_E(x)$ is the generalized Jacobian at x (Chiappori et al., 2017). For the exact form the transported density achieves under the co-area formula, we refer the reader to McCann and Pass (2020), Section 2. In general cases concerning transformations between variables, the surplus multiplicand is rather vol$[J_E(x)]$:= product of singular values of the $k\times d$ Jacobian matrix J_E (Ben-Israel, 1999). This boils down to the more familiar $|det(\nabla E(x))|$, given $E : \mathbb{R}^d \to \mathbb{R}^d$ is injective on its domain and continuously differentiable. Consequently, our desired distribution p_ρ turns out to be the density corresponding to a k-marginal of $E_{\#}\mu$* [2]*. The existence, let alone the regularity of such a map is not automatically guaranteed. In case both μ and ρ are nonatomic, such that μ vanishes on all Lipschitz $(d-1)$-surfaces, there exists a unique E (μ a.e.), that offers a lossless encoding (McCann, 1995). Brenier's theorem (Brenier, 1991) argues the same under the additional assumption that the distributions have finite variance. Though obtained under restrictive scenarios, a Brenier map pushing forward the standard Gaussian measure to a uniformly log-concave target distribution would be locally Lipschitz (Caffarelli, 2000; Colombo and Fathi, 2021). While this belongs to the class of possible encoders under a special case, it is rather challenging to verify that minimizing the WAE loss attains such a solution. Moreover, given a sample (semi-discrete) problem like ours, the optimal map is likely to be discontinuous. Even if they are not, the injectivity will be sacrificed when the supports are unbounded, since they cannot be continuously embedded at the same time.*

[2] Such transformations can merely be of the Rosenblatt type. Given that $p_\rho \in \mathcal{C}_R^{m_z}(\Omega_z)$, for some $m_z \in \mathbb{R}_{>0}$; E can be shown to be smooth in the sense of Hölder-Zygmund (Asatryan et al., 2023).

The practical intractability of lossless encodings compels us to focus on finding the tolerable margins of associated losses instead. The exploration of deterministic upper bounds rests on the decomposition (2.3). To obtain the same in the case of WAE-GANs, first notice that given $\rho_1, \rho_2 \in \mathcal{P}(\mathcal{Z})$, equipped with corresponding densities such that $\rho_1 \ll \rho_2$,

$$\mathrm{JS}(\rho_1, \rho_2) \leq \left[\pi \ln\left(\frac{1}{\pi}\right) + (1-\pi)\ln\left(\frac{1}{1-\pi}\right)\right] d_{\mathrm{TV}}(\rho_1, \rho_2) \leq \ln(2)\, d_{\mathrm{TV}}(\rho_1, \rho_2)$$

(this form is also called the *Information Transmission Rate* (Topsoe, 2000)), $0 \leq \pi \leq 1$. While easier to calculate, dealing with JS from a density estimation perspective is often technically challenging. By convention, it is assigned the value $+\infty$ in case the underlying distributions do not have densities, and as a result, it does not metrize $\mathcal{P}(\mathcal{Z})$ in general. However, when taken under the square root, JS follows the triangle inequality (Endres and Schindelin, 2003). Now, given a Lipschitz encoder E, let us consider the realized latent loss under JS-divergence due to the RIK estimator $\tilde{\mu}_n$ (as discussed in Theorem 2.1), defined as $\frac{d\tilde{\mu}_n}{dx} = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x}{h}, \frac{x_i}{h}\right)$, $x \in \Omega_x$ where $h_n \to 0$. Fragmenting the same based on unique sources of variation yields,

$$f(\pi)^{-1}\,\mathrm{JS}(E_{\#}\tilde{\mu}_n, \rho) - \Delta_{E,n} \leq d_{\mathrm{TV}}\left(E_{\#}\tilde{\mu}_n, \widehat{(E_{\#}\mu)}_n\right) + d_{\mathrm{TV}}\left(\widehat{\rho}_n, \rho\right), \tag{2.4}$$

given that $\Delta_{E,n} = d_{\mathrm{TV}}\big(\widehat{(E_{\#}\mu)}_n, \widehat{\rho}_n\big)$, which essentially (in asymptotic regime) determines how much latent regularization can be tolerated. As such, the choice of the optimal encoder must be the one that minimizes $\Delta_{E,n}$. It coincides with the minimum distance estimator (Devroye and Lugosi (2001), Theorem 6.4) of ρ amongst encoded candidates ($E_n^* = \operatorname{argmin} d_{\mathrm{TV}}(\widehat{(E_{\#}\mu)}_n, \widehat{\rho}_n)$). Since the other error terms shrink arbitrarily asymptotically ($n \to \infty$ with fixed d, k), given $t \in \mathbb{R}_{>0}$ (as in 2.2), we only need to ensure that $\Delta_{E_n^*,n} < t$.

Remark 2.7 (Cost to the Scheffé Tournament)**.** *Given the smoothness of the target distributions, substituting the plug-in estimates $\widehat{(E_{\#}\mu)}_n$ and $\widehat{\rho}_n$ with smoother alternatives might be computationally beneficial. A successful search for the minimum distance-achieving encoder takes quadratic time (Acharya et al., 2014). Instead, let us consider estimators $\widetilde{(E_{\#}\mu)}$ and $\tilde{\rho}$ respectively, such that $\frac{d\widetilde{(E_{\#}\mu)}}{dz}, \frac{d\tilde{\rho}}{dz} \in L_1(\mathbb{R}^k)$. The benefit is rooted in viewing the problem from an OT standpoint. Let us write,*

$$\begin{aligned} 2d_{TV}\big(\widetilde{(E_{\#}\mu)}_n, \tilde{\rho}_n\big) &= ||\widetilde{(E_{\#}\mu)}_n - \tilde{\rho}_n||_1 \\ &\leq \underbrace{||\widetilde{(E_{\#}\mu)}_n - K_h(\widetilde{(E_{\#}\mu)}_n)||_1}_{(i)} + \underbrace{||K_h(\widetilde{(E_{\#}\mu)}_n) - K_h(\tilde{\rho}_n)||_1}_{(ii)} + \underbrace{||K_h(\tilde{\rho}_n) - \tilde{\rho}_n||_1}_{(iii)}, \end{aligned} \tag{2.5}$$

where given $\rho_1 \in \mathcal{P}(\mathcal{Z})$,

$$K_h(\rho_1) = \int_{\Omega_z} K_h(.,y)\rho_1(dy) = \frac{1}{h^k}\int_{\Omega_z} K(\frac{\cdot}{h}, \frac{y}{h})\rho_1(dy)$$

defines the convolution with RI kernel K. *Also,* $\frac{y}{h} = (\frac{y_1}{h}, ..., \frac{y_k}{h})'$, *for* $h > 0$. *The terms (i) and (iii) both* $\to 0$ *as* $h \to 0$ *(Giné and Nickl (2021), Proposition 4.3.31). On the other hand,*

$$||K_h(\widetilde{(E_{\#}\mu)}_n) - K_h(\tilde{\rho}_n)||_1 \leq \int \Big\{\frac{1}{h^k}\int \Big|K(\frac{z}{h}, \frac{y}{h}) - K(\frac{z}{h}, \frac{y'}{h})\Big|dz\Big\}d\Pi(y, y') \tag{2.6}$$

$$= \int \Big\{\frac{\int |K(z', \frac{y}{h}) - K(z', \frac{y'}{h})|dz'}{||y - y'||}\Big\}||y - y'||d\Pi(y, y')$$

$$\lesssim \frac{1}{h}\int ||y - y'||d\Pi(y, y'), \tag{2.7}$$

where (2.6) is due to Jensen's inequality and Π *denotes a coupling between* $\widetilde{(E_{\#}\mu)}_n$ *and* $\tilde{\rho}_n$. *The invariance of* K *implies the inequality (2.7), which holds for all such measure couples. Hence, given* $c_z \equiv L_2$, *the quantity (ii)* $\lesssim \frac{1}{h} d_{\mathcal{L}^1_{c_z}}(\widetilde{(E_{\#}\mu)}_n, \tilde{\rho}_n)$. *As such, to obtain an optimal encoder* E^*_n— *achieving latent consistency*— *it is sufficient to compute* $\Delta'_{E^*_n,n} = \inf_E d_{\mathcal{L}^1_{c_z}}\big(\widetilde{(E_{\#}\mu)}_n, \tilde{\rho}_n\big)$ *instead. This is highly maneuverable computationally due to the sheer attention the problem has received in recent years. One can achieve a complexity of* $\tilde{O}\big(\frac{n^{\frac{9}{4}}}{t} \wedge \frac{n^2}{t^2}\big)$ *(Dvurechensky et al., 2018), even beyond what Sinkhorn's algorithm offers. This gives us an estimate of the cost associated with finding a latent-consistent encoder from a density estimation perspective.*

Inequality (2.4) provides a clear pathway to a non-asymptotic upper bound to the realized latent loss in a WAE-GAN setup. Given that Assumption 2.1 and 2.2 hold, we begin with $\tilde{\mu}_n$, an RIK density estimate (strongly invariant) of μ based on bandwidth $h \equiv h_n \to 0$ as $n \to \infty$, such that $\frac{nh_n^d}{|\log h_n^d|} \to \infty$. Corollary 2.2 implies the existence of a positive constant E_1'' such that

$$d_{\text{TV}}\left(g_{\#}\tilde{\mu}_n, \widehat{(g_{\#}\mu)}_n\right) \leq t + \mathcal{O}(h^{m_x} \vee \sqrt{v_z}n^{-\frac{1}{2}}) \tag{2.8}$$

holds with probability $\geq 1 - E_1'' \exp\{-E_3(1 \wedge h^d)nt^2\}$, whenever $t \geq \sqrt{\frac{|\log h_n|}{nh_n^d}}$ (Giné and Guillou, 2002) and $g \in \mathscr{F}_U(\mathcal{X}, \mathcal{Z})$. This eventually determines the rate associated with the probabilistic statement

$$\text{JS}(g_{\#}\tilde{\mu}_n, \rho) - f(\pi)\Delta_{U,n} = o_{\mathbb{P}}(1),$$

obtained as a consequence of (2.4) and assuming VC-dim$[\mathcal{Y}(\mathcal{P}(\mathcal{Z}))] = v_z < \infty$. Here, $\Delta_{U,n} = \inf_{g \in \mathscr{F}_U(\mathcal{X},\mathcal{Z})} d_{\text{TV}}(\widehat{(g_{\#}\mu)}_n, \widehat{\rho}_n)$. If one employs instead a NN-based encoder $\phi \in \Phi(W, L)^k_d$ according to our previous prescriptions (e.g., Example 1 or 2), an additional estimation error is duly incurred. This, along with the realized $\Delta_{\Phi,n}$ contributes to the extent of tolerable

latent loss.

Remark 2.8. *There are some interesting implications in the WAE-GAN regime, if along with Assumption 2.2, there exists $m_z \in \mathbb{R}_{>0}$, such that $p_\rho \in \mathcal{C}^{m_z}_{R'}(\Omega_z)$, $R' > 0$. By definition*

$$JS(E_{\#}\tilde{\mu}_n, \rho) \geq \frac{1}{2}\left[\pi d^2_{TV}\left(E_{\#}\tilde{\mu}_n, \mathcal{M}_{E_{\#}\tilde{\mu}_n,\rho}(\pi)\right) + (1-\pi) d^2_{TV}\left(\rho, \mathcal{M}_{E_{\#}\tilde{\mu}_n,\rho}(\pi)\right)\right] \tag{2.9}$$

$$= \frac{1}{2}\pi(1-\pi) d^2_{TV}\left(E_{\#}\tilde{\mu}_n, \rho\right) \tag{2.10}$$

$$\geq \frac{1}{8}\pi(1-\pi)\|E_{\#}\tilde{\mu}_n - \rho\|^2,$$

where we define the mixture as $\mathcal{M}_{E_{\#}\tilde{\mu}_n,\rho}(\pi) = \pi E_{\#}\tilde{\mu}_n + (1-\pi)\rho$, $\pi \in [0,1]$. The step (2.9) is due to Pinsker's inequality. We reach (2.10) using

$$\begin{aligned} d_{TV}\left(E_{\#}\tilde{\mu}_n, \mathcal{M}_{E_{\#}\tilde{\mu}_n,\rho}(\pi)\right) &= \sup_{\omega \in \Sigma_{\mathcal{Z}}} |E_{\#}\tilde{\mu}_n(\omega) - \pi E_{\#}\tilde{\mu}_n(\omega) - (1-\pi)\rho(\omega)| \\ &= (1-\pi) \sup_{\omega \in \Sigma_{\mathcal{Z}}} |E_{\#}\tilde{\mu}_n(\omega) - \rho(\omega)| = (1-\pi) d_{TV}\left(E_{\#}\tilde{\mu}_n, \rho\right). \end{aligned}$$

Typically, the value of π is taken to be $1/2$. Now,

$$\|E_{\#}\tilde{\mu}_n - \rho\|^2 \geq \inf_{\tilde{\rho}_n} \|\tilde{\rho}_n - \rho\|^2,$$

where the infimum is taken over the class of RIK density estimates based on n i.i.d. samples from p_ρ. Such estimators, under the L_2 loss tend to have the optimal convergence rate, i.e. $\inf_{\tilde{\rho}_n} \mathbb{E}\|\tilde{\rho}_n - \rho\|^2 \gtrsim n^{-\frac{2m_z}{2m_z+k}}$ (van der Vaart (2000), Theorem 24.4). As such, this gives us a sharp lower bound for latent performance.

Before showing latent consistency, we need to acknowledge another aspect of ρ. Since it embodies the hidden representation in input images, it must also remain invariant to certain deformations. For example, latent codes corresponding to an image of a lesion and its rotated counterpart ($\mathrm{SO}(k)$) should ideally appear equiprobably. Generative models achieve this by ensuring group symmetry in the target space (Birrell et al., 2022). Now, given a group Θ (an ordered pair of a nonempty set and a binary operation, satisfying the group axioms), a *group action* φ on $\mathcal{Z}$ is an automorphism defined as $\varphi_\theta = \varphi(\theta, \cdot) : \mathcal{Z} \to \mathcal{Z}$, $\forall \theta \in \Theta$, also satisfying $\varphi_{\theta_1} \circ \varphi_{\theta_2} = \varphi_{\theta_1 \cdot \theta_2}$, $\forall \theta_1, \theta_2 \in \Theta$. The following definition completes the characterization of ρ.

Definition 2.11 (Invariant Distributions)**.** *Given a group Θ, the class of Θ-invariant probability distributions on $\mathcal{Z}$ is defined as*

$$\mathcal{P}_\Theta(\mathcal{Z}) = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) : \mathbb{P} = (\varphi_\theta)_{\#}\mathbb{P}, \forall \theta \in \Theta\}.$$

Throughout the chapter, we only consider finite groups, i.e., $|\Theta| < \infty$. This makes the representation of the underlying space under group actions much easier. To that end, let us

introduce the *fundamental domain* $\mathcal{Z}_0 \subset \mathcal{Z}$, which is defined such that the subsets $\theta\mathcal{Z}_0$, $\theta \in \Theta$ form a locally finite cover of $\mathcal{Z}$ without sharing common interior points. This translates to saying that there exists a unique $z_0 \in \mathcal{Z}_0$ corresponding to each $z \in \mathcal{Z}$ such that $z = \theta z_0$ (Chen et al., 2023c). In order to adapt to the estimation of Θ-invariant latent distributions, we make additional assumptions for the underlying kernels in MMDs.

Assumption 2.3 (Group Invariant Kernels)**.** *The kernel* $\kappa(\cdot,\cdot)$ *satisfies* $\forall\theta(\neq id) \in \Theta$

(i) Given $z, z^{'} \in \mathcal{Z}$, $\kappa(\theta z, \theta z^{'}) = \kappa(z, z^{'})$, *and*

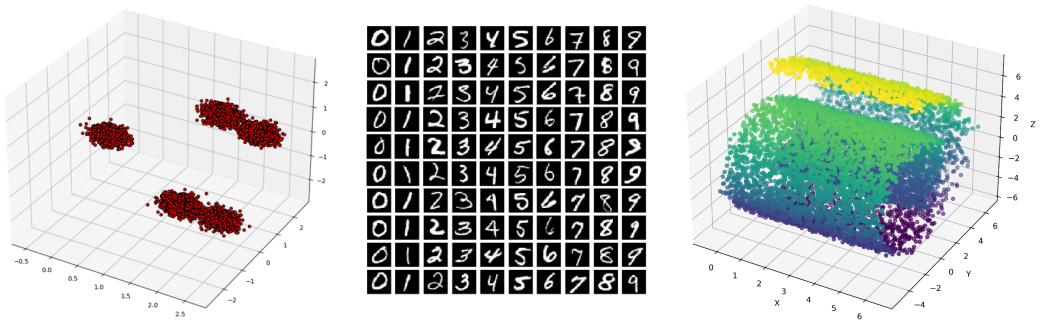
(ii) There exists $0 < \varsigma_{\kappa,\Theta} < 1$ *such that* $\kappa(\theta z, z) \leq \varsigma_{\kappa,\Theta} C_\kappa$, *where* $z \in \mathcal{Z}_0$.

Theorem 2.4 (Latent Space Consistency in WAE-MMD under Invariance)**.** *Let,* $\rho \in \mathcal{P}_\Theta(\Omega_z)$ *such that* $|\Theta| < \infty$. *Also, let* $\kappa(\cdot,\cdot)$ *be strongly invariant satisfying Assumption 2.3 such that* $\sup_{z\in\Omega_z} \kappa(z,z) \leq C_\kappa$, *for* $C_\kappa > 0$. *Then, there exists a probabilistic encoder* $\phi \in \Phi(W, L)_d^k$, *based on ReLU activations with* $W = \mathcal{O}(d\lfloor N_1^{\frac{1}{d}} \rfloor \vee N_1 + 1)$ *and* $L = \mathcal{O}(N_2)$ *such that given* $\delta > 0$, *we have with probability* $1 - \delta$

$$d_{\mathcal{H}_\kappa}\left(\phi_{\#}\hat{\mu}_n, \rho\right) - \sqrt{C_\kappa}\sup_{\rho^{\otimes n}} \Delta_{\Phi,n} \leq \sqrt{c_{g,n}}\left(\frac{\max\{B_x^2, 4C_\kappa[1 + \varsigma_{\kappa,\Theta}\left(|\Theta| - 1\right)]\}}{2n} \ln\left(\frac{2}{\delta}\right)\right)^{\frac{1}{4}}$$
$$+ \mathcal{O}(\sqrt{c_{g,n}}(d^2 n)^{-\frac{1}{2d}}) + \mathcal{O}(\sqrt{dD_n} N_1^{-\frac{2}{d}} N_2^{-\frac{2}{d}}) + \sqrt{\frac{2C_\kappa}{n}}\left[1 + \sqrt{\frac{1 + \varsigma_{\kappa,\Theta}\left(|\Theta| - 1\right)}{|\Theta|}}\right],$$

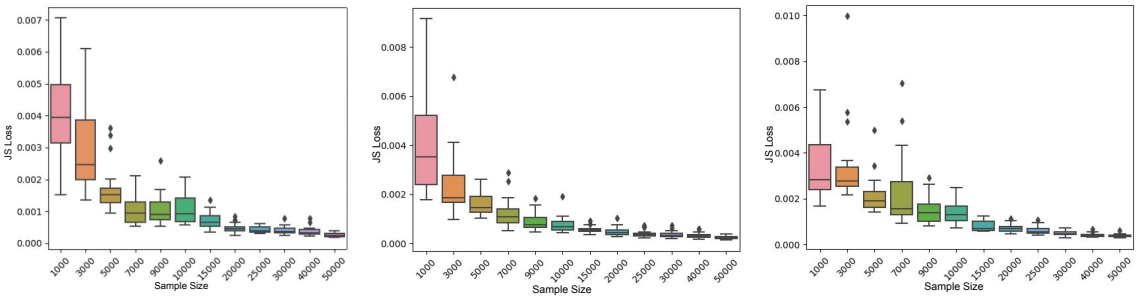
where both D_n *and* $c_{g,n}$ *are sequences based on* n *that* $\searrow 0$ *almost surely as* $n \to \infty$.

The result says that the realized latent loss in a WAE-MMD converges (in $\mathbb{P}$) to a small data-dependent error margin, which indicates the extent to which the constraint in (2.2) is satisfied.

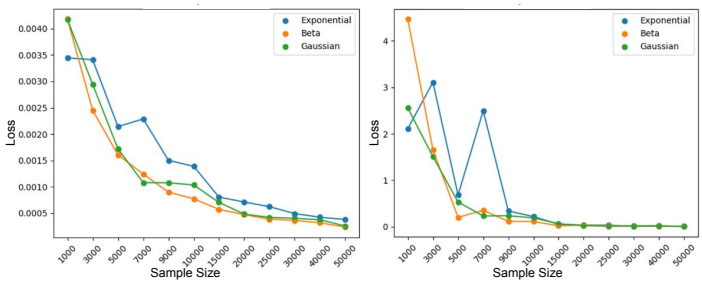


(a) (b) (c)

Figure 2.1: The (a) Five-Gaussian, (b) MNIST, and (c) Swiss roll data sets.

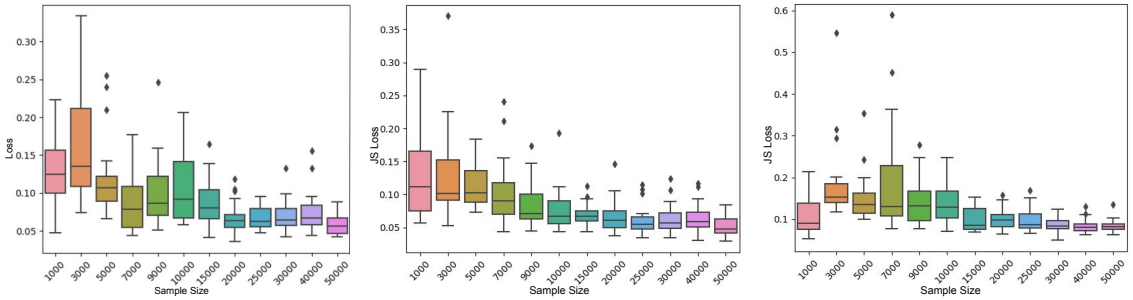


(a) Gaussian (b) Beta copula (c) Exponential copula



(d) Mean and Variance of latent loss

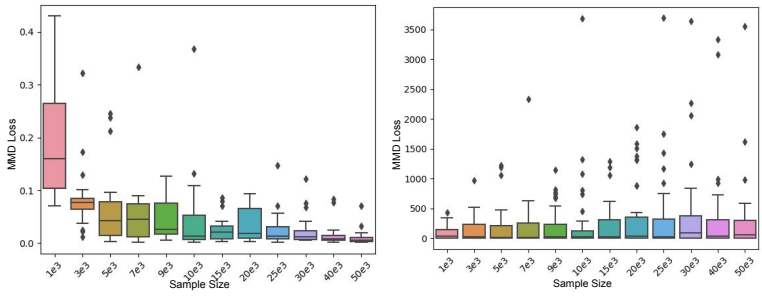
Figure 2.2: Latent loss corresponding to Five-Gaussian data under Jensen-Shannon divergence using ReLU encoders. The Lagrangian weight assigned to the latent space, as given in (2.1), remains $\lambda = 0.2$. We consider both Gaussian and Exponential marginal densities as standard. The parameters for Beta marginals are taken as $(0.5, 0.8)$.



(a) Gaussian (b) Beta (c) Exponential

Figure 2.3: Propagation of sample corrected $(\times n)$ latent JS loss for Five-Gaussian data under ReLU encoders.

Remark 2.9 (Mitigating Curse of Dimensionality)**.** *The term contributing to a slower convergence rate (second on the RHS) due to its dependence on d is rooted in the estimation error under the Wasserstein metric (see proof). While we do not allow the input dimension d to grow as a function of n, it being inherently large degrades the sharpness of the non-asymptotic bound. In search of a remedy, we recall the solution showed in Chakrabarty and Das (2021), namely the 1-upper Wasserstein dimension (d_1^*). It is typically smaller com-*



(a) (b)

Figure 2.4: Propagation of (a) sample MMD losses and (b) sample corrected ($\times n^{\frac{1}{2}}$) MMD losses corresponding to Five-Gaussian, trained with the Lagrangian parameter $\lambda = 0.2$ using ReLU encoders.

pared to the Minkowski–Bouligand dimension. However, in case μ is essentially supported on a 'latent' regular space, e.g. a compact d'-dimensional differentiable manifold ($d' < d$), we have $d_1^* = d'$ (Weed and Bach, 2019). This suits our discussion since such a phenomenon regularly occurs in image datasets (Pope et al., 2021). The definition of d_1^* goes as follows

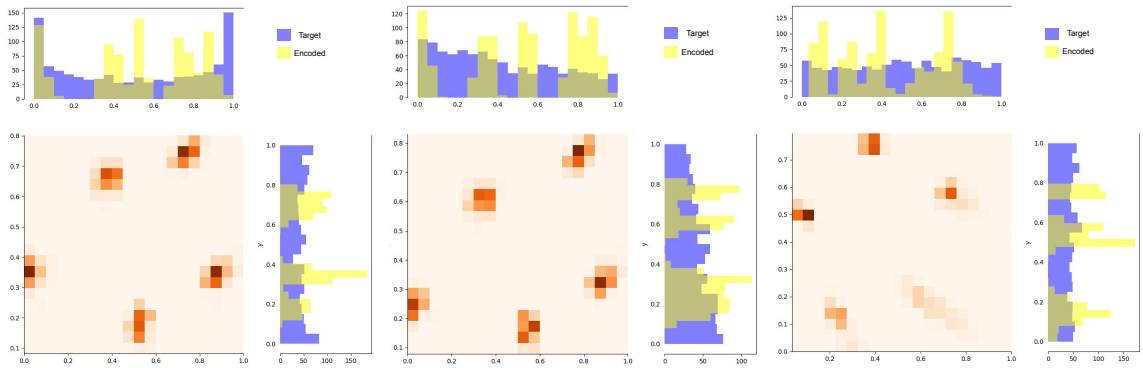
$$d_1^*(\mu) = \inf\left\{s \in (2,\infty) : \limsup_{\varepsilon\to 0} \frac{\mathcal{N}_\varepsilon(\mu, \varepsilon^{\frac{s}{s-2}})}{-\log(\varepsilon)} \leq s\right\},$$

where $\mathcal{N}_\varepsilon$ denotes the covering number. Now, if $s > d_1^*(\mu)$, the second term on the right hand of the inequality in Theorem 2.4 can be replaced by $\mathcal{O}(\sqrt{c_{g,n}} n^{-\frac{1}{2s}})$. This approach also generalizes that of Chakraborty and Bartlett (2024). We also point out that the other term carrying d in the exponent does not contribute to asymptotic rates since the encoders constructed are usually of fixed proportions. Since the upper bound becomes $o(1)$, using the Borel-Cantelli lemma, the latent WAE-MMD error deviated from $\sqrt{C_\kappa}\sup_{\rho^{\otimes n}} \Delta_{\Phi,n}$ vanishes almost surely.

2.4.3 Simulations

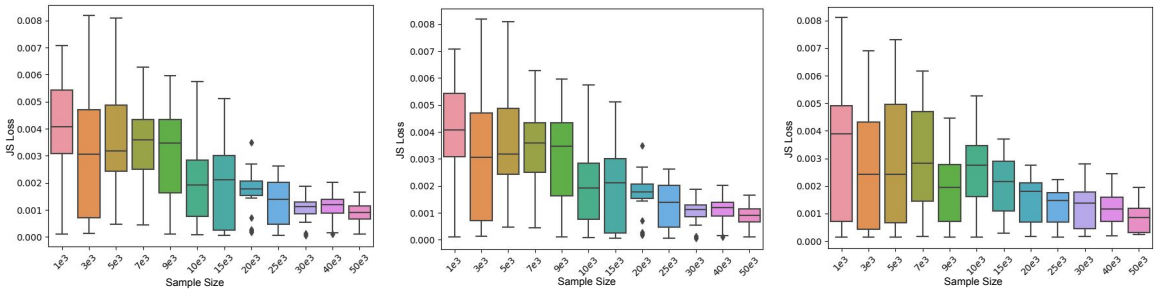
To validate our findings empirically, we carry out experiments on both real and synthetic data [Fig. 2.1]. The existing data set we work on is MNIST (LeCun et al.), consisting of 70,000 2D images of hand-written digits. The 'Five-Gaussian' data set is a collection of 50,000 observations drawn independently at random out of five trivariate Gaussians with unit dispersion and mean at five vertices of a unit cube. To corroborate the dependence of the convergence rate solely on the latent dimension, we consider the 'Swiss roll' data. It consists of 50,000 observations lying on a curved 2D plain. We run both WAE-GAN and WAE-MMD to reconstruct observations from the data sets. All the experiments are carried out on an RTX 3090 GPU.

Five-Gaussian. The encoders we use in the case of Five-Gaussian data map the points to a 2D latent space. The first kind we deploy is a 4-deep, ReLU-activated network. However, the last layer uses an additional rescaling to span the target support and mitigate zero inflation. To suit our theoretical specifications, we experiment with diverse latent distributions, namely, bivariate standard Gaussian and the classes of bivariate distributions having Beta and Exponential marginals, respectively. We call them Beta and Exponential *copulas*. This way, we encompass unbounded supports, multimodality, and skewed densities. Firstly, we train the model based on the entire sample ($n = 50,000$) to obtain the nearest estimate of the population loss. Our goal is then to observe the propagation of the losses as we gradually increase the sample size $n = (1000, 3000, 5000, \cdots, 50000)$, drawn uniformly at random. To account for the variation due to sampling, we carry out 20 runs corresponding to each n. The following discussion interprets our findings.



(a) (b) (c)

Figure 2.5: Bin estimates of ReLU-encoded (yellow) vs latent distribution (blue) in case of the Five-Gaussian data. Under the JS loss, we observe (a) Beta $(0.5, 0.8)$ and (b) standard Exponential copula, (c) shows standard Gaussian under MMD loss. (Effective range of values scaled to aid visualization)



(a) Gaussian (b) Beta (c) Exponential

Figure 2.6: Latent JS loss under Groupsort encoders of grouping size 2 for Five-Gaussian data.

The illustrations show that for all three latent laws, the sample losses approach their population counterpart with diminishing variance [Fig. 2.2]. In search of the sharpest asymptotic rate associated— even beyond the theoretically achievable $\mathcal{O}(n^{-\frac{1}{2}})$— we observe the movement of the loss multiplied by n. In our experiments [Fig. 2.3], such a sequence also tends to converge to a small constant, namely, the population error margin. This is an empirical guarantee to the impression Asatryan et al. (2023) [Remark 4.6] had ($\hat{d}_{\text{JS}} \sim n^{-1}$) in a parametric GAN (Goodfellow et al., 2014) generation. However, using a GroupSort activated encoder (grouping size 2) one may observe the approximate rate of $\mathcal{O}(n^{-\frac{1}{2}})$ in diminishing MMD losses [Fig. 2.22 (b), (c)]. Note that the regularizing parameter λ is chosen based on a trade-off between quality reconstruction and latent performance.

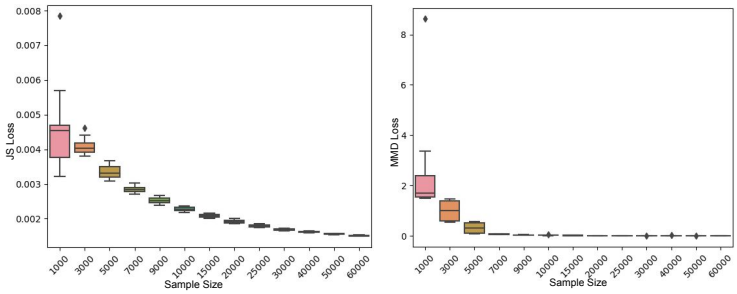
We follow the same experimental protocol for WAE-MMD [Fig. 2.4], taking the latent distribution as bivariate standard Gaussian. Here as well, a similar trait is noticed. The observed MMD losses gradually decrease to their population counterpart with diminishing variability. The rate of convergence however, becomes comparable to $\mathcal{O}(n^{-\frac{1}{2}})$, attesting the theoretical result. While the latent loss— asymptotically at the least— moves close to nullity, the bin estimates corresponding to the target and the encoded law must differ. This discrepancy, as we have already discussed, is rooted in the information preserved.

We study the concentration of encoded bins in contrast with the latent ones over regular intervals of 200 training epochs. It is fascinating to observe the rearrangement of density as the losses slowly diminish. We present a detailed commentary on the same in the Appendix [Fig. 2.16, 2.17]. Here instead, we show the encoded estimates after the completion of 2000 epochs [Fig. 2.5]. The information retention can be readily identified from the high-density areas representing the distinct clusters. The quantile-quantile (QQ) plots [Fig. 2.20a] between the empirical encoded distribution and the targets also tells the same story. To check the extent of approximation, we also perform multivariate goodness-of-fit tests (see Appendix 2.8).

For checking the efficacy of GroupSort activations in encoding, we repeat the experiment in a WAE-GAN setup [Fig. 2.6]. The regularization remains at $\lambda = 0.2$, and the grouping size is taken as 2 (OPLU). Quite similar to previous observations, the losses tend to decrease at a familiar rate.

MNIST. The individual images in MNIST are of size (28×28). In vectorized form, the input observations are reduced to $d = 512$. Here also, we deploy a 4-deep ReLU encoder with layers of width 512-256-128-64=k. During decoding, the output tensor is reshaped to have a size $(\text{batch size}, 1, 28, 28)$, enabling us to calculate the reconstruction loss. Keeping in mind the high dimensionality of the latent space, we only consider a multivariate standard Gaussian target. The findings from training runs on both WAE-GAN and WAE-MMD with regularization $\lambda = 0.2$ are obtained as in Fig. 2.7.

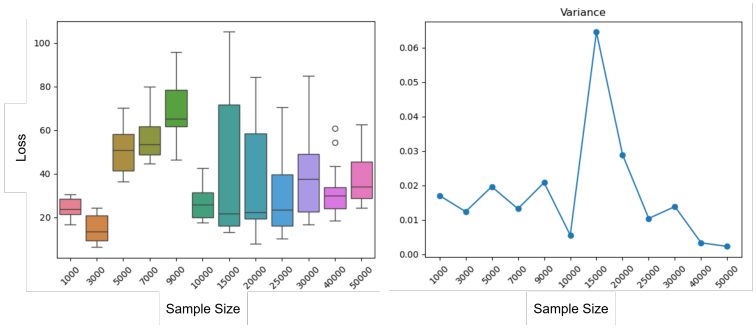
Swiss roll. Following Theorem 2.4, we construct the encoder with intermediate widths



(a) (b)

Figure 2.7: Latent (a) JS and (b) MMD loss for MNIST data set with Gaussian targets. Both losses tend to converge to the population benchmark at a sharp rate.

given as 3-6-12-32-64-32-12-2=k. The activations remain ReLU throughout, with rescaling in the encoded output. We keep the target latent distribution to be Gaussian and the regularization $\lambda = 0.2$. In this setup also, the sample latent losses converge to the population near estimate with diminishing variance [Fig. 2.8]. It is fascinating to observe that the convergence rate adapts to the intrinsic dimension ($n^{-\frac{1}{2}}$) rather than $d = 3$ [Fig. 2.8a]. We also get a clearer look at IP and its gradual realization over epochs [Fig. 2.9]. The encoded observations simultaneously retain the intrinsic pattern of the data and meet the marginal constraints [Fig. 2.18]. Here, one may observe a more prominent mode-covering effect in the marginals.



(a) (b)

Figure 2.8: Propagation of (a) sample corrected ($\times n^{\frac{1}{2}}$) MMD losses and (b) corresponding variances for Swiss roll data.

2.5 Reconstruction Consistency

With a clearer understanding of WAEs' performance towards meeting the constraint it was formulated under, we move on to its main objective. If we position ourselves along the flow of information— first through the encoder and now at the footsteps of the decoder— we have

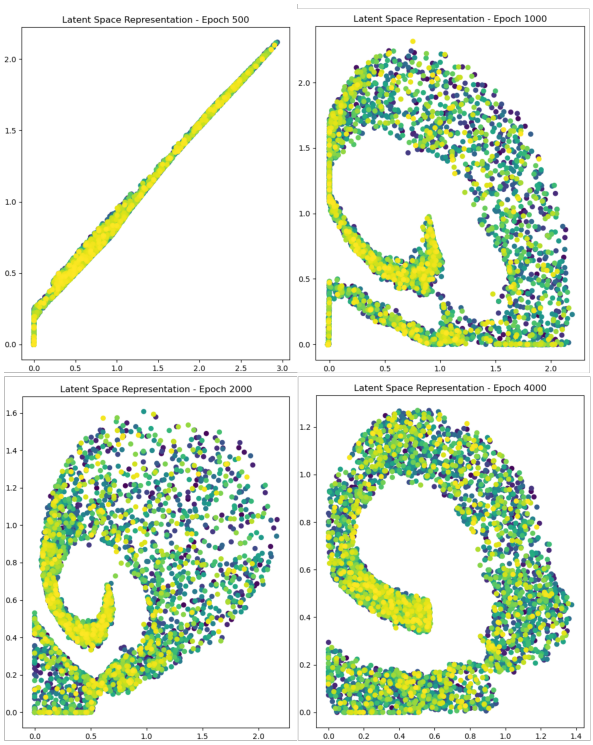


Figure 2.9: Information preservation in encoded observations over epochs $(500, 1000, 2000, 4000)$ (left to right from top) in WAE-MMD for Swiss roll data.

ourselves a density estimation task in higher dimensions. This is typical of inverse models, and the underlying goal is to utilize the encoded information at one's disposal to reach as close as to $\hat{\mu}_n$. It comes as a consequence of the error $W_{c_x}^1(\hat{\mu}_n, (D \circ E_n^*(t))_{\#}\hat{\mu}_n)$ being minimized, given the optimal encoder $E_n^*(t)$ incurring latent loss $\leq t$.

In spirit, the role of the decoder is somewhat similar to a generative map. The aspects in which they differ from those in a GAN architecture are mainly twofold. Firstly, there is no dynamic critic in the form of a discriminator to guide its learning. The role is taken up by $\mathcal{L}_{c_x}^1$ only. However, on the upside, while the latent distribution in GANs is non-informative, WAEs have latent laws with input information 'preserved'. During decoding, this very information needs to be utilized with the utmost efficacy. Since the resemblance between spaces of different dimensions in a sample problem lies in the local geometry based on pairwise distances between samples, quasi-isometries or bi-Lipschitz maps can be identified as ideal decoders (Chakrabarty and Das, 2021). However, $\mathbb{R}^k$ is not quasi-isometric to $\mathbb{R}^d$, $d > k$ in general. There may only exist bi-Lipschitz maps from $\Omega_z \to \mathbb{R}^l$, $l \leq d$, which can thus be used to form outer extensions mapping $\mathbb{R}^k \to \mathbb{R}^d$ (Mahabadi et al., 2018). Such extensions typically preserve distortions up to constant factors and, as a result, do not depreciate the asymptotic behavior of estimates post-translation. Another technique to ensure bi-Lipschitzness is to search for *regular* maps $f : \mathbb{R}^k \to \mathbb{R}^d$, defined as a Lipschitz map $f : (\mathcal{Z}, c_z) \to (\mathcal{X}, c_x)$

if there exists a constant $C > 0$ such that given any ball B in $\mathcal{Z}$, $f^{-1}(B)$ can be covered by at most C balls, each of radius $C \cdot \mathrm{rad}(B)$ (David and Semmes, 2000). These, in turn, make the restriction of f to $\overline{\Omega}_z$ (closure) to be BL. While bi-Lipschitz transforms acting as decoders automatically enforce non-degeneracy in reconstructed signals, it is sufficient to have Lipschitz decoders to obtain an upper bound on the associated error. Most theoretical studies on GANs tend to impose this restriction on generators. The existence of such a benchmark Lipschitz transform, however, is readily guaranteed if encoders are considered according to Lemma 2.3. This also supports the practical convention of building the decoder as the exact inverse of $E^*(t)$.

We, on the contrary, prescribe constructing a decoder map that utilizes the information encapsulated in the latent law wholly and efficiently. To demonstrate the notion, let us fragment the reconstruction loss, given an encoder E, as follows

$$W^1_{c_x}(\mu, (D \circ E)_{\#}\hat{\mu}_n) \leq W^1_{c_x}((D \circ E)_{\#}\hat{\mu}_n, D_{\#}\rho) + \underbrace{W^1_{c_x}(D_{\#}\rho, \hat{\mu}_n)}_{\text{Decoder Translation Error}} + W^1_{c_x}(\hat{\mu}_n, \mu). \quad (2.11)$$

Observe that the first term on the right is essentially the propagated disagreement in the latent space. If the metric measuring the discrepancy follows a data-processing inequality, we can expect a non-asymptotic upper bound of the same order as that obtained on the latent error. As such, D must be constructed keeping in mind the sole aim of minimizing the semi-discrete translation error. The following lemma provides the backbone of the construction.

Lemma 2.5 (Yang et al. (2022)). *Let ν be an univariate absolutely continuous distribution and $\hat{\mu}_n \in \mu^{\otimes n}$. Given $W \geq 7d + 1$ and $L \geq 2$, there exists a NN transform based on ReLU activation $\phi^{'} \in \Phi(W, L)^d_1$ such that $\forall \varepsilon > 0$*

$$W^1_{c_x \equiv L_1}(\hat{\mu}_n, \phi^{'}_{\#}\nu) \leq \varepsilon,$$

whenever $n \leq \frac{W-d-1}{2}\lfloor \frac{W-d-1}{6d} \rfloor \lfloor \frac{L}{2} \rfloor + 2$.

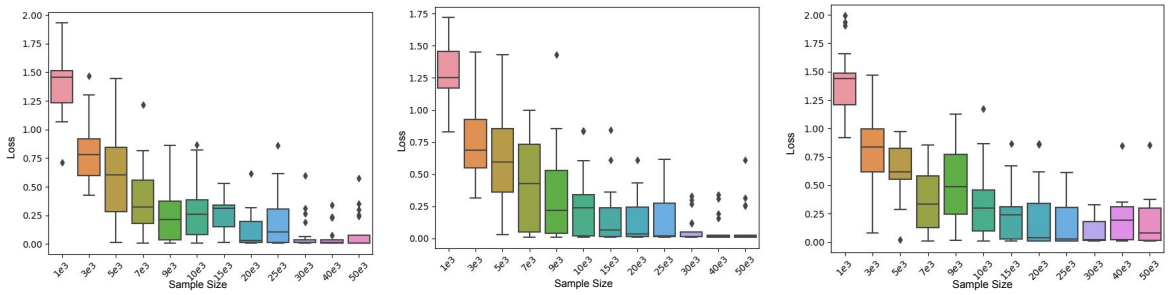
The transport hinted in the lemma is essentially a piecewise linear map, Lipschitz continuous on bounded balls. The restriction on n indicates the number of breakpoints. Since the result holds for all probability measures ν having densities, the only modification required in our case is projecting ρ onto $\mathbb{R}$, using linear maps beforehand. Given such a linear map $D_0 : \Omega_z \to \mathbb{R}$, and $\phi^{'}$ according to lemma 2.5, the desired decoder is given by $D = \phi^{'} \circ D_0$. This operation also preserves the Lipschitz continuity in the resultant decoder. Since there is no unique way of selecting the linear transform, one may use instead a pooled distribution. As such, $D = \phi^{'} \circ \sum_{i=1}^{N_3} D_i \equiv \sum_{i=1}^{N_3} \phi^{'} \circ D_i$, where $D_1, \cdots, D_{N_3}$ are individual linear maps aimed at preserving different aspects of ρ. This is especially useful when $|\Theta| > 1$. In this case also, D turns out to be Lipschitz (since the property itself stems from individual components

and summation— that too without scaling— only changes the associated constant). The following result formalizes our discussion.

Theorem 2.5 (Reconstruction Consistency in a Latent-Consistent WAE)**.** *Given a margin of latent error $t > 0$, let $E_n^*(t)$ be an optimal encoder satisfying latent consistency under the metric d_{TV}. Then, there exists a decoder $D \in \Phi(W, L)_k^d$, $d \geq 3$ based on ReLU activations, with $W \geq 7d + 1$ and $L \geq 3$ such that*

$$\mathbb{E}\left[W_{c_x}^1(\mu, (D \circ E_n^*(t))_{\#}\hat{\mu}_n)\right] - \mathcal{O}(t) \lesssim n^{-\frac{1}{d}},$$

where $n = \mathcal{O}(\frac{W^2 L}{d})$.



(a) Gaussian (b) Beta copula (c) Exponential copula

Figure 2.10: Wasserstein reconstruction loss for Five-Gaussian data corresponding to three latent distributions, under MMD using ReLU encoders. The penalization on the latent loss is kept at $\lambda = 0.2$.

The theorem reveals the extent to which the realized latent loss can potentially amplify during reconstruction. The corresponding excess error always stays $\mathcal{O}(n^{-\frac{1}{d\vee 2}})$, with high probability (using McDiarmid's inequality). A similar result for general f-WAE reconstructions can be shown assuming invertibility of D (Chakrabarty and Das, 2021). The result also allows for the formulation of WAEs to be made general by replacing $W_{c_x}^1$ with $W_{c_x}^p$, $p \geq 1$. In such a case, the observation: $W^p(\mu_1, \mu_2) \leq B_x^{\frac{p-1}{p}} W^1(\mu_1, \mu_2)^{\frac{1}{p}}$, $\mu_1, \mu_2 \in \mathcal{P}(\mathcal{X})$ coupled with Theorem 2.5 provides the regeneration guarantee.

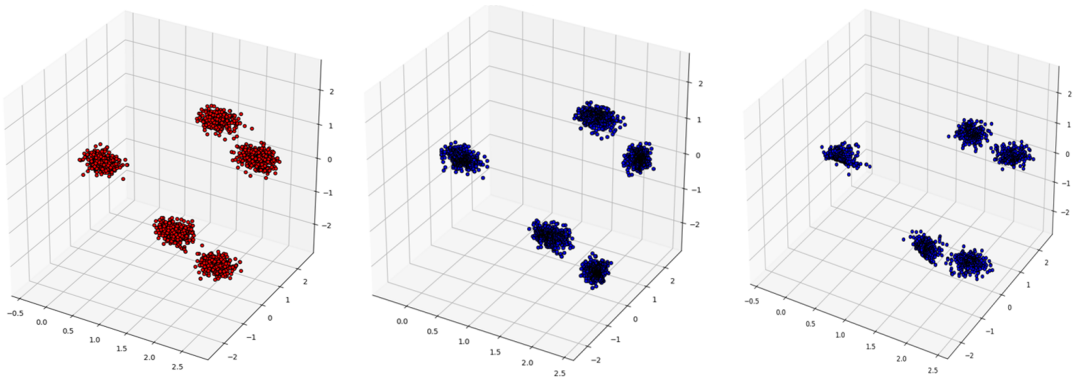
Remark 2.10 (Reconstruction as a Consequence of Latent Consistency in WAE-MMD)**.** *Considering a latent error $d_{TV}(E_n^*(t)_{\#}\hat{\mu}_n, \rho) \leq t$, one may show reconstruction consistency for both WAE models (WAE-GAN and WAE-MMD) based on the fact that TV is a natural upper bound to both JS and MMD. However, if the bounded kernel κ applied in a latent space (in WAE-MMDs) is strictly positive definite and follows strong invariance, we have the partial inequality— given Assumption 2.1 and 2.2— as follows*

$$W_{c_x}^1((D \circ E)_{\#}\hat{\mu}_n, D_{\#}\rho) \lesssim W_{c_z}^1(E_{\#}\hat{\mu}_n, \rho) \leq C_\varepsilon \, d_{\mathcal{H}_\kappa}(E_{\#}\hat{\mu}_n, \rho) + \varepsilon,$$

for some $C_\varepsilon > 0$ and $\forall \varepsilon > 0$ (Modeste and Dombry (2024), Proposition 3.9). The first inequality is due to the Lipschitz continuity of D. Now, if MMD, equipped with the same κ is applied on the input space as well, the same D constructed so far satisfies $d_{\mathcal{H}_\kappa}(D_{\#}\rho, \hat{\mu}_n) < \varepsilon$, $\forall \varepsilon > 0$ (Yang et al. (2022), Lemma 3.3). As a result, the right-hand side of inequality (2.11) can be written entirely in terms of MMD. Hence, Theorem 2.4 can be readily plugged in to obtain a deterministic upper bound to the realized reconstruction error.

2.5.1 Simulations

We continue with the earlier experimental setup to provide empirical validation. In fact, reconstruction outputs are obtained simultaneously with latent results. Conforming to our previous prescription, we employ 4-deep decoders for the Five-Gaussian data set (decoder architecture: k=2-22-16-16-3). On the other hand, to reconstruct observations corresponding to MNIST, we take a pragmatic approach while choosing network widths (k=64-128-256-512). The final output tensor is suitably reshaped to have the size of an image (batch size, 1, 28, 28). For Swiss roll, following Theorem 2.5, we deploy a ReLU-activated decoder with intermediate widths k=2-32-64-32-3.

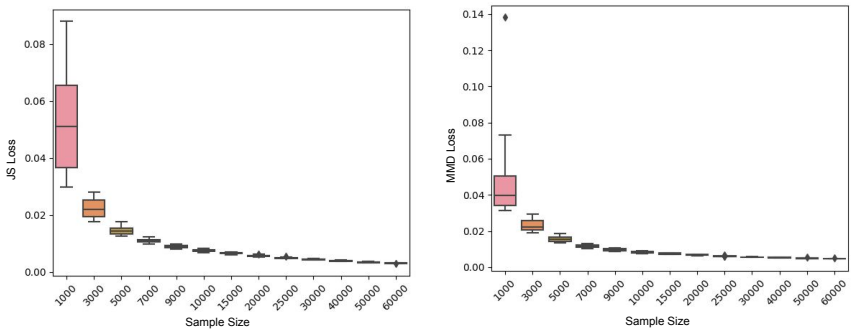


(a) Input data (b) Gaussian (c) Beta copula

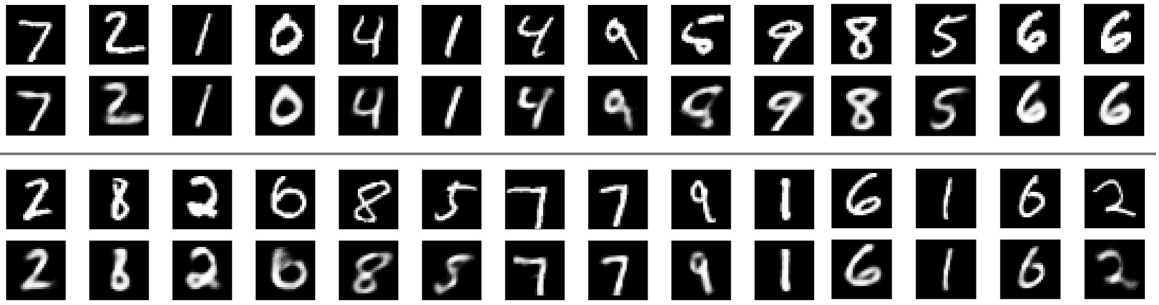
Figure 2.11: Reconstructed samples (n = 10000) from the Five-Gaussian dataset under JS latent loss for given latent distributions, using ReLU encoders after 1500 epochs.

The diminishing trait of error values with shrinking variance is evident in Fig. 2.10. The limiting margin of error, where the sequence converges (for all latent distributions under consideration), remains well below the tolerable latent loss. This further attests to our theoretical bound. The optimizations not only make the error eventually vanish but also produce perceptually alike samples [Fig. 2.11]. Reconstruction errors corresponding to Five-Gaussian in a WAE-MMD setup also tend to follow a convergence rate $\mathcal{O}(n^{-\frac{1}{2}})$ [Fig. 2.22]. This corroborates Remark 2.10, even under the deployment of GroupSort encoders. Reconstructions of MNIST also result in photo-realistic copies of the input samples [Fig. 2.12].

The corresponding errors exhibit sharply decaying behavior under both WAE architectures. Reconstruction errors for Swiss roll in a WAE-MMD setup also exhibit parametric decay towards the population benchmark [Fig. 2.13]. Corresponding reconstructions turn out to be near-perfect on convergence [Fig. 2.19].



(a) (b)



(c) Reconstruction under latent JS loss

Figure 2.12: MNIST reconstruction error given Gaussian latent laws under (a) JS and (b) MMD latent loss, using ReLU encoders. In (c), the odd rows hold the input digits, and the even ones are their reconstructed counterparts.

Even though the decoder model specification from Theorem 2.5 outlines a sufficient condition for reconstruction consistency, we examine the effect of a D violating the criteria on the loss. In contrast to the earlier architecture, if we alter the width of the decoder network, given Five-Gaussian data, as follows: k=2-18-16-8-3, i.e., $W = 18 \not\geq 7d + 1$, the reconstruction errors tend to diverge under sample correction ($\times n^{\frac{1}{3}}$, since $d = 3$) [Fig. 2.21(b)]. Notice that the modified decoder remains comparable to the former in the number of parameters. The corresponding regenerated samples also fail to recover variations in the clusters [Fig. 2.21(c)].

2.6 Robustness to Data Corruption

The quality of deep generative model outputs is often marred by contamination in the data. Images, and consequently WAEs, are very much susceptible to such adversarial corruption.

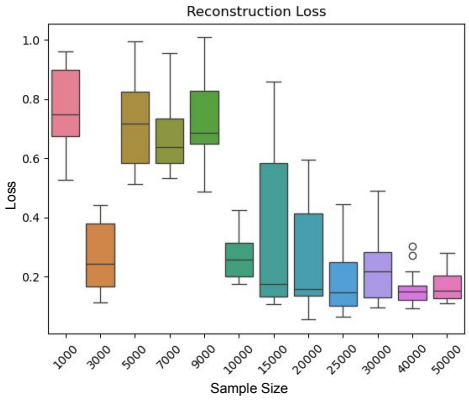


Figure 2.13: Swiss roll reconstruction error in a WAE-MMD given Gaussian latent law, using ReLU encoders.

In our model recommendation, we have prioritized the free flow of information through them. As such, a corrupted set of input samples runs the risk of spoiling all the downstream tasks. Thus, to get a comprehensive look at the machinery of WAE, one must test its innate capability to preserve regeneration quality under the influence of adversaries. Some of the well-recognized models in robust statistics include *Adaptive*, *Oblivious*, and the *Huber contamination model* (Chen et al., 2016; Zhu et al., 2022). In Huber contamination, instead of assuming independent replicates from the input density p_μ, it is assumed that there lies a probability $\epsilon > 0$ that the sample comes from a *contamination* $p_c \in \mathcal{P}(\mathcal{X})$. The density p_c is independent of the input law and remains unknown. As such, input observations

$$X_1, \cdots, X_n \sim \tilde{p} := (1-\epsilon)p_\mu + \epsilon p_c \tag{2.12}$$

are what we have at hand. Under such a setup, Liu and Gao (2019) showed that given $p_\mu \in \mathcal{C}_R^s(\Omega_x)$, kernel density estimates incur euclidean losses $\mathcal{O}(n^{-\frac{2s}{2s+1}} \vee \epsilon^{\frac{2s}{s+1}})$. The underlying kernels are considered to be square-integrable and bounded, with centered moments. This result is particularly motivating since given that $D \circ E \approx \mathrm{id}$ a.e., it gives us a bound on the regeneration error for VAEs.

The regime we consider in our following discussion is closer to oblivious contamination. We assume that the *contaminated input distribution* $\tilde{p}$ is such that $\mathbb{E}_{X\sim\tilde{p},Y\sim p_\mu} c_x(X,Y) \leq \epsilon$, a more general notion compared to Huber. Observe that such a criterion automatically implies 1-Wasserstein contamination under metric c_x (Liu and Loh, 2022). No additional assumption on the regularity of $\tilde{p}$ is assumed. The goal remains the same: reconstructing p_μ based on an input estimator. In other words, in the absence of additional regularization, we check for the extent of inherent distributional robustness WAEs possess.

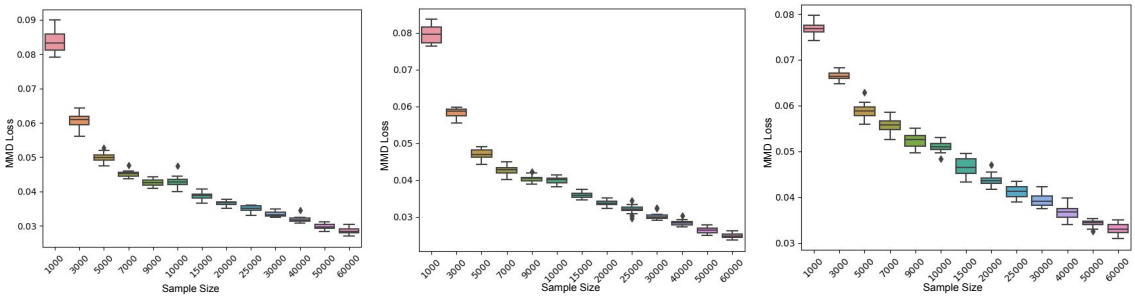
We have already seen that both the encoder and decoder under careful construction can follow Lipschitz continuity. As a result, their composition behaves similarly. To generalize

such composite maps, in this section, we consider maps $\mathcal{G} : \mathcal{X} \to \mathcal{X}$ that induce group actions. Now, if $G \in \mathcal{G}$ satisfies information preservation, it is approximately equivalent to constructing an estimator based on translated observations rather than translating a pre-constructed estimator to approach p_μ. As such, given replicates $X_1, \cdots, X_n \sim \tilde{p}$, we essentially need to look for upper bounds to the loss $W^1_{c_x}(\hat{p}_n, p_\mu) \lesssim ||\hat{p}_n - p_\mu||_1$, where $\hat{p}_n \in L_1(\mathbb{R}^d)$ is based on $\{G(X_i)\}_{i=1}^n$.

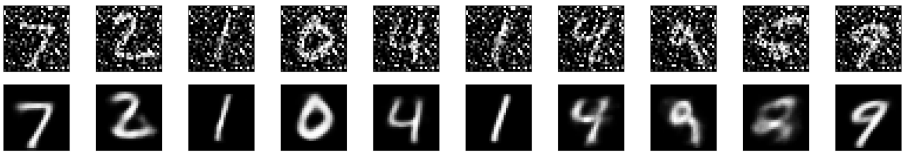
To cope with the adversary, first, we modify the properties of the regularly invariant kernels we utilized earlier. We call a kernel $\kappa(x, y) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *transformation invariant* if given action $G \in \mathcal{G}$, $\kappa(G(x), y) = \kappa(x, G^{-1}(y))$ (Liu et al., 2022). The following theorem provides the extent of WAEs' resilience based on such kernel estimates.

Theorem 2.6 (Reconstruction Consistency under Contamination)**.** *Let the contaminated distribution $\tilde{p}$ be such that $\sup_{\Omega_x} \tilde{p}(x) < \infty$. Also, let the kernel κ be regular, translation invariant with respect to L_1 (2.8) and transformation invariant which satisfies $\int_{\mathcal{X}} \kappa^2(v, v - u)du < \infty$. Then, given any $G \in \mathcal{G}$, a kernel density estimate $\hat{p}_h \equiv \hat{p}_{h,n}$ based on κ satisfies*

$$\mathbb{E}|\hat{p}_h(0) - p_\mu(0)| \lesssim n^{-\frac{m_x}{d+2m_x}} \vee \epsilon^{\frac{m_x}{2d+m_x}}.$$



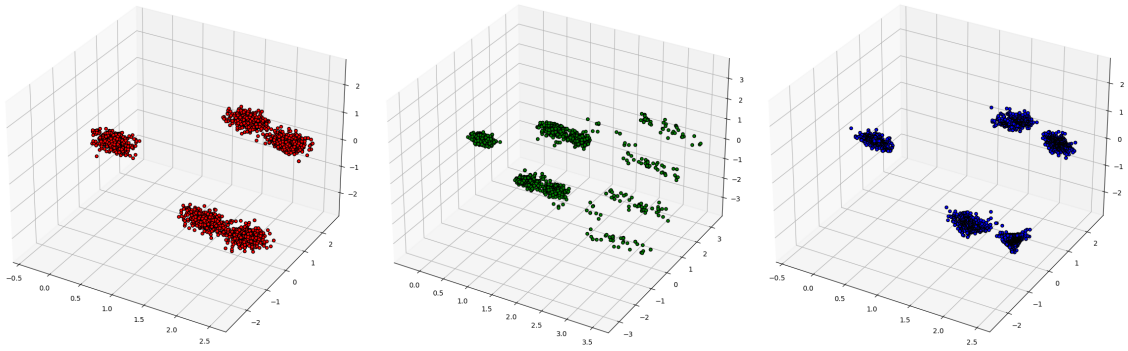
(a) Gaussian (b) Cauchy (c) Dirichlet



(d) Reconstruction under Cauchy contamination

Figure 2.14: Reconstruction errors incurred by a ReLU-induced WAE-MMD for MNIST, under different contaminating distributions at level 0.2. In all the experiments, the latent distribution is kept standard Gaussian. In (d), the first row represents contaminated samples (standard Cauchy at level 0.2), and the second row contains their reconstructed counterparts.

The result can be interpreted as the following: observations regenerated using WAEs under contamination are distributed sufficiently accurately following p_μ such that it can be recovered with high precision by only deploying a suitable kernel.



(a) Actual Data (b) Contaminated Data (c) Reconstructed Data

Figure 2.15: Reconstructed samples ($n = 10,000$) from the Five-Gaussian dataset with half the observations contaminated at level 0.2, under JS latent loss. The corrupting distribution is taken to be Dirichlet$(5, 3, 5)$.

2.6.1 Simulations

During empirical validation, based on the diverse natures of their support, the thickness of tails, and central tendencies, we select three potential contaminating distributions: Standard Gaussian, Cauchy, and Dirichlet. While the Five-Gaussian data set is corrupted with the latter two, we apply all three on MNIST. For utmost rigor in our experiments, we adopt an elaborate contaminating regime. We not only vary the proportion of observations getting corrupted but also regulate the extent of it. For example, there may be a set of input observations $\{X_i\}_{i=1}^n$, out of whom $\lfloor \frac{n}{2} \rfloor$ are replaced by $(0.8)X_i + (1 - 0.8)Y_i$, where $\{Y_i\}_{i \in \mathscr{C}}$ are replicates from a contaminating law. $\mathscr{C}$ denotes the set of indices suffering corruption, i.e., $|\mathscr{C}| = \lfloor \frac{n}{2} \rfloor$. We refer to the mixing proportion as *level* (α). This regime generalizes the entire contamination landscape in statistics. The specific choice of parameters for Dirichlet is taken as $(5, 3, 5)$. Perhaps the most interesting observation from our huge body of experiments is some of the near-accurate reconstructions.

In the case of the MNIST data set, even at a significant level of contamination, the reconstruction errors continue to converge to a near-zero value. The corresponding reconstructed samples are of remarkably sound resolution, given the regenerative capability of WAEs in general [Fig. 2.14]. We also study the effect of varying α on reconstructed image quality [Fig. 2.25], which hints at tolerable levels of contamination. Simulations on Five-Gaussian and Swiss roll data also showcase the inherent denoising capability of WAE-MMD [Fig. 2.23, 2.24], even in the presence of extreme outliers. Remarkably, based on our prescribed networks, WAEs recover denoised reconstructions of Five-Gaussian samples under skewed noise [Fig. 2.15].

2.7 Discussion

In this chapter, we provide statistical guarantees regarding the concurrent tasks Wasserstein autoencoders carry out, i.e., achieving both the latent and reconstruction benchmark laws. Under probabilistic characterization of the input data, we establish deterministic upper bounds to both losses. Our non-parametric estimation approach caters to both WAE architectures, namely WAE-MMD and WAE-GAN. In the process, we find out sufficient properties an encoder must possess to become information preserving. This notion further enables us to prescribe to a practitioner the architectural specifications of an ideal encoder network. Deployment of such a network, in turn, aids the latter process of reconstruction. In a WAE-MMD framework, we explore the sufficient conditions the deployed kernel needs to satisfy to ensure latent consistency. We put to test our theoretical findings in simulations based on real and synthetic data sets. The phenomenon of information preservation in the latent space is fascinating to witness in the flesh. The encoded distributions tend to maximize their alignment with the latent target without losing the local geometric properties of the input. Similarly, we recommend decoder frameworks that achieve near-perfect regeneration. Such results are also substantiated by accompanying numerical experiments that show sharply decaying losses over varying sample sizes. Finally, in a density estimation setup, we test the degree of robustness WAEs hold naturally against distribution shifts without additional regularization.

We dedicate the rest of the section to pointing out possibilities that our theoretical framework spawns. The first question arises regarding the modes of the distributions involved. Regenerated samples using a WAE are not known to be plagued by 'mode collapse' as severely as vanilla GAN outputs. However, it is not uncommon for real data distributions (μ) to have non-convex support. In such a case, the OT (due to Brenier) map between the latent distribution and μ will mostly be discontinuous. Moreover, NN-based transforms often fail to universally approximate such discontinuous functions. This may lead to significant mismatches between the supports of μ and $D_{\#}\rho$, however good the representative samples may be. As such, there always lies an innate possibility of missing out on modes of μ, even if the effect is benign to the eye. The question also involves the role of underlying divergences. In generative modeling, they are typically judged based on their capacity of *mode covering* and *mode seeking* (Li and Farnia, 2023). A mode-covering divergence tends to prioritize the spread of masses to all target modes and, as a result, generates out-of-sample observations. While mode-seeking distances avoid doing so, their conservative mass assignment leads to the model missing out on one or several modes. Unfortunately, both TV (weakly mode-seeking) and JS (uniformly mode-seeking) lean toward the latter characterization. As such, in case ρ is multi-modal and there exists a mismatch between the number of modes of μ (unknown) and ρ, WAEs operate under a heightened risk of losing information on some modes. In our non-parametric setup, we do not specify the modality of input and latent distributions.

This creates an interesting prospect to study the effect of varying numbers of modes in ρ on information preservation and reconstruction. Future work may also look into the tolerable modality of input laws in a WAE-MMD before mode collapse transcends benignity. This is particularly intriguing since the mode-seeking properties of MMD remain unexplored.

Code availability

All codes, along with implementation details, can be found in the following repository https://github.com/Thecoder1012/Decons_Wae.

2.8 Appendix: Proofs and Experimental Details

2.8.1 A Note on the Optimal Latent Dimension

The problem that unites theorists and practitioners in shared discomfort is the precise prescription of the latent dimension k. The question remains simple: *Given a set of samples from a distribution, what should be the extent of dimensionality reduction (DR) such that they can be reconstructed?*. In generative exercises, however, the data distribution lies unknown, unlike the ambient dimension of its support. The answer should be multifaceted since there lie several entwined aspects that contribute to the complexity.

The first hint comes from the input data dimension itself. Signals from naturally occurring events are mere instantiations of underlying random processes. Variability in a set of observations is rooted in this very idea. The explanatory attributes and their corresponding directions encapsulate this variation, giving rise to the notion of 'dimensionality' of the data. As characterized by Bennett (1969), this quantity is formally known as the *embedding* (ambient) dimension (Eneva et al., 2002). However, dealing with high-dimensional real datasets (e.g., images), we have come to observe that such a space tends to have a lower-dimensional structure (typically submanifolds $\mathcal{M}$) where most of the variation lies, with a high probability (Fefferman et al., 2016). The smaller set of directions this 'Manifold Hypothesis' points at is called the *intrinsic dimension* (ID). While there is significant disagreement between authors regarding the exact definition of the same (e.g., Minkowski dimension), we recognize ID as the topological dimension of $\mathcal{M}$. Several attempts have been made to estimate its ID given the data distribution (Facco et al., 2017; Levina and Bickel, 2004; Pope et al., 2021). We emphasize the importance of such an intrinsic pattern of the signal to reflect on the encoded law as well. If one goes by the notion of ID being the set of independent dimensions that capture most of the variation in the dataset, k should be a near-estimate of it.

Since WAEs are restricted to reconstructing input observations, they must preserve as much *information* as possible while encoding. In our density estimation regime, the notion of 'information' is somewhat different from that offered by geometry. While the responsibility

to preserve local and broader geometric signatures (based on topologies) lies with the encoder transform, our impression of the statistical information being conserved is that 'the estimates perform with comparable accuracy even after being pushed forward'. Observe that the necessity to learn a latent representation puts an upper bound on the encoded dimension. At the same time, the need to preserve information hints at the existence of a lower bound. As such, the dimensions in between these two extremities invoke a trade-off between the accuracy of achieved *representation* and the amount of information lost.

Along with this discussion comes the call to clarify what we mean by a good representation. Though WAEs prioritize the task of regeneration, one must not forget the roots of its predecessors in learning a *disentangled* representation. Without a robust definition, the idea of disentanglement is marred by subjectivity. Most, however, deem it as the process of compartmentalizing information into groups of independent 'semantic' attributes (Bengio et al., 2013; Higgins et al., 2018). The underlying assumption being $\mathcal{M} = \bigcup_{j=1}^{v} \mathcal{M}_j$, where v is the number of such groups and $\mathcal{M}_j$ are the support submanifolds. The notion of independence may be softened to 'uncorrelatedness' in case group actions are linear (Higgins et al., 2018). Yu et al. (2020) argues that, additionally, such representations should be between-class heterogeneous and within-class homogeneous to the greatest extent. However, based on human perception, no measurement of this extent can summarize the whole picture (Do and Tran, 2020). This discussion finds great motivation in Rubenstein et al. (2018)'s experiments showing WAEs as efficient representation learners. With further regularization on the latent space, we may expect to enhance the efficiency in both static (Gaujac et al., 2021) and dynamic (Han et al., 2021) data regimes. From a statistical viewpoint, we understand disentanglement as the process of attaining a distribution with a block diagonal (axis-aligned as a special case) dispersion matrix. This should ideally contribute to the characterization of the latent distribution. In other words, a disentangled law will be our key to the latent dimension. However, finding such a law, devoid of inductive biases, in an unsupervised setting is theoretically impossible (Locatello et al., 2019).

It is evident that a typical WAE model, during encoding, performs a nonlinear dimensionality reduction. The standard convolutional architecture carries out a feature aggregation in the process that is intractable and is not expected to attain a disentangled law without additional regularization (Kim and Mnih, 2018; Mathieu et al., 2019). Thus, instead of pursuing the optimal value of k directly, we turn our focus to the transformation induced by the encoder. The resilience of such functions against *distortion* along with their regularity becomes paramount in our discussion.

2.8.2 Testing Encoded vs. Latent Distributions

To check the efficacy of a WAE-encoding statistically, we perform two-sample non-parametric tests of equality on target latent and encoded observations. Peacock (1983) suggested a multi-

dimensional generalization to the well-known Kolmogorov–Smirnov test, which, however, has high computational complexity. In our study, we identify the test suggested by Fasano and Franceschini (1987) (FF) as a suitable alternative based on its manageable complexity without sacrificing the power and consistency[3]. Many suggestions have been made to improve the reliability of two-sample tests in higher dimensions since. However, given the ease of implementation, we use FF's version of the KS test. To ascertain our findings, we additionally carry out a referral test of equality of distributions, based on kernelized distances between pairs of observations, namely the Cramér test (Baringhaus and Franz, 2004). Unlike FF, the test statistic corresponding to Cramér[4] is not distribution-free, and requires bootstrapping to obtain the p-value. We utilize two of such methods, namely the usual *Monte-Carlo* (MC) and calculating the approximate *eigenvalues* (EV) as the weights to the limiting distribution (of the statistic). Test results on the Five-Gaussian data at 5% level of significance, given $\lambda = 0.8$, are as reported in Table 2.1.

Table 2.1: Two-sample tests of equality on latent and encoded distributions.

Architecture	Latent Distribution	KS test	Cramér test	
			MC	EV
WAE-GAN	Gaussian	✗	✗	✗
	Exponential	✗	✗	✗
WAE-WAE	Gaussian	✗	✗	✗
	Exponential	✗	✗	✗

The decisions 'Accept' and 'Reject' against the null hypothesis that the two distributions are equal are denoted by the symbols (✓) and (✗) respectively.

Though we only establish an upper bound to the latent loss, it is apparent that there lies an optimization error due to the minimum distance estimate. As such, under a metrizing measure of discrepancy, the target latent law and the encoded estimate must be distinct in distribution. The rejection of the null hypothesis corroborates the same observation. It is not unusual to arrive at such a contrasting conclusion, as optimization errors in deep generative models often become large enough to overwhelm significance testing, e.g., in GANs, transformations between standard random variables often fail to meet the desired test output despite generating reliable pseudo-random replicates (Dutta et al., 2024).

It becomes even clearer looking at the histograms corresponding to samples from the two, overlaid. The interesting observation from Fig. 2.16, 2.17 is the visual manifestation of information preservation. Semantic information, in the form of cluster structures originally present in the dataset, remains intact in encoded distributions while trying to maximize similarity with their target counterparts. The evolution of this 'maximization' is clear from

[3]We follow the fasano.franceschini.test implementation (Puritz et al., 2021) on R.

[4]We implement the cramer package (Franz, 2006) in R with underlying kernel specified to $\kappa(z) = \sqrt{z}/2$.

the histograms obtained over epochs. Another viewpoint that attests to this finding is the quantile-quantile plot [Fig. 2.20] of the marginals.

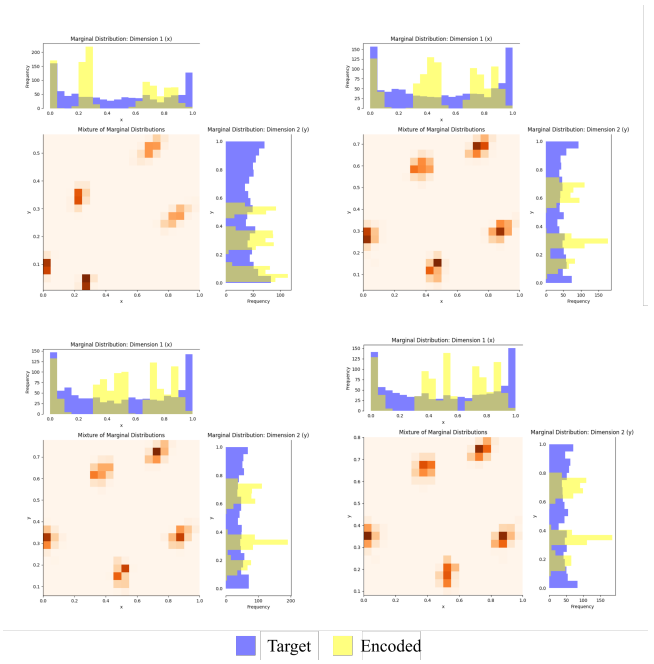


Figure 2.16: Concentration of bin estimates corresponding to Five-Gaussian data under ReLU encoders (yellow) against target latent Beta$(0.5, 0.8)$ copula (blue), over epochs $(200, 800, 1400, 2000)$ (left to right from top) in a WAE-GAN setup.

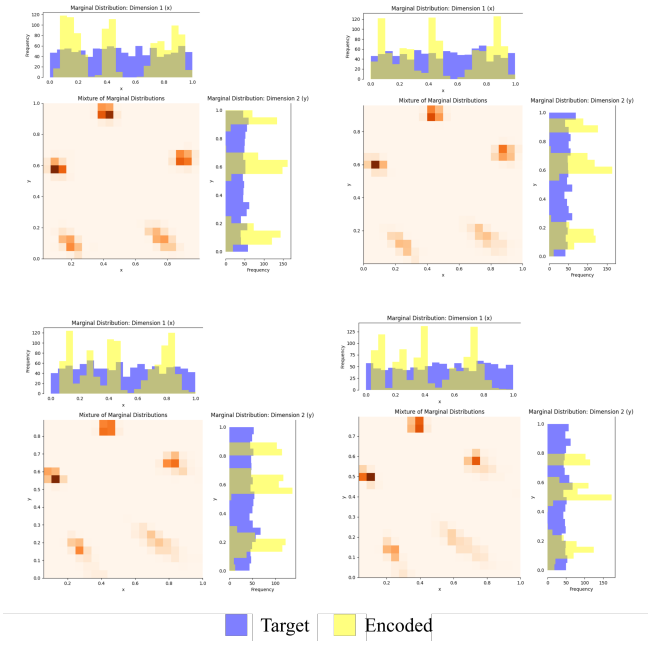


Figure 2.17: Concentration of bin estimates corresponding to Five-Gaussian under ReLU encoders (yellow), given latent bivariate Gaussian distribution (blue), over epochs $(200, 800, 1400, 1800)$ (left to right from top) in a WAE-MMD setup with regularization $\lambda = 0.1$.

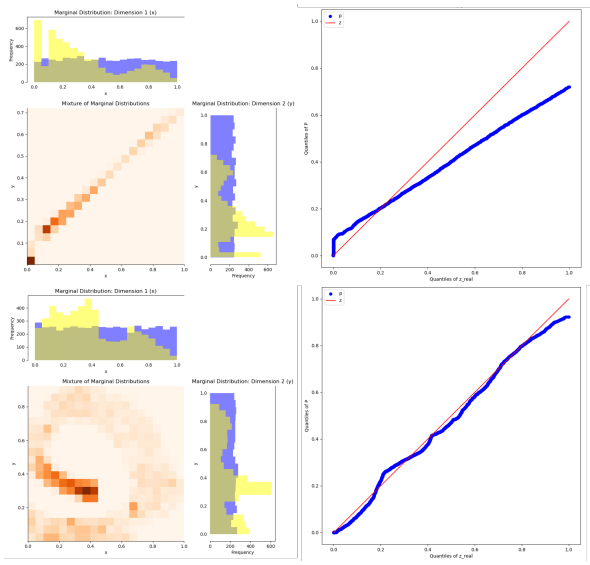


Figure 2.18: Concentration of bin estimates (yellow) against latent Gaussian distribution (blue) and corresponding QQ plots of marginals (upper), for epochs 500 (top row) and 4000 (bottom row) for Swiss roll data. Evidently, the encoded distribution preserves information from the input data and matches the target marginals simultaneously.

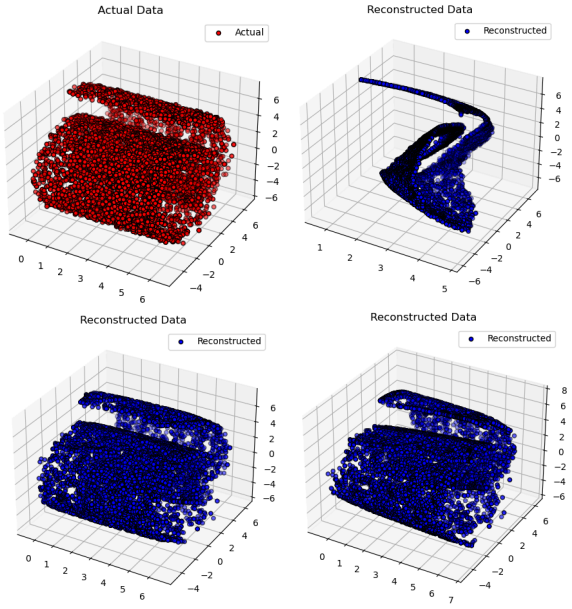
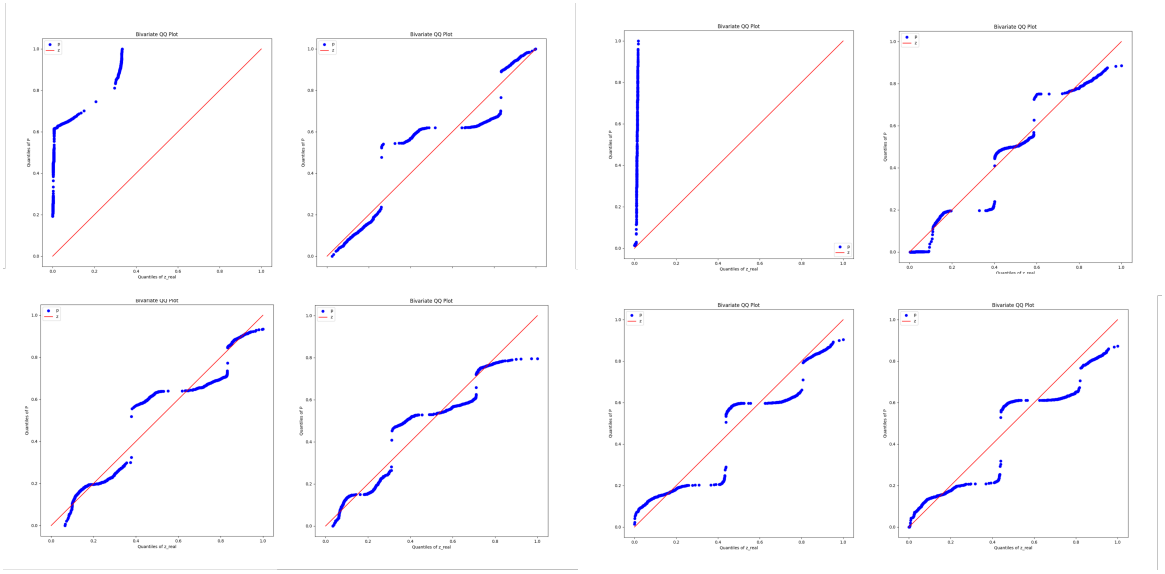
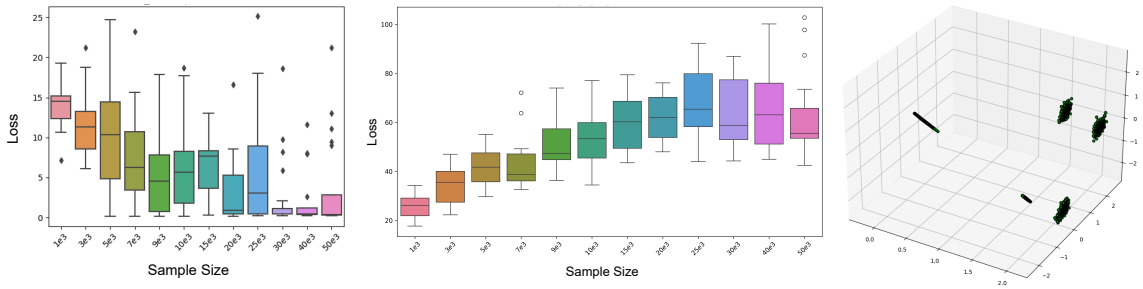


Figure 2.19: Actual Swiss roll data (top left) vs reconstructed samples ($n = 10000$) after epochs $(1000, 4000, 8000)$ (clockwise from top right) under MMD latent loss.



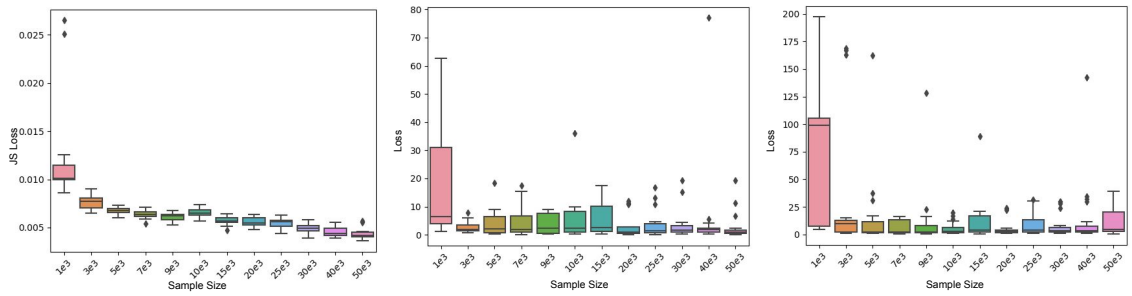
(a) Gaussian (b) Beta Copula

Figure 2.20: Evolution of information preservation over epochs $(0, 200, 1000, 1800)$ (clockwise) based on the propagation of quantile-quantile (QQ) plots of marginals corresponding to encoded (blue) vs latent distribution (red) under ReLU encoders given Five-Gaussian input data, in a WAE-MMD setup with regularization $\lambda = 0.1$.



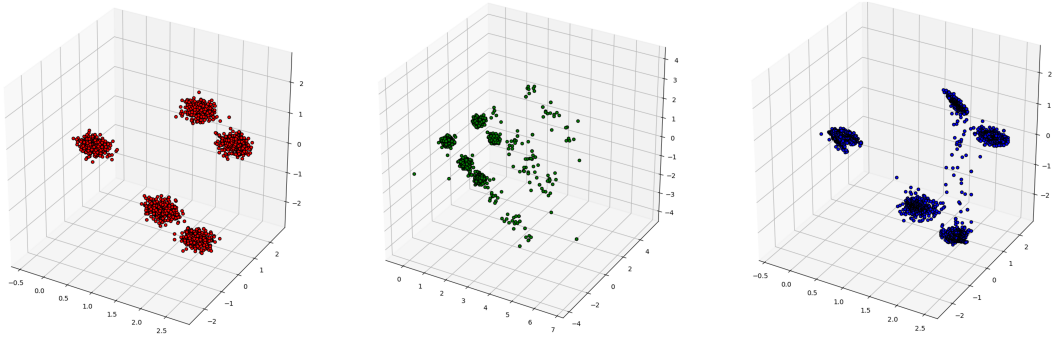
(a) (b) (c)

Figure 2.21: Sample corrected ($\times n^{\frac{1}{3}}$) Wasserstein reconstruction error corresponding to WAE-MMD for Five-Gaussian input data using (a) a decoder that follows the architecture of Theorem 5.2, and (b) one that violates the width criteria therein, having a comparable number of parameters. (c) Regenerated sample from the latter model after 4000 epochs. The second model does not exhibit accurate reconstruction, and the associated errors follow a much slower convergence rate in the process.



(a) Gaussian (b) Beta (c) Exponential

Figure 2.22: Reconstruction error of Five-Gaussian data under (a) JS and (b), (c) sample corrected ($\times n^{\frac{1}{2}}$) MMD latent loss, using GroupSort encoders (grouping 2).



(a) Actual Data (b) Contaminated Data (c) Reconstructed Data

Figure 2.23: Reconstructed samples ($n = 10000$) from Five-Gaussian dataset with 10% observations contaminated at level 0.2, under MMD latent loss. The corrupting distribution remains standard tri-variate Cauchy.

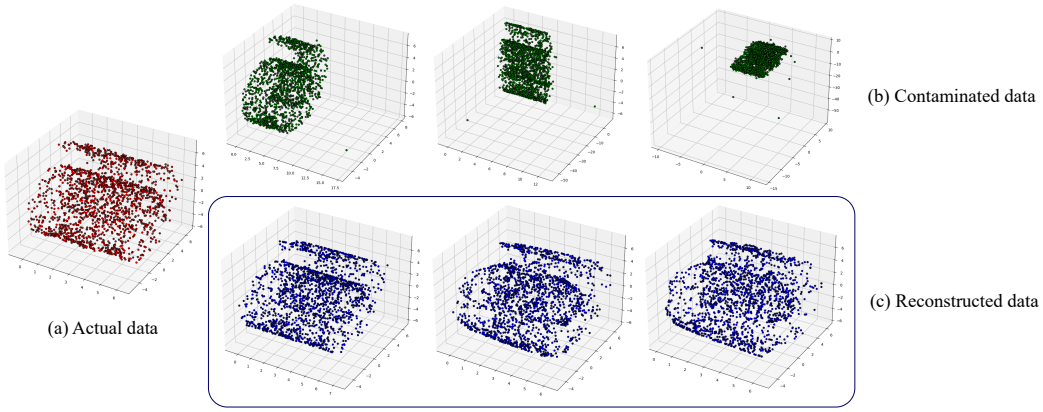
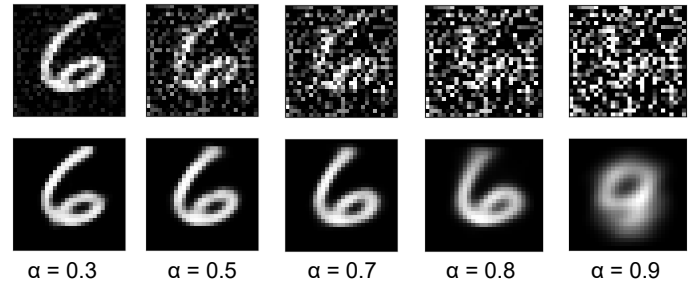
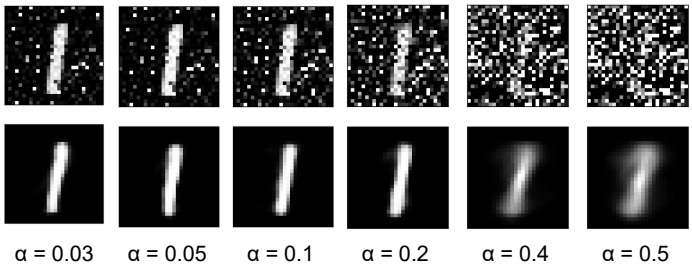


Figure 2.24: Reconstructed samples ($n = 2000$) from Swiss roll dataset with 10% observations contaminated at level 0.01 & 0.1 (left to right) and 20% observations contaminated at level 0.1, under MMD latent loss. The corrupting distribution is taken to be standard tri-variate Cauchy.



(a) Gaussian



(b) Cauchy

Figure 2.25: Reconstructed samples ($n = 10,000$) from Five-Gaussian dataset with 10% observations contaminated at level 0.2, under MMD latent loss. The corrupting distribution is taken to be standard tri-variate Cauchy.

2.8.3 Proof of Lemma 2.1

Let $E : \mathcal{X} \to \mathcal{Z}$ and $D : \mathcal{Z} \to \mathcal{X}$ be measurable maps— namely a encoder-decoder pair— that satisfy

$$W_{c_x}^1(\mu, (D \circ E)_{\#}\mu) + \lambda.\Omega(E_{\#}\mu, \rho) = 0, \tag{2.13}$$

given $\lambda > 0$. Since Ω is a divergence metric metrizing the underlying class of probability distributions, it is implied that $E_{\#}\mu = \rho$. As such, we observe a lossless encoding in the population sense. Now,

$$E_{\#}\mu = \rho \implies (D \circ E)_{\#}\mu = D_{\#}\rho \overset{(1)}{\implies} \mu = D_{\#}\rho,$$

where (1) is due to (2.13). This hints towards an absolute information preservation in reconstruction based on the fact that $D \circ E = \mathrm{id}_X$ a.e.

Given a probability space automorphism φ, $\forall A \in \mathcal{Z}$

$$\rho(\varphi(A)) = \rho(\varphi^{-1}(\varphi(A))) = \rho(A),$$

since φ^{-1} also becomes an automorphism. Hence,

$$\begin{aligned} & W_{c_x}^1(\mu, (D \circ \varphi) \circ (\varphi^{-1} \circ E)_{\#}\mu) + \lambda.\Omega((\varphi^{-1} \circ E)_{\#}\mu, \rho) \\ = & W_{c_x}^1(\mu, (D \circ (\varphi \circ \varphi^{-1}) \circ E)_{\#}\mu) + \lambda.\Omega(\varphi_{\#}^{-1}(E_{\#}\mu), \rho) \\ = & W_{c_x}^1(\mu, (D \circ E)_{\#}\mu) + \lambda.\Omega(\varphi_{\#}^{-1}\rho, \rho) = 0. \end{aligned}$$

As such, the encoder-decoder pair $(\varphi^{-1} \circ E, D \circ \varphi)$ is also a zero-solution of the population loss function. Also, in case $\varphi \neq \mathrm{id}$, the pair clearly differs from (E, D) a.e.

2.8.4 Proof of Theorem 2.1

Given a transform $g \in \mathscr{F}_U(\mathcal{X}, \mathcal{Z})$, the information dissipated can be decomposed using the triangle inequality as follows

$$d_{\mathcal{H}}\left(g_{\#}\tilde{\mu}_n, \widehat{(g_{\#}\mu)}_m\right) \leq d_{\mathcal{H}}\left(g_{\#}\tilde{\mu}_n, g_{\#}\mu\right) + d_{\mathcal{H}}\left(g_{\#}\mu, \widehat{(g_{\#}\mu)}_m\right). \tag{2.14}$$

Here, $\tilde{\mu}_n$ denotes the RIK estimator (see Definition 8), defined as $\frac{d\tilde{\mu}_n}{dx} = \frac{1}{nh^d}\sum_{i=1}^n \kappa\left(\frac{x}{h}, \frac{x_i}{h}\right) = \hat{p}_h(x)$, $x \in \Omega_x$ where h is the bandwidth. Also, since the kernels are bounded, there exists $B > 0$ such that $\sup_{\Omega_x} \kappa(\cdot, \cdot) = B \leq 1$. In most cases, with choices of kernels being distributions themselves, the modal values tend to satisfy this criterion. Since members of $\mathcal{H}$ are bounded,

total variation turns out to be a natural upper bound for the associated loss. In particular,

$$\begin{aligned} d_{\mathcal{H}}\left(g_{\#}\tilde{\mu}_n, g_{\#}\mu\right) &= \sup_{h\in\mathcal{H}} \int h(z)\, d(g_{\#}\tilde{\mu}_n - g_{\#}\mu) \\ &= U \sup_{h'\in\frac{1}{U}\mathcal{H}\circ g} \int h^{'}(x)\, d(\tilde{\mu}_n - \mu) \\ &\leq \frac{U}{2} \int |\hat{p}_h(x) - p_\mu(x)| dx \qquad (2.15) \\ &\leq \frac{U}{2} \left\{ \int |\hat{p}_h(x) - \mathbb{E}[\hat{p}_h(x)]| dx + \left\| \mathbb{E}[\hat{p}_h(x)] - p_\mu \right\|_1 \right\}, \qquad (2.16) \end{aligned}$$

where $\mathcal{H}\circ g = \{h\circ g : h\in\mathcal{H}\}$ such that $\|h\circ g\|_\infty = \max_x |h(g(x))| < \infty$. The first inequality is obtained by taking the supremum over all bounded real-valued functions on $\mathcal{X}$. As such, it is sufficient to find the concentration of $\hat{p}_h$ around p_μ under the essential supremum norm.

The bias term under the norm satisfies

$$\begin{aligned} \mathbb{E}[\hat{p}_h(x)] - p_\mu(x) &= \frac{1}{h^d} \int \kappa\left(\frac{x}{h}, \frac{y}{h}\right) p_\mu(y) dy - p_\mu(x) \\ &= \int \kappa\left(\frac{x}{h}, \frac{x}{h} - u\right) [p_\mu(x - hu) - p_\mu(x)] du \qquad (2.17) \\ &= \int \kappa\left(\frac{x}{h}, \frac{x}{h} - u\right) \sum_{|\alpha|\leq m_x - 1} \frac{D^\alpha p_\mu(x)}{\alpha!} (-uh)^\alpha du \qquad (2.18) \\ &\quad + m_x \int \kappa\left(\frac{x}{h}, \frac{x}{h} - u\right) \sum_{|\alpha| = m_x} \frac{(-uh)^\alpha}{\alpha!} \int_0^1 (1-t)^{m_x - 1} D^\alpha p_\mu(x - tuh) dt \\ &= \int \int_0^1 \kappa\left(\frac{x}{h}, \frac{x}{h} - u\right) (-u)^{m_x} h^{m_x} \frac{(1-t)^{m_x-1}}{(m_x-1)!} D^{m_x} p_\mu(x - tuh) dt\, du, \qquad (2.19) \end{aligned}$$

where (2.17) is obtained using the change in variables $\frac{y}{h} = \frac{x}{h} - u$. In (2.18), we use Taylor's expansion for multivariate functions, in particular, $p_\mu \in \mathcal{W}_R^{m_x, p}(\Omega_x)$. The first part of the sum vanishes due to the regularity of underlying kernels (see Definition 8). Now, using Minkowski's inequality for integrals, given $p = 1$ we get

$$\begin{aligned} \left\| \mathbb{E}[\hat{p}_h(x)] - p_\mu \right\|_1 &\leq h^{m_x} \|D^{m_x} p_\mu\|_1 \int \sup_v |\kappa(v, v-u)| |u|^{m_x} du \int_0^1 \frac{(1-t)^{m_x-1}}{(m_x-1)!} dt \\ &\lesssim h^{m_x}, \qquad (2.20) \end{aligned}$$

again due to the regularity of κ and Assumption 1.

Let us define $\mathscr{K} = \left\{ f(x,\cdot) = \frac{1}{h^d}\kappa\left(\frac{x}{h}, \frac{\cdot}{h}\right) : x\in\Omega_x \right\}$. Now, given the invariance of the

underlying kernels κ, the bracketing number of $\mathscr{K}$ turns out to satisfy

$$\mathcal{N}_{[]}(\mathscr{K}, L_1(\mu), \epsilon) \leq E_\kappa \left(\frac{U\sqrt{d}B_x}{h^{d+1}\epsilon} \right)^d,$$

where $E_\kappa > 0$ is an universal constant depending on d (van der Vaart (2000), Example 19.7). Also, observe that

$$\mathrm{Var}(f) \leq \int f^2 d\mu \leq \|f\|_\infty \int |f| d\mu \leq \frac{B}{h^d},$$

where the last inequality utilizes the fact $\sup_f \mathbb{E}|f| \leq \mathbb{E}[\sup_f |f|]$. Hence, using Bernstein's inequality (see Yukich (1985) for details regarding such a bracketing argument) we infer that

$$\mathbb{P}^n \left(\sup_x |\hat{p}_h(x) - \mathbb{E}[\hat{p}_h(x)]| > \epsilon \right) \leq 4E_\kappa \left(\frac{U\sqrt{d}B_x}{h^{d+1}\epsilon} \right)^d \exp\left\{ -\frac{En\epsilon^2 h^d}{B} \right\}, \tag{2.21}$$

where $E > 0$ is an universal constant[5] and $0 < \epsilon \leq \frac{2}{3}$. This enables us to state a concentration bound for $d_{\mathcal{H}}\left(g_\# \tilde{\mu}_n, g_\# \mu\right)$ readily using (2.16) and (2.20).

Bounding the expected estimation error based on the translated data (second term in 2.14) is comparatively straightforward. We present the refined Dudley's entropy integral, which becomes the cornerstone of the rest of the proof.

Lemma 2.6. *Given a symmetric class of functions $\mathcal{H}$, satisfying $\sup_{h\in\mathcal{H}} \|h\|_\infty \leq M$, $M > 0$ we have*

$$\mathbb{E}[d_{\mathcal{H}}\left(\hat{\rho}_m, \rho\right)] \leq 2 \inf_{\delta \in (0,M)} \left(2\delta + \frac{12}{\sqrt{m}} \int_\delta^M \sqrt{\log \mathcal{N}\left(\mathcal{H}, \|\cdot\|_\infty, \epsilon\right)} d\epsilon \right),$$

where $\rho \in \mathcal{P}(\mathcal{Z})$.

As such, given that the entropy corresponding to the underlying critic functions satisfy polynomial discrimination, we obtain an upper bound corresponding to the infimum as follows

$$\mathbb{E}\left[d_{\mathcal{H}}\left(g_\# \mu, \widehat{(g_\# \mu)}_m \right) \right] \lesssim m^{-\frac{1}{q\vee 2}}. \tag{2.22}$$

To obtain a probabilistic concentration inequality corresponding to the same error, observe that

$$d_{\mathcal{H}}\left(g_\# \mu, \widehat{(g_\# \mu)}_m \right) = \sup_{h\in\mathcal{H}} \left\{ \frac{1}{m} \sum_{i=1}^m h(Y_i) - \mathbb{E}_{g_\# \mu} h \right\} := W(Y_1, \cdots, Y_m),$$

[5] Lafferty et al. (2008) derive the sharp value of E under Lipschitz continuous kernels.

given $Y_1, \cdots, Y_m \sim g_{\#}\mu$. As such, for $y_1, \cdots, y_m, y_m^{'}$

$$\begin{aligned}
&\left|W(y_1, \cdots, y_m) - W(y_1, \cdots, y_m^{'})\right| \\
=&\frac{1}{m}\left|\sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \left(h(y_i) - \mathbb{E}_{g_{\#}\mu} h\right) - \sup_{h^{'}\in\mathcal{H}} \sum_{i=1}^{m-1} \left(h^{'}(y_i) - \mathbb{E}_{g_{\#}\mu} h^{'}\right) + h^{'}(y_m^{'}) - \mathbb{E}_{g_{\#}\mu} h^{'}\right| \\
\leq&\frac{1}{m}\left|\sup_{h\in\mathcal{H}} h(y_m) - h(y_m^{'})\right| \leq \frac{2M}{m}.
\end{aligned}$$

Applying McDiarmid's inequality,

$$d_{\mathcal{H}}\left(g_{\#}\mu, \widehat{(g_{\#}\mu)}_m\right) \leq \epsilon + \mathcal{O}(m^{-\frac{1}{q\vee 2}})$$

holds with probability $\geq 1 - \exp\left\{-\frac{n\epsilon^2}{2M^2}\right\}$, where $\epsilon > 0$. This bound, along with (2.21) ensure the existence of constants l, E_1, E_2 and $E_3 > 0$ that proof the theorem. Observe that the bound satisfies for arbitrary choices of h. To ensure that the realized error is indeed $o(1)$ with high probability, we specify $h := h_n = (\frac{1}{n})^{\xi}$ such that $\xi \geq \frac{1}{d}$.

2.8.5 Proof of Lemma 2.2

Given any $\phi \in \Phi(W, L)_d^k$ and $\mu_1, \mu_2 \in \mathcal{P}(\mathcal{X})$, by the definition of IPMs

$$\begin{aligned}
&d_{\mathcal{L}_{c_z}^1}(\phi_{\#}\mu_1, \phi_{\#}\mu_2) \\
=&\sup_{l\in\mathcal{L}_{c_z}^1} \mathbb{E}_{\mu_1}[l\circ\phi] - \mathbb{E}_{\mu_2}[l\circ\phi] \\
=&\sup_{l\in\mathcal{L}_{c_z}^1} \{\mathbb{E}_{\mu_1}[l\circ\phi] - \mathbb{E}_{\mu_1}[l\circ g] + \mathbb{E}_{\mu_2}[l\circ g] - \mathbb{E}_{\mu_2}[l\circ\phi] + \mathbb{E}_{\mu_1}[l\circ g] - \mathbb{E}_{\mu_2}[l\circ g]\} \\
\leq&\sup_{l\in\mathcal{L}_{c_z}^1} \{\mathbb{E}_{\mu_1}|l\circ\phi - l\circ g| + \mathbb{E}_{\mu_2}|l\circ g - l\circ\phi| + \mathbb{E}_{\mu_1}[l\circ g] - \mathbb{E}_{\mu_2}[l\circ g]\} \\
\leq&2\|\phi - g\|_{\infty} + \sup_{l\in\mathcal{L}_{c_z}^1} \mathbb{E}_{\mu_1}[l\circ g] - \mathbb{E}_{\mu_2}[l\circ g],
\end{aligned}$$

where the first inequality is due to Jensen's inequality and in the second one we use the fact that $l \in \mathcal{L}_{c_z}^1$, given $c_z \equiv L_1$. The arbitrary choice of $g \in \mathscr{F}(\mathcal{X}, \mathcal{Z})$ makes the result hold for the infimum as well. As such,

$$d_{\mathcal{L}_{c_z}^1}(\phi_{\#}\mu_1, \phi_{\#}\mu_2) \leq 2 \inf_{g\in\mathscr{F}(\mathcal{X},\mathcal{Z})} \|\phi - g\|_{\infty} + d_{\mathcal{L}_{c_z}^1}(g_{\#}\mu_1, g_{\#}\mu_2). \tag{2.23}$$

Now, under the critic $\mathcal{F}$, $\forall \varepsilon > 0$ there exists $f_\varepsilon \in \mathcal{F}$ such that

$$\begin{aligned} d_{\mathcal{F}}(\phi_{\#}\mu_1, \phi_{\#}\mu_2) &\leq \mathbb{E}_{\phi_{\#}\mu_1}[f_\varepsilon] - \mathbb{E}_{\phi_{\#}\mu_2}[f_\varepsilon] + \varepsilon \\ &= \mathbb{E}_{\phi_{\#}\mu_1}[f_\varepsilon - l] + \mathbb{E}_{\phi_{\#}\mu_2}[l - f_\varepsilon] + \mathbb{E}_{\phi_{\#}\mu_1}[l] - \mathbb{E}_{\phi_{\#}\mu_2}[l] + \varepsilon \\ &\leq 2\|f_\varepsilon - l\|_\infty + \sup_{l \in \mathcal{L}^1_{c_z}} \mathbb{E}_{\phi_{\#}\mu_1}[l] - \mathbb{E}_{\phi_{\#}\mu_2}[l] + \varepsilon. \end{aligned}$$

Similarly, taking infimum over all such choices of l

$$d_{\mathcal{F}}(\phi_{\#}\mu_1, \phi_{\#}\mu_2) \leq 2\mathcal{E}(\mathcal{F}, \mathcal{L}^1_{c_z}) + d_{\mathcal{L}^1_{c_z}}(\phi_{\#}\mu_1, \phi_{\#}\mu_2) + \varepsilon. \tag{2.24}$$

The inequalities (2.23) and (2.24) together proof the lemma.

2.8.6 Proof of Theorem 2.2

Given any NN-induced map ϕ, using the triangle inequality on MMDs, one can write

$$d_{\mathcal{H}_\kappa}\left(\phi_{\#}\hat{\mu}_n, \widehat{(\phi_{\#}\mu)}_m\right) \leq \underbrace{d_{\mathcal{H}_\kappa}\left(\phi_{\#}\hat{\mu}_n, \phi_{\#}\mu\right)}_{\text{(i)}} + \underbrace{d_{\mathcal{H}_\kappa}\left(\phi_{\#}\mu, \widehat{(\phi_{\#}\mu)}_m\right)}_{\text{(ii)}}. \tag{2.25}$$

Since the underlying kernels are bounded to begin with, an immediate upper bound for (ii) might be: $d_{\mathcal{H}_\kappa}(\phi_{\#}\mu, \widehat{(\phi_{\#}\mu)}_m) \leq \sqrt{\sup_{z\in\Omega_z} \kappa(z,z)}\, d_{\mathrm{TV}}(\phi_{\#}\mu, \widehat{(\phi_{\#}\mu)}_m)$ (Sriperumbudur et al. (2009), Theorem 14 (ii)). Being larger in general, TV may enforce information preservation onto MMDs. However, the very property of boundedness of the kernels enables us to show the concentration of empirical measures under MMD as well. Observe that, for bounded kernels, MMD satisfies the bounded difference inequality with the universal upper bound $2m^{-1}\sqrt{\sup_{z\in\Omega_z} \kappa(z,z)}$. As such, using McDiarmid's inequality

$$\mathbb{P}\left(d_{\mathcal{H}_\kappa}(\phi_{\#}\mu, \widehat{(\phi_{\#}\mu)}_m) \leq \mathbb{E}[d_{\mathcal{H}_\kappa}(\phi_{\#}\mu, \widehat{(\phi_{\#}\mu)}_m)] + t\right) \geq 1 - e^{-\frac{mt^2}{2C_\kappa}}, \tag{2.26}$$

where C_κ is a positive constant such that $\sup_{z\in\Omega_z} \kappa(z,z) \leq C_\kappa$. Furthermore, we observe that $\mathbb{E}\left[d_{\mathcal{H}_\kappa}(\phi_{\#}\mu, \widehat{(\phi_{\#}\mu)}_m)\right] \leq \left[\mathbb{E}\, d^2_{\mathcal{H}_\kappa}(\phi_{\#}\mu, \widehat{(\phi_{\#}\mu)}_m)\right]^{\frac{1}{2}} \leq \sqrt{\frac{2C_k}{m}}$ (Briol et al. (2019), lemma 2). In pursuit of establishing an upper bound to (i), one needs additional enforcement. The first of which is presented as the following lemma.

Lemma 2.7. *Given arbitrary $\phi, g \in \mathscr{F}(\mathcal{X}, \mathcal{Z})$ and $\mu_1, \mu_2 \in \mathcal{P}_\kappa(\mathcal{X})$ such that the underlying kernel function κ is strongly invariant, there exists a constant $D > 0$ (dependant on κ and the latent dimension) for which*

$$d^2_{\mathcal{H}_\kappa}(\phi_{\#}\mu_1, \phi_{\#}\mu_2) \leq D\|\phi - g\|_\infty + d^2_{\mathcal{H}_\kappa}(g_{\#}\mu_1, g_{\#}\mu_2).$$

Proof of lemma 2.7. Observe that

$$
\begin{aligned}
&\left|d^2_{\mathcal{H}_\kappa}(\phi_\#\mu_1,\phi_\#\mu_2)-d^2_{\mathcal{H}_\kappa}(g_\#\mu_1,g_\#\mu_2)\right| \\
=&\left|\int_{\Omega_x\times\Omega_x}\left[\kappa(\phi(x),\phi(y))-\kappa(g(x),g(y))\right](\mu_1-\mu_2)\otimes(\mu_1-\mu_2)(dxdy)\right| \\
=&\left|\int_{\Omega_x\times\Omega_x}\left[\kappa(\phi(x),\phi(y))-\kappa(g(x),\phi(y))+\kappa(g(x),\phi(y))-\kappa(g(x),g(y))\right](\mu_1-\mu_2)\otimes(\mu_1-\mu_2)(dxdy)\right| \\
\overset{(1)}{=}&\left|\int_{\Omega_x}\left[K(\phi_\#\mu_1-\phi_\#\mu_2)(\phi(x))-K(\phi_\#\mu_1-\phi_\#\mu_2)(g(x))\right](\mu_1-\mu_2)(dx)\right. \\
&\qquad\left.+\int_{\Omega_x}\left[K(g_\#\mu_1-g_\#\mu_2)(\phi(y))-K(g_\#\mu_1-g_\#\mu_2)(g(y))\right](\mu_1-\mu_2)(dy)\right| \\
\leq&\int_{\Omega_x}|K(\phi_\#\mu_1-\phi_\#\mu_2)(\phi(x))-K(\phi_\#\mu_1-\phi_\#\mu_2)(g(x))||\mu_1-\mu_2|(dx) \\
&\qquad+\int_{\Omega_x}|K(g_\#\mu_1-g_\#\mu_2)(\phi(y))-K(g_\#\mu_1-g_\#\mu_2)(g(y))||\mu_1-\mu_2|(dy),
\end{aligned}
$$

where (1) is due to the fact

$$
\begin{aligned}
&\int\kappa(\phi(x),\phi(y))(\mu_1-\mu_2)\otimes(\mu_1-\mu_2)(dxdy) \\
=&\int\kappa(\phi(x),y)(\mu_1-\mu_2)\otimes(\phi_\#\mu_1-\phi_\#\mu_2)(dxdy)=\int K(\phi_\#\mu_1-\phi_\#\mu_2)(\phi(x))(\mu_1-\mu_2)(dx).
\end{aligned}
$$

Now,

$$
\begin{aligned}
&|K(\phi_\#\mu_1-\phi_\#\mu_2)(\phi(x))-K(\phi_\#\mu_1-\phi_\#\mu_2)(g(x))| \\
&\leq\int_{\Omega_z}|\kappa(\phi(x),y)-\kappa(g(x),y)||\phi_\#\mu_1-\phi_\#\mu_2|(dy) \\
&\overset{(2)}{\leq}\int_{\Omega_z}\|K(\phi(x))-K(g(x))\|\sqrt{\kappa(y,y)}|\phi_\#\mu_1-\phi_\#\mu_2|(dy) \\
&\overset{(3)}{\lesssim}\|\phi(x)-g(x)\|\int_{\Omega_z}\sqrt{\kappa(y,y)}|\phi_\#\mu_1-\phi_\#\mu_2|(dy),
\end{aligned}
$$

where (2) is due to the reproducing kernel property, coupled with the Cauchy-Schwartz inequality. The strong invariance of the underlying kernel inspires (3). Noticing the quantity under the integral to be finite, we obtain

$$
\begin{aligned}
&\int_{\Omega_x}|K(\phi_\#\mu_1-\phi_\#\mu_2)(\phi(x))-K(\phi_\#\mu_1-\phi_\#\mu_2)(g(x))||\mu_1-\mu_2|(dx) \\
&\lesssim\int_{\Omega_x}\|\phi(x)-g(x)\||\mu_1-\mu_2|(dx)\lesssim\|\phi-g\|_\infty,
\end{aligned}
$$

since the dominating measure is sigma-finite. The suppressed constant is, namely, k (the latent dimension). Similarly, observing $\int_{\Omega_z} \sqrt{\kappa(y,y)}|g_{\#}\mu_1 - g_{\#}\mu_2|(dy) < \infty$ in addition, we conclude

$$\left|d^2_{\mathcal{H}_\kappa}(\phi_{\#}\mu_1, \phi_{\#}\mu_2) - d^2_{\mathcal{H}_\kappa}(g_{\#}\mu_1, g_{\#}\mu_2)\right| \lesssim \|\phi - g\|_\infty.$$

As such, there indeed exists a constant that satisfies the lemma. □

Now, let us choose in particular $g \in \mathscr{F}_U(\mathcal{X}, \mathcal{Z})$. For ease of understanding, we continue with the distributions $\mu_1, \mu_2 \in \mathcal{P}_\kappa(\mathcal{X})$. Using the reproducing property again, we get

$$\begin{aligned} d^2_{\mathcal{H}_\kappa}(g_{\#}\mu_1, g_{\#}\mu_2) &= \int_{\Omega_x \times \Omega_x} \kappa(g(x), g(y))(\mu_1 - \mu_2) \otimes (\mu_1 - \mu_2)(dxdy) \\ &= \int_{\Omega_x} K(g_{\#}\mu_1 - g_{\#}\mu_2)(g(x))(\mu_1 - \mu_2)(dx). \end{aligned}$$

While proving Lemma 2.7, we have observed that the function $K(g_{\#}\mu_1 - g_{\#}\mu_2)$ is Lipschitz continuous with accompanying constant $c_g^{(\mu_1,\mu_2)} = \int_{\Omega_z} \sqrt{\kappa(y,y)}|g_{\#}\mu_1 - g_{\#}\mu_2|(dy)$. We mention that for Energy kernels, the same function rather turns out to be Hölder continuous. As such,

$$\begin{aligned} d^2_{\mathcal{H}_\kappa}(g_{\#}\mu_1, g_{\#}\mu_2) &\leq c_g^{(\mu_1,\mu_2)} \sup_{f \in \mathcal{L}^1_{c_z \equiv L_2}} \left[\int_{\Omega_x} f(g(x))(\mu_1 - \mu_2)(dx)\right] \\ &\overset{(4)}{=} c_g^{(\mu_1,\mu_2)} d_{\mathcal{L}^1_{c_z}}(g_{\#}\mu_1, g_{\#}\mu_2) \leq c_g^{(\mu_1,\mu_2)} U d_{\mathcal{L}^1_{c_x}}(\mu_1, \mu_2), \end{aligned}$$

where (4) is due to the Kantorovitch-Rubinstein duality. Hence, we may write

$$d^2_{\mathcal{H}_\kappa}(\phi_{\#}\hat{\mu}_n, \phi_{\#}\mu) \leq D_n \|\phi - g\|_\infty + c_g^{(\hat{\mu}_n,\mu)} U d_{\mathcal{L}^1_{c_x}}(\hat{\mu}_n, \mu), \tag{2.27}$$

where D_n and $c_g^{(\hat{\mu}_n,\mu)}$ are no longer constants, but sequences based on n that converge to 0 almost surely as $n \to \infty$ (by dominated convergence theorem). In particular,

$$c_g^{(\hat{\mu}_n,\mu)} = c_{g,n} = \int_{\Omega_z} \sqrt{\kappa(y,y)}|g_{\#}\hat{\mu}_n - g_{\#}\mu|(dy) = \int_{\Omega_x} \sqrt{\kappa(g(y),g(y))}|\hat{\mu}_n - \mu|(dy)$$

and $D_n = o(c_{g,n} \vee c_{\phi,n})$. Applying the concentration of $\hat{\mu}_n$ around μ under the metric $d_{\mathcal{L}^1_{c_x}}$, along with the inequalities 2.25 and 2.26 we infer that given $t > 0$

$$\mathbb{P}\left(d_{\mathcal{H}_\kappa}\left(\phi_{\#}\hat{\mu}_n, \widehat{(\phi_{\#}\mu)}_m\right) \leq t + \sqrt{\frac{2C_\kappa}{m}} + \sqrt{D_n\|\phi - g\|_\infty} + \sqrt{\mathcal{O}(c_g^{(\hat{\mu}_n,\mu)}(d^2 n)^{-\frac{1}{d}}) + c_g^{(\hat{\mu}_n,\mu)} U t}\right)$$

remains at least $1 - 2\exp\left\{\left(-\frac{2(m\wedge n)t^2}{B}\right)\right\}$, where $B = \max\{B_x^2, 4C_\kappa\}$. The quantity B_x

symbolises the diameter of Ω_x under the metric c_x.

2.8.7 Proof of Theorem 2.3

The proof takes inspiration from Theorem 3.5 of Lee et al. (2017). First, let us construct $\phi_{1:L+1} = \phi_1 \circ \cdots \circ \phi_{L+1}$ where the individual functions are defined as

$$\phi_i : \mathbb{R}^{N_{i-1}} \to \mathbb{R}^{N_i} \text{ such that } (\phi_i(x))_j = c_{ij0} + \sum_{k=1}^{r_i} c_{ijk}\sigma\left(\langle m_{ijk}, x\rangle + b_{ijk}\right),$$

where $c_{ijk}, b_{ijk} \in \mathbb{R}$ and $m_{ijk} \in \mathbb{R}^{N_{i-1}}$ are model parameters in accordance with Definition 6. Since these are shallow networks, the quantity $r_i, 1 \leq i \leq L+1$ denotes the number of nodes they have. Observe that, following Barron's argument, $f_1 : \mathbb{R}^d \to \mathbb{R}^{N_1}$ can be approximated using ϕ_1 (under the L_2 norm, with respect to μ). The same result holds for all intermediate individual pieces, with respect to measures at their respective domains. Our goal is to approximate the composition of them all. To invoke an induction argument, let us first consider a sequence of nested sets $\{S_i\}_{i=1}^{L+1} \subseteq \mathbb{R}^d$ such that $S_i = S_{i-1} \cap \{x : \phi_{1:i-1}(x) \in \Omega_{i-1}^{s,N_{i-1}}\}$. The measure μ, restricted onto S_i, when pushed-forward by the map $\phi_{1:i-1}$ yields $\mu^i = (\phi_{1:i-1})_{\#}(\mathbb{1}_{S_i}\mu)$ (need not be a probability measure). The support of μ^i is thus obtained to be $\phi_{1:i-1}(S_i) \subseteq \Omega_{i-1}^{s,N_{i-1}}$. Let us denote the Lipschitz constant corresponding to f_i by $\|f_i\|_{\mathcal{B}}$ (Wojtowytsch et al., 2022). Now, given any $\varepsilon > 0$, define $r_i = \lceil \frac{4C_i^2 N_i}{\varepsilon^2} \rceil$. Now,

$$\begin{aligned}
&\left(\int_{\mathbb{R}^d} \mathbb{1}_{S_i}\|f_{1:i} - \phi_{1:i}\|^2 d\mu\right)^{\frac{1}{2}} \\
&\overset{(1)}{\leq} \left(\int_{\mathbb{R}^d} \mathbb{1}_{S_i}\|f_i \circ f_{1:i-1} - f_i \circ \phi_{1:i-1}\|^2 d\mu\right)^{\frac{1}{2}} + \left(\int_{\mathbb{R}^{N_{i-1}}} \|f_i - \phi_i\|^2 d(\phi_{1:i-1})_{\#}(\mathbb{1}_{S_i}\mu)\right)^{\frac{1}{2}} \\
&\leq \|f_i\|_{\mathcal{B}} \left(\int_{\mathbb{R}^d} \mathbb{1}_{S_i}\|f_{1:i-1} - \phi_{1:i-1}\|^2 d\mu\right)^{\frac{1}{2}} + \varepsilon \\
&\overset{(2)}{\leq} \|f_i\|_{\mathcal{B}} \left(\int_{\mathbb{R}^d} \mathbb{1}_{S_{i-1}}\|f_{1:i-1} - \phi_{1:i-1}\|^2 d\mu\right)^{\frac{1}{2}} + \varepsilon \\
&\leq \left[1 + \|f_i\|_{\mathcal{B}} + \|f_i\|_{\mathcal{B}}\|f_{i-1}\|_{\mathcal{B}} + \cdots + \prod_{j=1}^{i} \|f_j\|_{\mathcal{B}}\right] \varepsilon \leq \frac{\left\{\bigvee_1^i \|f_j\|_{\mathcal{B}}\right\}^{i+1} - 1}{\bigvee_1^i \|f_j\|_{\mathcal{B}} - 1}\varepsilon,
\end{aligned}$$

where (1) is due to the triangle inequality. The second term in the same stage turns out to be $\leq \varepsilon$ using Barron's theorem (Barron, 1993; Lee et al., 2017) given our specific choice of r_i. The inequality (2) is based on the filtration offered by S_i. Observe that, in case the maximum of the Lipschitz constants is exactly 1, the upper bound becomes $i\varepsilon$. This implies

that

$$\mu\left(S_{i-1} \cup \{x : \|f_{1:i-1}(x) - \phi_{1:i-1}(x)\| \geq s\}\right) \leq \left[\frac{\left\{\bigvee_1^{i-1} \|f_j\|_{\mathcal{B}}\right\}^{\overline{i-1}+1} - 1}{\bigvee_1^{i-1} \|f_j\|_{\mathcal{B}} - 1}\right]^2 \cdot \frac{\varepsilon^2}{s^2}.$$

As such,

$$\begin{aligned}\mu(S_i) &= \mu(S_{i-1}) - \mu(S_{i-1} \cup \{x : \|f_{1:i-1}(x) - \phi_{1:i-1}(x)\| \geq s\}) \\ &\geq \mu(S_{i-1}) - \left[\frac{\left\{\bigvee_1^{i-1} \|f_j\|_{\mathcal{B}}\right\}^{i} - 1}{\bigvee_1^{i-1} \|f_j\|_{\mathcal{B}} - 1}\right]^2 \cdot \frac{\varepsilon^2}{s^2} \geq 1 - \frac{\varepsilon^2}{s^2} \sum_{l=1}^{i-1} \left[\frac{\left\{\bigvee_1^{l} \|f_j\|_{\mathcal{B}}\right\}^{l+1} - 1}{\bigvee_1^{l} \|f_j\|_{\mathcal{B}} - 1}\right]^2 \end{aligned} \tag{2.28}$$

In other words, there exists $\phi = \phi_{1:L+1}$ such that $\left(\int_{\mathbb{R}^d} \mathbb{1}_S \|f_{1:L+1} - \phi\|^2 d\mu\right)^{\frac{1}{2}} \leq \mathcal{O}(\varepsilon)$ for a set $S \subset \mathbb{R}^d$ satisfying inequality 2.28. Now, owing to the compactness of the base domain and its regularity, the range of $\phi_i = (\phi_{i1}, \phi_{i2}, \cdots, \phi_{iN_i})$ is always contained in a space of finite diameter. Hence,

$$\begin{aligned}\int_{\Omega_x} \|f_{1:i} - \phi_{1:i}\|^2 d\mu &\leq \int_{S_i} \|f_{1:i} - \phi_{1:i}\|^2 d\mu + \int_{S_i^c} \|f_{1:i} - \phi_{1:i}\|^2 d\mu \\ &\leq \left[\frac{\left\{\bigvee_1^{i} \|f_j\|_{\mathcal{B}}\right\}^{i+1} - 1}{\bigvee_1^{i} \|f_j\|_{\mathcal{B}} - 1}\right]^2 \cdot \varepsilon^2 + \mathcal{O}\left(\frac{\varepsilon^2}{s^2}\right).\end{aligned}$$

Taking the square root and choosing i corresponding to the ultimate layer, we prove the result.

2.8.8 Proof of Theorem 2.4

Given an encoder $\phi \in \Phi(W, L)_d^k$ and an empirical distribution corresponding to μ (based on n i.i.d. replicates), let us fragment the realized latent WAE-MMD loss as usual,

$$d_{\mathcal{H}_\kappa}\left(\phi_{\#}\hat{\mu}_n, \rho\right) \leq d_{\mathcal{H}_\kappa}\left(\phi_{\#}\hat{\mu}_n, \widehat{(\phi_{\#}\mu)}_n\right) + d_{\mathcal{H}_\kappa}\left(\widehat{(\phi_{\#}\mu)}_n, \hat{\rho}_n\right) + d_{\mathcal{H}_\kappa}\left(\hat{\rho}_n, \rho\right) \tag{2.29}$$

which holds for all empirical $\hat{\rho}_n$, based on n i.i.d. samples from ρ. Observe that the first quantity is the information dissipated during encoding and can be put under a deterministic upper bound using Theorem 4.5. The second quantity, due to the observation $\left\|\kappa(\cdot, z) - \kappa(\cdot, z')\right\|_{\mathcal{H}_\kappa} \leq 2\sqrt{C_\kappa}\mathbb{1}_{z \neq z'}$, satisfies

$$d_{\mathcal{H}_\kappa}\left(\widehat{(\phi_{\#}\mu)}_n, \hat{\rho}_n\right) \leq \sqrt{C_\kappa} d_{\text{TV}}\left(\widehat{(\phi_{\#}\mu)}_n, \hat{\rho}_n\right). \tag{2.30}$$

Now, recall that $\rho \in \mathcal{P}_\Theta(\Omega_z)$. As such, the observations from ρ, forming its empirical counterpart, are first projected onto the subset of Θ-invariant distributions. This is termed *symmetrization* and the corresponding operator enabling it, $S^\Theta : \mathcal{P}(\mathcal{Z}) \to \mathcal{P}(\mathcal{Z})$ is defined as

$$\mathbb{E}_{S^\Theta[\rho]} f = \int_{\mathcal{Z}} \left[\int_\Theta f(\varphi_\theta(z)) \mu_\Theta(d\theta) \right] d\rho(z) = \mathbb{E}_\rho \mathbb{E}_{\mu_\Theta}[f \circ \varphi_\theta],$$

where μ_Θ is the Haar measure on the compact group Θ and f denotes any bounded measurable function. In other words, the recipe to obtain samples from $S^\Theta[\rho]$ in general is to draw i.i.d. $\{z_i\}_1^n \sim \rho$ and $\{\theta_i\}_1^n \sim \mu_\Theta$ independently, followed by the operation $\varphi_{\theta_j}(z_i)$ (Birrell et al., 2022). This, in turn, enables us to narrow down the class of critic functions under MMD to its Θ-invariant subset $\mathcal{H}_\kappa^\Theta$. We use the same idea to obtain an upper bound to the remaining estimation error in (2.29) as follows. Let,

$$\begin{aligned}
\gamma(z_1, z_2, \cdots, z_n) &= d_{\mathcal{H}_\kappa}(\hat{\rho}_n, \rho) = d_{\mathcal{H}_\kappa^\Theta}(\hat{\rho}_n, \rho) \\
&= \sup_{\|f\|_{\mathcal{H}_\kappa} \leq 1} \left\{ \mathbb{E}_{S^\Theta[\hat{\rho}_n]} f - \mathbb{E}_\rho f \right\} \\
&= \sup_{\|f\|_{\mathcal{H}_\kappa} \leq 1} \left\{ \frac{1}{n|\Theta|} \sum_{i=1}^{n} \sum_{j=1}^{|\Theta|} f(\theta_j z_i) - \mathbb{E}_\rho f \right\} \\
&\leq \sup_{\|f\|_{\mathcal{H}_\kappa} \leq 1} \left| \frac{1}{n|\Theta|} \sum_{i=1}^{n} \sum_{j=1}^{|\Theta|} f(\theta_j z_i) - \mathbb{E}_\rho f \right|. \qquad (2.31)
\end{aligned}$$

To look for the specific constant satisfying the bounded difference property, observe that

$$\begin{aligned}
&\left| \gamma(z_1, \cdots, z_i, \cdots, z_n) - \gamma(z_1, \cdots, z_i^{'}, \cdots, z_n) \right| \\
&\leq \sup_{\|f\|_{\mathcal{H}_\kappa} \leq 1} \left| \frac{1}{n|\Theta|} \sum_{j=1}^{|\Theta|} f(\theta_j z_i) - f(\theta_j z_i^{'}) \right| \\
&\leq \frac{1}{n|\Theta|} \left\{ \left\| \sum_{j=1}^{|\Theta|} K(\theta_j z_i) \right\|_{\mathcal{H}_\kappa} + \left\| \sum_{j=1}^{|\Theta|} K(\theta_j z_i^{'}) \right\|_{\mathcal{H}_\kappa} \right\}, \qquad (2.32)
\end{aligned}$$

where

$$\left\| \sum_{j=1}^{|\Theta|} K(\theta_j z_i) \right\|_{\mathcal{H}_\kappa} = \left[\sum_{j=1}^{|\Theta|} \kappa(\theta_j z_i, \theta_j z_i) + \sum_{j \neq l} \kappa(\theta_j z_i, \theta_l z_i) \right]^{\frac{1}{2}}$$

$$= \left[\sum_{j=1}^{|\Theta|} \kappa(\theta_j z_i, \theta_j z_i) + \sum_{\theta_j \neq \mathrm{id}} \kappa(\theta_j z_i, z_i) \right]^{\frac{1}{2}} \tag{2.33}$$

$$\leq \sqrt{C_\kappa |\Theta|} \left[1 + \varsigma_{\kappa,\Theta} \left(|\Theta| - 1 \right) \right]^{\frac{1}{2}}. \tag{2.34}$$

As such, using only the one-sided McDiarmid's inequality, for every $t > 0$ we have with probability $\geq 1 - \exp\left\{ - \frac{n|\Theta|t^2}{2C_\kappa \left[1 + \varsigma_{\kappa,\Theta}(|\Theta|-1) \right]} \right\}$

$$d_{\mathcal{H}_\kappa}(\hat{\rho}_n, \rho) - \mathbb{E}[d_{\mathcal{H}_\kappa}(\hat{\rho}_n, \rho)] \leq t. \tag{2.35}$$

In order to upper bound the expectation, we use the symmetrization trick as follows

$$\mathbb{E}_{\{Z_i\}_1^n \sim \rho} \sup_{\|f\|_{\mathcal{H}_\kappa} \leq 1} \left| \frac{1}{n|\Theta|} \sum_{i=1}^{n} \sum_{j=1}^{|\Theta|} f(\theta_j z_i) - \mathbb{E}_\rho f \right|$$

$$= \mathbb{E}_Z \sup_{\|f\|_{\mathcal{H}_\kappa} \leq 1} \left| \frac{1}{n|\Theta|} \sum_{i=1}^{n} \sum_{j=1}^{|\Theta|} f(\theta_j z_i) - \mathbb{E}_{\{Z_i'\}_1^n \sim \rho} \left(\frac{1}{n|\Theta|} \sum_{i=1}^{n} \sum_{j=1}^{|\Theta|} f(\theta_j z_i') \right) \right|$$

$$\leq \mathbb{E}_{Z,Z'} \sup_{\|f\|_{\mathcal{H}_\kappa} \leq 1} \left| \frac{1}{n|\Theta|} \sum_{i=1}^{n} \sum_{j=1}^{|\Theta|} f(\theta_j z_i) - f(\theta_j z_i') \right|$$

$$= \mathbb{E}_{Z,Z',\xi} \sup_{\|f\|_{\mathcal{H}_\kappa} \leq 1} \left| \frac{1}{n|\Theta|} \sum_{i=1}^{n} \xi_i \sum_{j=1}^{|\Theta|} f(\theta_j z_i) - f(\theta_j z_i') \right| \tag{2.36}$$

$$\leq 2 \sqrt{\frac{C_\kappa \left[1 + \varsigma_{\kappa,\Theta} \left(|\Theta| - 1 \right) \right]}{n|\Theta|}}, \tag{2.37}$$

where $(\xi_1, \xi_2, \cdots, \xi_n) \sim$ i.i.d. standard Rademacher and (2.37) is due to Chen et al. (2023c), Lemma A.14. As such, 2.35 implies that

$$d_{\mathcal{H}_\kappa}(\hat{\rho}_n, \rho) \leq 2 \sqrt{\frac{C_\kappa \left[1 + \varsigma_{\kappa,\Theta} \left(|\Theta| - 1 \right) \right]}{n|\Theta|}} \left(1 + \sqrt{\frac{1}{2} \ln \frac{1}{\delta}} \right)$$

holds with probability $1 - \delta$ for $\delta > 0$. Observe that, the proof of this concentration bound acts as a generalization to the usual MMD, which arises in case $|\Theta| = 1$.

Going back to 2.29, we write

$$d_{\mathcal{H}_\kappa}\left(\phi_{\#}\hat{\mu}_n, \rho\right) - \sqrt{C_\kappa} \sup_{\mathcal{P}_n(\rho)} \Delta_{\Phi,n} \leq d_{\mathcal{H}_\kappa}\left(\phi_{\#}\hat{\mu}_n, \widehat{(\phi_{\#}\mu)}_n\right) + d_{\mathcal{H}_\kappa}\left(\hat{\rho}_n, \rho\right),$$

where the supremum is taken over possible estimators based on replicates from ρ of size n. It becomes evident that the quantity on the left has finite expectation and is essentially bounded in probability. Based on the concentration found in Theorem 4.5, along with the bound as in Example 1, we conclude that given $t > 0$

$$\begin{aligned} d_{\mathcal{H}_\kappa}\left(\phi_{\#}\hat{\mu}_n, \rho\right) - \sqrt{C_\kappa} \sup_{\mathcal{P}_n(\rho)} \Delta_{\Phi,n} &\leq \sqrt{t}(\sqrt{c_{g,n}U} + 2\sqrt{t}) + \mathcal{O}(\sqrt{c_{g,n}}(d^2 n)^{-\frac{1}{2d}}) \\ &+ \sqrt{\frac{2C_\kappa}{n}}\left[1 + \sqrt{\frac{1 + \varsigma_{\kappa,\Theta}\left(|\Theta| - 1\right)}{|\Theta|}}\right] + \mathcal{O}(\sqrt{dD_n} N_1^{-\frac{2}{d}} N_2^{-\frac{2}{d}}) \end{aligned} \tag{2.38}$$

holds with probability at least $1 - 2\exp\Big(-\frac{2nt^2}{\max\left\{B_x^2, 4C_\kappa\left[1+\varsigma_{\kappa,\Theta}(|\Theta|-1)\right]\right\}}\Big)$. Observe that, here $c_{g,n}$ is as described in Theorem 4.5 and typically behave as $\mathcal{O}(n^{-\frac{1}{2}})$ due to the strong law. N_1 and N_2 specify the width $W = \mathcal{O}(d\lfloor N_1^{\frac{1}{d}} \rfloor \vee N_1 + 1)$ and length $L = \mathcal{O}(N_2)$ of the encoder. Applying a change in variables on (2.38) we prove the theorem.

2.8.9 Proof of Theorem 2.5

Given an optimal encoder $E_n^*(t)$ that incurs latent loss $d_{\mathrm{TV}}\left(E_n^*(t)_{\#}\hat{\mu}_n, \rho\right) \leq t$, fragmenting the reconstruction error according to (2.11) yields

$$W_{c_x}^1(\mu, (D \circ E_n^*(t))_{\#}\hat{\mu}_n) \leq W_{c_x}^1((D \circ E_n^*(t))_{\#}\hat{\mu}_n, D_{\#}\rho) + W_{c_x}^1(D_{\#}\rho, \hat{\mu}_n) + W_{c_x}^1(\hat{\mu}_n, \mu).$$

The decoder transform is constructed as $D = \phi^{'} \circ D_0$, where $\phi^{'}$ is according to lemma 5.1 and $D_0 : \Omega_z \to \mathbb{R}$ is a linear map (or an ensemble of several). D can be equivalently written as $D = \phi^{'} \circ \sigma_I \circ D_0$, where σ_I is the identity activation, applied componentwise. As such, based on our definition, D indeed belongs to $\Phi(W, L)_k^d$ with depth ≥ 3. As discussed earlier, the resultant D thus turns out to be Lipschitz continuous. If c_x is taken as L_1, then observe

$$\begin{aligned} W_{c_x}^1\left((D \circ E_n^*(t))_{\#}\hat{\mu}_n, D_{\#}\rho\right) &\leq B_x d_{\mathrm{TV}}\left((D \circ E_n^*(t))_{\#}\hat{\mu}_n, D_{\#}\rho\right) \\ &= B_x \sup_{\omega \in \Sigma_{\mathcal{X}}} \left|E_n^*(t)_{\#}\hat{\mu}_n(D^{-1}(\omega)), \rho(D^{-1}(\omega))\right| \\ &\leq B_x \sup_{\omega^{'} \in \Sigma_{\mathcal{Z}}} \left|E_n^*(t)_{\#}\hat{\mu}_n(\omega^{'}), \rho(\omega^{'})\right| \\ &= B_x d_{\mathrm{TV}}\left(E_n^*(t)_{\#}\hat{\mu}_n, \rho\right) \leq t B_x, \end{aligned}$$

where the latter inequality is reached by taking supremum over all measurable sets belonging to the Borel sigma-algebra on $\mathcal{Z}$ instead of the particular path directed by D^{-1}. Also, the definition of D implies that given arvitrary $\varepsilon > 0$, $W^1_{c_x}(D_{\#}\rho, \hat{\mu}_n) < \varepsilon$ (lemma 5.1).

As such, the concentration of the reconstruction error is essentially determined by the statistical estimation error in the input space. Taking expectation over the samples, we write

$$\mathbb{E}\left[W^1_{c_x}(\mu, (D \circ E^*_n(t))_{\#}\hat{\mu}_n)\right] - tB_x \leq \mathcal{O}(n^{-\frac{1}{d}}),$$

whenever $d \geq 3$. Using the naive estimator, this is the sharpest rate one can achieve. However, to remove the influence of the dimensionality d, here, also Weed and Bach (2019)'s device can be applied (see Remark 8).

The bound, based on the empirical estimator, does not appreciate the smoothness of the input density p_μ. On the other hand, given $m > 0$, if $m_x > m$ we have $\mathcal{W}_R^{m_x,p} \subset \mathscr{B}^m_{pq}(\tilde{R})$, for some $\tilde{R} > 0$, where $1 \leq q \leq \infty$ and $1 \leq p < \infty$ i.e. belonging to the general Besov space (Giné and Nickl (2021), Section 4.3). Thus, if wavelet estimates (as in Weed and Berthet (2019)) are deployed instead, the rate can be expected to be $\mathcal{O}(n^{-\frac{1+m}{d+2m}})$ given $d \geq 3$.

2.8.10 Proof of Theorem 2.6

Let us consider the kernel $\kappa(x, y)$ to be of the form $\kappa(x-y)$, without loss of generality. Given replicates $X_1, \cdots, X_n \sim \tilde{p}$, the density estimate based on transformed (due to $G \in \mathcal{G}$) samples is given as

$$\hat{p}_h(x) = \frac{1}{nh^d}\sum_{i=1}^{n} \kappa\left(\frac{G(X_i) - x}{h}\right),$$

where h is the bandwidth. Decomposing the L_1 reconstruction error yields

$$|\hat{p}_h(x) - p_\mu(x)| \leq |\hat{p}_h(x) - \mathbb{E}[\hat{p}_h(x)]| + \left|\mathbb{E}[\hat{p}_h(x)] - p_\mu(G^{-1}(x))\right| + \left|p_\mu(x) - p_\mu(G^{-1}(x))\right|. \tag{2.39}$$

Now, observe that

$$\begin{aligned}
\mathbb{E}|\hat{p}_h(x) - \mathbb{E}[\hat{p}_h(x)]| &\leq \sqrt{\mathbb{E}\left(\hat{p}_h(x) - \mathbb{E}[\hat{p}_h(x)]\right)^2} \\
&= \sqrt{\frac{1}{n}\mathrm{Var}_{\tilde{p}}\left[\frac{1}{h^d}\kappa\left(\frac{G(X) - x}{h}\right)\right]} \\
&\overset{(1)}{=} \sqrt{\frac{1}{n}\mathrm{Var}_{\tilde{p}}\left[\frac{1}{h^d}\kappa\left(\frac{X - G^{-1}(x)}{h}\right)\right]} \\
&\overset{(2)}{\lesssim} \sqrt{\frac{1}{nh^d}},
\end{aligned}$$

where (1) is due to the transformation invariance of the kernel. The inequality (2) is obtained as a result of $\int \kappa^2(u)du < \infty$. The suppressed constant in the process is the upper bound to the density $\tilde{p}$. For the bias term, we write

$$\begin{aligned}
\left|\mathbb{E}[\hat{p}_h(x)] - p_\mu(G^{-1}(x))\right| &= \left|\frac{1}{h^d}\mathbb{E}_{\tilde{p}}\left[\kappa\left(\frac{G(X_i)-x}{h}\right)\right] - p_\mu(G^{-1}(x))\right| \\
&\leq \left|\frac{1}{h^d}\mathbb{E}_{\tilde{p}}\left[\kappa\left(\frac{X-G^{-1}(x)}{h}\right)\right] - \frac{1}{h^d}\mathbb{E}_{p_\mu}\left[\kappa\left(\frac{Y-G^{-1}(x)}{h}\right)\right]\right| \\
&\quad + \left|\frac{1}{h^d}\mathbb{E}_{p_\mu}\left[\kappa\left(\frac{Y-G^{-1}(x)}{h}\right)\right] - p_\mu(G^{-1}(x))\right| \\
&\overset{(3)}{\lesssim} \frac{1}{h^d}\mathbb{E}_{\tilde{p},p_\mu}\left|\frac{X-Y}{h}\right| + h^{m_x} \lesssim \frac{\epsilon}{h^{2d}} \vee h^{m_x},
\end{aligned}$$

where in (3) we use the invariance of κ for the first term, and the second bound is obtained using Giné and Nickl (2021), Proposition 4.3.33. We emphasize the fact that kernels satisfying Lipschitz continuity are commonly taken in practice, which are readily invariant to translations. The latter inequality is due to the contamination model. As such, assuming without loss of generality that the transform G preserves the origin, we get

$$\mathbb{E}|\hat{p}_h(0) - p_\mu(0)| \lesssim \sqrt{\frac{1}{nh^d}} \vee \frac{\epsilon}{h^{2d}} \vee h^{m_x}.$$

Taking $h = n^{-\frac{1}{d+2m_x}} \vee \epsilon^{\frac{1}{2d+m_x}}$, we prove the theorem.

Chapter 3

Translation and Cycle-Consistency of Cross-domain Generative Models

Summary

The task of unpaired image-to-image translation has witnessed a revolution with the introduction of the cycle-consistency loss to Generative Adversarial Networks (GANs). Numerous variants, with Cycle-Consistent Adversarial Network (CycleGAN) at their forefront, have shown remarkable empirical performance. The involvement of two unalike data spaces and the existence of multiple solution maps between them are some of the facets that make such architectures unique. In this chapter, we investigate the statistical properties of such unpaired data translator networks between distinct spaces, bearing the additional responsibility of cycle-consistency. In a density estimation setup, we derive sharp non-asymptotic bounds on the translation errors under suitably characterized models. This, in turn, extends the sufficient regularity conditions that maps must obey to carry out successful translations. We further show that cycle-consistency is achieved as a consequence of the data being generated with sufficient smoothness in each space based on observations from the other. In a first-of-its-kind attempt, we also provide deterministic bounds on the cumulative reconstruction error. In the process, we establish tolerable upper bounds on the discrepancy responsible for ill-posedness in such networks.

3.1 Introduction

The overwhelming number of variants GAN (Goodfellow et al., 2014) has inspired, while catering to its vast application domains, is a testament to its versatility. One such family of progenies, having remarkable accolades of its own, owes its genesis to the cycle-consistency constraint. Possibly the most influential one belonging to this group is CycleGAN (Zhu et al., 2017). It offers an unsupervised image-to-image (I2I) translation framework for unpaired observations hailing from unrelated data spaces. In terms of the architecture, both DualGAN

(Yi et al., 2017) and DiscoGAN (Kim et al., 2017) are immediate relatives to CycleGAN. In the chassis of such networks lie two concurrent adversarial generation processes, commonly termed *translations*, regularized by a cyclic loss. This penalization ensures the reconstruction of input data from either space post-translation. In addition, models such as DTN (Taigman et al., 2017) and UNIT (Liu et al., 2017a) assume the existence of a shared latent space between the domains. This allows the restructuring of the model without altering the objective. By stacking multiple translator networks, SCAN (Li et al., 2018) promises significant performance improvement, especially for high-resolution images. Some members of the family ((Zhu et al., 2017), U-GAT-IT (Kim et al., 2020), Moriakov et al. (2020)) also deploy an additional identity loss to remove tilt-shift in generated images. We call this broad class of I2I translation machines 'cycle-consistent networks'. The constraint of cycle-consistency should be primarily credited for the masterly generative capability of such models, from which tasks like style transfer, object transfiguration (Zhu et al., 2017), and data augmentation (Sandfort et al., 2019) benefit immensely.

In this chapter, we intend to rise above the empirical evidence by providing a statistical backbone to the fact that cycle-consistent networks can simultaneously translate data both ways without losing the capacity to reconstruct. We call the two maps, operating in opposite directions, *Translators*. Both underlying distributions portraying purposeful image data, in the absence of conventional latent laws, call for such transformations to differ from usual generators used in vanilla GANs. Unlike GANs, there may exist non-unique solution maps bringing about 'zero' realized loss in this case (Moriakov et al., 2020). As such, searching for translators that minimize the error is not sufficient. This fact motivates us to study the desirable regularities of the maps that facilitate the statistical convergence of output measures, which in turn define our notion of *consistency*. Theoretically, the concept of 'cycle-consistency' is analogous to 'regeneration' (Chapter 2) in case of Variational Autoencoders (Kingma and Welling, 2014b). In such inverse problems, maps reconstructing the input signal become sensitive to slight perturbations due to noise. A noisy output from earlier translations contributes to this ambiguity in the inverse generation process, formally known as 'ill-posedness' (Sim et al., 2020). Theoretical insights regarding the source and admissible error margins of ill-posedness remain absent to date. Confronted with such challenges, this chapter provides a fresh perspective on the theoretical machinery of the cycle-consistent adversarial networks. Our contributions can be summarized in the following way:

- We show that translators, based on deep ReLU networks, prevent the information provided by input empirical laws from dissipating during generation cycles. In Theorem 3.1 and Corollary 3.1, we prove that the same translators not only achieve zero generation loss asymptotically, but the generated sequence of distributions also converges to the target density almost surely.

- Under Sobolev-smooth input laws, we establish that the uses of L_1 norm and 1-Wasserstein

distance in the cyclic loss are equivalent, attesting to Zhu *et al.*'s Zhu et al. (2017) observation that the latter does not improve performance [Theorem 3.3].

- Furthermore, we prove that a network deploying the aforementioned translators may achieve cycle-consistency as a consequence of translation consistency in both directions [Theorem 3.3, (3.4)], a fact that does not hold in general.

3.2 Background

Playing catch-up to earlier empirical success, theoretical scrutiny of GANs fostered a series of notable works in recent years. Liu et al. (2017b) characterized the objective functions of several GAN architectures (f-GAN (Nowozin et al., 2016), WGAN (Arjovsky et al., 2017), etc.) as *adversarial divergences*. This allowed them to analyze the convergence of generated distributions towards the target law in a unified framework. Meanwhile, Arora et al. (2017) explored the expressiveness of generator networks and the generalization performance of GANs under the *neural net distance*. In a later work, however, we observed Zhang et al. (2018) show improved results over both. Convergence and related asymptotic properties of the density estimates in a GAN setup can also be found in the parametric approach of Biau et al. (2020). On the other hand, error decomposition of the GAN-objective under both parametric and non-parametric regimes may lead to non-asymptotic concentration bounds. Several works followed this approach with various smoothness assumptions on the data distributions and the transformations involved (Chen et al., 2020; Huang et al., 2021; Liang, 2021). A more recent study of the same nature also focused on learning from low-dimensional latent laws using smooth maps (Schreuder et al., 2021). One may also come across several GAN variants inspiring similar pursuits. Biau et al. (2021) presented a comprehensive study of the convergence and related asymptotic properties of the parametric density estimates in a WGAN setup. From a non-parametric viewpoint, Haas and Richter (2020) derived deterministic upper bounds on the expected WGAN loss, under both conditional and unconditional generation processes. Lately, a non-asymptotic approach of a similar spirit has been utilized to establish risk bounds on the realized Bidirectional-GAN (BiGAN) error (Liu et al., 2021b).

Cycle-consistent networks, despite marking a triumph in deep generative modeling, have not received such independent attention yet. This scarcity makes the existing attempts even more meaningful. Moriakov et al. (2020) proved that multiple solutions to the CycleGAN problem exist, as a consequence of the existence of nontrivial automorphisms in either data space. Tiao et al. (2018) pointed out that the cycle-consistency loss boils down to an expected posterior log-likelihood in a Bayesian setup. On the contrary, the CycleGAN objective can also be recognized as the Unbalanced Gromov-Monge Divergence (UGMD), when the transformations are assumed to be isometric (Zhang et al., 2022). However, all the above studies refrain from exploring the statistical guarantees a cycle-consistent translator aims to provide

by following its concurrent objectives. Our work is an original attempt to fill this gap.

3.3 Preliminaries

3.3.1 Notations

We follow notational conventions from the previous chapter. In this section, we reiterate some for ease of reading. The two data spaces involved $\mathcal{X}$ and $\mathcal{Y}$, equipped with respective distances c and $c^{'}$, are considered to be Polish (i.e., separable and completely metrizable). A simple characterization of the same might be $\mathbb{R}^d$ for $d \geq 1$. Let $\mathcal{P}(\mathcal{X})$ denote the space of probability measures defined on $\mathcal{X}$. We refer to the set of measurable functions mapping $\mathcal{X}$ to $\mathcal{Y}$ as $\mathscr{F}(\mathcal{X}, \mathcal{Y})$. The 'forward' and 'backward' translator maps between the spaces are denoted by F and G, respectively. Observe that a probabilistic forward translator belongs to $\mathscr{F}(\mathcal{X}, \mathcal{P}(\mathcal{Y}))$, as the translated observations conditioned on the inputs follow a probability distribution. Similarly, $G \in \mathscr{F}(\mathcal{Y}, \mathcal{P}(\mathcal{X}))$. The two discriminator networks at both ends induce functions D_X and D_Y, which play the role of critics in the two simultaneous adversarial games. Let us now revisit some concepts that we frequently utilize in the upcoming discussion.

Definition 3.1 (Wasserstein Distance)**.** *For a metric $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ and measures $P, Q \in \mathcal{P}(\mathcal{X})$, the r^{th} Wasserstein Distance between P and Q is defined as*

$$W_c^r(P, Q) = \inf_{\gamma \in \Gamma(P,Q)} \Big\{ \int_{\mathcal{X} \times \mathcal{X}} [c(x, y)]^r d\gamma(x, y) \Big\}^{\frac{1}{r}},$$

where $\Gamma(P, Q) = \big\{\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) : \int_{\mathcal{X}} \gamma(x, y)dy = P,\ \int_{\mathcal{X}} \gamma(x, y)dx = Q\big\}$ is the set of all measure couples between P and Q; $r \in [1, \infty)$.

Remark 3.1. *In our analysis, we make extensive use of a particular case of this discrepancy measure, namely when $r = 1$. We also reiterate the fact that W_c^1 can be written as $W_c^1(P, Q) = \sup_{l \in \mathscr{L}_c^1} \big\{\int_{\mathcal{X}} l(x)dP(x) - \int_{\mathcal{X}} l(x)dQ(x)\big\}$, where $\mathscr{L}_c^1 :=$ class of 1-Lipschitz functions with respect to c [Remark 6.5 in Villani (2009)]. However, we adopt the notation $d_{\mathscr{L}_c^1}$ instead to maintain consistency with the other Integral Probability Metrics (IPMs).*

3.3.2 Problem Setup

Throughout our discussion, we denote the distributions at both ends by $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ respectively. The adversarial loss that the backward generation process ($\mu \xleftarrow{G} \nu$) tries to minimize is given by,

$$\mathcal{L}_{D_X}(\mu, \nu, G) = \mathbb{E}_{x \sim \mu}[D_X(x)] - \mathbb{E}_{y \sim \nu}[D_X(G(y))].$$

The same convention leads to the forward generation ($\mu \xrightarrow{F} \nu$) loss, $\mathcal{L}_{D_Y}(\nu, \mu, F)$. The string tying these two processes together comes in the form of the cyclic loss. Based on our notations, it can be written as

$$\mathcal{L}_{cyc}(\mu, \nu, F, G) = \mathbb{E}_{x\sim\mu}\Big[\|x - G(F(x))\|_1\Big] + \mathbb{E}_{y\sim\nu}\Big[\|y - F(G(y))\|_1\Big],$$

where $\|\cdot\|_1$ represents the L^1 norm. We point out that this specific choice of the norm is based on the recommendation of Zhu et al. (2017). According to them, the usage of an adversarial loss instead does not improve the regenerated image quality. Through the following illustration (Figure 3.1), we offer the reader a glimpse of our idea of concurrent translations and reconstructions.

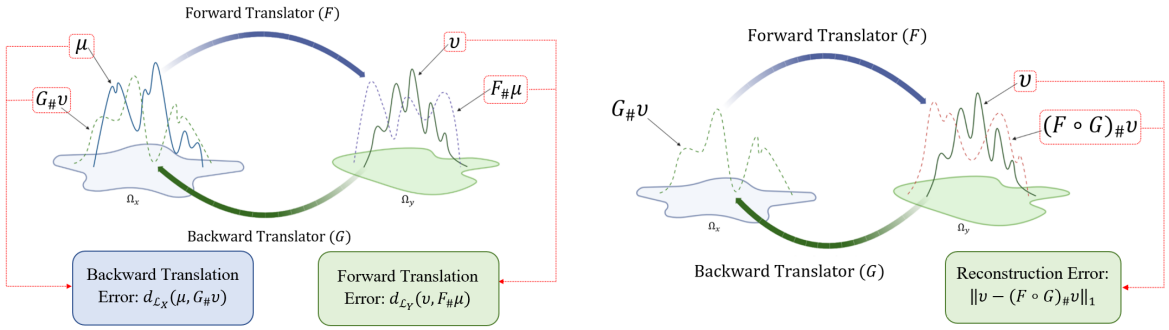


Figure 3.1: (a) Forward and backward translations with corresponding errors, (b) Reconstruction in the space $\mathcal{Y}$, all viewed through the glass of density estimation.

A typical CycleGAN (Zhu et al., 2017), or equivalently DiscoGAN (Kim et al., 2017) formulation, carries out the following optimization task:

$$\inf_{\substack{F\in\mathscr{F}(\mathcal{X},\mathcal{P}(\mathcal{Y}))\\ G\in\mathscr{F}(\mathcal{Y},\mathcal{P}(\mathcal{X}))}} \sup_{\substack{D_X\in\mathscr{L}_X\\ D_Y\in\mathscr{L}_Y}} \Big\{\mathcal{L}_{cyc}(\mu, \nu, F, G) + \lambda_1\mathcal{L}_{D_X}(\mu, \nu, G) + \lambda_2\mathcal{L}_{D_Y}(\nu, \mu, F)\Big\}, \tag{3.1}$$

where $\mathscr{L}_X$ and $\mathscr{L}_Y$ are classes of discriminator functions and the maps F, G are sculpted using translator networks. Also, $\lambda_1, \lambda_2 > 0$. Observe that, (3.1) can be rewritten as

$$\begin{aligned}
&\inf_{\substack{F\in\mathscr{F}(\mathcal{X},\mathcal{P}(\mathcal{Y}))\\ G\in\mathscr{F}(\mathcal{Y},\mathcal{P}(\mathcal{X}))}} \Big\{\mathcal{L}_{cyc}(\mu, \nu, F, G) + \lambda_1 \sup_{D_X\in\mathscr{L}_X} \mathcal{L}_{D_X}(\mu, \nu, G) + \lambda_2 \sup_{D_Y\in\mathscr{L}_Y} \mathcal{L}_{D_Y}(\nu, \mu, F)\Big\}\\
\equiv &\inf_{\substack{F\in\mathscr{F}(\mathcal{X},\mathcal{P}(\mathcal{Y}))\\ G\in\mathscr{F}(\mathcal{Y},\mathcal{P}(\mathcal{X}))}} \Big\{\mathcal{L}_{cyc}(\mu, \nu, F, G) + \lambda_1 d_{\mathscr{L}_X}(\mu, G_{\#}\nu) + \lambda_2 d_{\mathscr{L}_Y}(\nu, F_{\#}\mu)\Big\},
\end{aligned} \tag{3.2}$$

given that $d_{\mathcal{F}}(P, Q) = \sup_{f\in\mathcal{F}}\{\mathbb{E}_P[f] - \mathbb{E}_Q[f]\}$. The only feature that differentiates the DualGAN (Yi et al., 2017) objective from (3.2) is the employment of the conventional generation

technique based on noise. Furthermore, in an attempt to preserve the colour composition in images, models such as the extended CycleGAN (Moriakov et al., 2020; Zhu et al., 2017), U-GAT-IT (Kim et al., 2020) impose a constraint that penalizes the translators' tendency to move away from the identity. It is given by,

$$\mathcal{L}_{id}(\mu, \nu, F, G) = \mathbb{E}_{x\sim\mu}\Big[\|x - F(x)\|_1\Big] + \mathbb{E}_{y\sim\nu}\Big[\|y - G(y)\|_1\Big].$$

Observe that such regularization is feasible only when the two data distributions are equi-dimensional. The map F is built with the fundamental motivation of transforming μ into ν. As such, the discrepancy $\|\mu - F_{\#}\mu\|_1$ should not be minimized beyond the difference between μ and ν. We provide a detailed discussion on the same in the Appendix.

3.4 Theoretical Analysis

3.4.1 Data Distributions

Depicting real images as observations from probability distributions is a completely theoretical construct. The representation provides practitioners with a refined view of the problem. Moreover, the transformed objective of density estimation has its benefits. Perhaps this is the idea that inspired the genesis of 'Roundtrip', a CycleGAN progeny (Liu et al., 2021a). In our study, we consider $\mathcal{X} \equiv \mathbb{R}^d$ and $\mathcal{Y} \equiv \mathbb{R}^k$; $d, k \in \mathbb{N}^+$. The two dimensions need not be equal in general. The consequences of the special case of equality will be discussed at a later stage. Let us now characterize the data distributions under consideration.

Particularly, we consider μ and ν to have corresponding densities p_μ and p_ν, with respect to Lebesgue measures, in their respective spaces. The following assumption provides coherence to their characterization.

Assumption 3.1. *(Regularity of distributions) There exists $m_x, m_y \in \mathbb{N}^+$ such that $p_\mu \in \mathcal{W}_L^{m_x,p}(\Omega_x)$ and $p_\nu \in \mathcal{W}_L^{m_y,q}(\Omega_y)$, where the supports $\Omega_x \subseteq \mathbb{R}^d$ and $\Omega_y \subseteq \mathbb{R}^k$ are both compact, $p, q \in [1, \infty)$.*

Feature-extracted image data, in vectorized form, tends to hail from bounded domains in each of its coordinates. Our characterization of input laws having compact support complements this fact. Perhaps it is the very reason that motivates Chen et al. (2020) and Liang (2021) to assume the same, contextually.

3.4.2 Class of Discriminator Functions

Functional classes $\mathscr{L}_X, \mathscr{L}_Y$ are characterized based on their ability to tell apart real and generated observations. Some of the notable choices of the same include functions defined over a Reproducing Kernel Hilbert Space (RKHS) (Dziugaite et al., 2015), Lipschitz (Arjovsky

et al., 2017), and Sobolev functions (Mroueh et al., 2018). In our work, we concentrate on two families, namely $\mathscr{L}_c^1$ and $\mathcal{W}^{m,\infty}$. Our first choice is motivated by the heightened generative quality that the Wasserstein distance brings along to deep models. It also offers a pathway to fend off mode collapse and vanishing gradients. On the other hand, the latter class of critics enables us to study the effect of improved smoothness on translation and regeneration.

So far, we have only discussed the Lagrangian formulation of the optimization problem at hand. In fact, (3.2) is the embodiment of the exact *Lagrange dual function*. The forthcoming analysis, however, relies on the 'constrained version' [Chapter 5 of Boyd and Vandenberghe (2004)] given as follows:

$$\inf_{\substack{F\in\mathscr{F}(\mathcal{X},\mathcal{P}(\mathcal{Y}))\\ G\in\mathscr{F}(\mathcal{Y},\mathcal{P}(\mathcal{X}))}} \Big\{\mathcal{L}_{cyc}(\mu,\nu,F,G)\Big\} \text{ subject to } d_{\mathscr{L}_X}(\mu, G_{\#}\nu)\le t_1 \text{ and } d_{\mathscr{L}_Y}(\nu, F_{\#}\mu)\le t_2, \quad (3.3)$$

where $t_1, t_2 \ge 0$. Solutions from (3.2) turn out to be lower bounds to those derived from (3.3), a fact that inspires the forthcoming theory. We say 'simultaneous successful translations have taken place' only when the constraints in (3.3) are met. The immediate inquiry that follows involves checking the feasibility of an architecture to achieve cycle-consistency. As supporting evidence for both phenomena, we produce deterministic upper bounds on the respective errors along with convergence guarantees of distributions.

3.4.3 Translation Guarantees

Let us concentrate on the backward translation ($\mu \xleftarrow{G} \nu$) first. Observe that, a realized sample counterpart of the objective turns out to be $d_{\mathscr{L}_X}(\hat{\mu}_{n_1}, G_{\#}\hat{\nu}_{n_2})$, where $\hat{\mu}_{n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1}\delta_{X_i}$ is the empirical distribution corresponding to μ, based on $n_1 \in \mathbb{N}^+$ i.i.d. samples $\{X_i\}_{i=1}^{n_1}$. Similarly, $\hat{\nu}_{n_2}$ stands for the same in case of ν, given $n_2 \in \mathbb{N}^+$ samples. As a consequence, any backward translator $G \in \mathscr{F}(\mathcal{Y}, \mathcal{P}(\mathcal{X}))$ should be recognized as $G(n_1, n_2)$. Non-uniqueness of the members residing in the kernel of CycleGAN loss is a well-known fact (Moriakov et al., 2020). Our goal is to prescribe real architectures that induce maps satisfying the first constraint in the sample version of (3.3). Such recommendations rely on the next definition.

Definition 3.2 (ReLU Neural Network)**.** *Given $L \in \mathbb{N}^+$, a L-deep Neural Network (NN) is defined as the collection of maps $\phi : \mathbb{R}^{N_0} \longrightarrow \mathbb{R}^{N_{L+1}}$, $\{N_i\}_{i=0}^{L+1} \in \mathbb{N}^+$ given by*

$$\phi(x) := A_L \circ \sigma \circ A_{L-1} \circ ... \circ \sigma \circ A_0(x),$$

where $A_i(y) = M_i y + b_i$; $M_i \in \mathbb{R}^{N_{i+1}\times N_i}$ and $b_i \in \mathbb{R}^{N_{i+1}}$, $i = 0, ..., L$ is an affinity. The activation $\sigma(y) = y \vee 0$, $y \in \mathbb{R}$. Under this setup, we call $W = \vee_{i=1}^{L} N_i$ the width of the network. Denote this collection by $\Phi(W, L)_{N_0}^{N_{L+1}}$.

It is fair to say that ReLU is the most commonly used activation function in modern deep

NNs. It is simple to use and also speeds up training, much to practitioners' delight. Moreover, it is superior at dealing with vanishing gradients compared to *sigmoid* or *tanh*. However, we draw inspiration from the remarkable approximation capability that ReLU-based networks offer (Yarotsky, 2017, 2018), especially towards smooth functions (Petersen and Voigtlaender, 2018).

Theorem 3.1. *There exist backward translators ϕ, based on ReLU neural network $\Phi(W, L)_k^d$ with width $W \geq 7d+1$ and depth $L \geq 3$, such that whenever $n_1 \leq \frac{W-d-1}{2}\lfloor\frac{W-d-1}{6d}\rfloor\lfloor\frac{L}{2}\rfloor + 2$, we have*

$$\mathbb{E}\big[d_{\mathscr{L}_c^1}(\hat{\mu}_{n_1}, \phi_{\#}\hat{\nu}_{n_2})\big] \precsim (k^2 n_2)^{-\frac{1}{k}} + \sqrt{k} W^{-\frac{2}{k}} L^{-\frac{2}{k}}.$$

This result enables us to formally present what we mean by 'translation guarantee'. The next corollary can be seen as an embodiment of the same idea.

Corollary 3.1 (Translation consistency)**.** *As* $\min(n_1, n_2) \to \infty$*, we have* $d_{\mathscr{L}_c^1}(\hat{\mu}_{n_1}, \phi_{\#}\hat{\nu}_{n_2}) \xrightarrow{a.s.} 0$.

In other words, given sufficient information from both the distributions, the backward translation method governed by a map ϕ satisfies the constraint in the sample version of (3.3). The corollary is an asymptotic statement that ensures the error eventually shrinks below any given $t_1 > 0$. A crucial observation in this context is that for $m \geq 1$, $\mathcal{W}_1^{m,\infty}$ is a sub-family of bounded Lipschitz functions. In our case, since the supports of the distributions are taken to be bounded, one may equivalently say $\mathcal{W}_1^{m,\infty} \subset \mathscr{L}_c^1$, $c \equiv L_1$. As such, $\mathcal{W}_1^{m,\infty}$ playing the role of the critic should produce results similar to Theorem 3.1.

Theorem 3.2. *For a backward translator ϕ of width W and depth L, as specified in Theorem 3.1*

$$\mathbb{E}\big[d_{\mathcal{W}_1^{m,\infty}}(\hat{\mu}_{n_1}, \phi_{\#}\hat{\nu}_{n_2})\big] \precsim {n_2}^{-\frac{m}{k}} + \frac{\log n_2}{\sqrt{n_2}} + \sqrt{k} W^{-\frac{2}{k}} L^{-\frac{2}{k}},$$

where $n_1 \leq \frac{W-d-1}{2}\lfloor\frac{W-d-1}{6d}\rfloor\lfloor\frac{L}{2}\rfloor + 2$ and $n_2 \in \mathbb{N}^+$.

One might wonder what makes Lipschitz transformations so relevant to this context. The first rather evident observation is that it restricts any further amplification of the distance between laws post-translation. The next reason, a particular consequence of the former, brings us back to the concept of *Information preserving transformations* (IPT).

Here, $\widehat{(I_{\#}\nu)}_n$ is an empirical counterpart of the translated law $I_{\#}\nu$ based on $n \in \mathbb{N}^+$ samples. As such, IPTs are maps that ensure the error committed while replacing $I_{\#}\hat{\nu}_n$ with $\widehat{(I_{\#}\nu)}_n$ (information dissipated) remains arbitrarily small, with a high probability.

Remark 3.2. *Recalling Chapter 2, maps induced by ReLU feed-forward networks can similarly pose as IPT, incurring an additional approximation error of order $\mathcal{O}(W^{-\frac{2}{k}} L^{-\frac{2}{k}})$ [Lemma 3.4]. This near-perfect behaviour makes the choice of ReLU-NNs, as suitable translators,*

rather inevitable. However, neural networks based on tanh (De Ryck et al., 2021), sigmoid (Langer, 2021), and GroupSort (Tanielian and Biau, 2021) activations have also been shown to approximate Lipschitz functions with high precision. We feel, a comparative analysis of the activations based on their effectiveness in the face of information dissipation may lead to improved prescriptions.

Remark 3.3 (Forward translation)**.** *We stress the fact that so far in our discussion, the data dimensions d, k have no restrictions put on them. As such, the same arguments hold true for the forward generation process ($\mu \xrightarrow{F} \nu$) as well. A forward translator map $\psi \in \Phi(W, L)_d^k$ can be similarly constructed that achieves translation consistency. In other words, one may easily check that the second constraint in (3.3) is also satisfied for arbitrary values of t_2.*

Cycle-consistent networks find themselves under the obligation to reconstruct the input signal following their translation. A major obstacle in the process, however, that often mars the quality of regenerated observations is ill-posedness. It stems from a translation belonging to the feasible set of solutions that results in noisy output, devoid of sufficient information to aid the reconstruction. Theoretically, the remedy to ill-posedness lies in the formation of a 'perfect' transport map between the measures. Having IPT (Lipschitz) as a reference, NN-based transports tend to overcome this issue asymptotically [Corollary 3.1]. However, measurable maps, in general, lack such approximation capability. Our following discussion sheds light on the same.

For this section, let us assume the dimensions of the two data domains to be equal, i.e., $d = k$. This occasion, in particular, has interesting consequences. Given that p_μ and p_ν have finite variance, Brenier's theorem (Brenier, 1991) ensures the existence of a unique solution $\gamma = (Id \times T)_\# \nu$ to the Kantotovich Optimal Transport (OT) problem: $W^1_{c \equiv L_2}$. In other words, we get hold of a map $T \in \mathscr{F}(\mathcal{Y}, \mathcal{P}(\mathcal{X}))$ such that $T_\# \nu = \mu$. It is clear that any function aspiring to subdue ill-posedness should lie in an ϵ-envelope of T, $\epsilon > 0$ being as small as possible. The larger the deviation, the greater is the extent of degradation in reconstructed image quality. Drawing inspiration from this fact, Lu et al. (2019); Sim et al. (2020) deploy OT-based regularizers to guide the solution map toward T. However, the regularity properties of T can only be determined under very specific assumptions on the data domains (Caffarelli, 1992; Colombo and Fathi, 2021). Moreover, we only have access to an empirical counterpart of the target law in the sample version of the problem. As a result, approximations of the transport map are likely to be noisy. Our next result aims at pointing out the tolerable error margin due to ill-posedness in a sample backward translation.

Lemma 3.1. *For a discriminator class $\mathscr{L}_X$, and a backward translator G*

$$d_{\mathscr{L}_X}(\hat{\mu}_{n_1}, G_\# \hat{\nu}_{n_2}) \leq \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3,$$

where $\frac{\mathcal{E}_1}{B_x} := \|\hat{\mu}_{n_1} - T_\# \nu\|_{TV}$ (Statistical approximation error in target space),

$\mathcal{E}_2 := B_x \left\| \Gamma_{n_1} - \widehat{(G_{\#}\nu)}_{n_2} \right\|_{TV} = \Lambda_{(n_1,n_2)}$; *given* $\Gamma_{n_1} = \text{argmin}_{\tau \in \mathcal{P}(\mathcal{X})} \|\tau - \hat{\mu}_{n_1}\|_{TV}$, $B_x = diam(\Omega_x)$ *with respect to the metric* c, *and* $\mathcal{E}_3 := d_{\mathscr{L}_X}\left(\widehat{(G_{\#}\nu)}_{n_2}, G_{\#}\hat{\nu}_{n_2}\right)$ *(Information dissipated).*

The quantity $\Lambda_{(n_1,n_2)}$ represents the cost incurred by $\widehat{(G_{\#}\nu)}_{n_2}$ for partaking in the Scheffe tournament (Devroye and Lugosi, 2001) to approach $\hat{\mu}_{n_1}$. Observe that it remains an admissible amount of deviation if $\lim_{\min(n_1,n_2)\to\infty} \Lambda_{(n_1,n_2)} \leq t_1$ (3.3). The maps, ensuring $\lim_{\min(n_1,n_2)\to\infty} \Lambda_{(n_1,n_2)} = 0$, belong to the set of 'pure' solutions (Moriakov et al., 2020). Theoretically, the negative effect of such maps on the regeneration quality would be benign. We elaborate on the same in Proposition 3.1. Also, observe that Lemma 3.1 re-emphasizes the necessity of a backward translator to be an IPT.

Remark 3.4 (Mode collapse)**.** *In real situations, supports of data distributions at both ends are often non-convex. This is an important feature that makes OT maps (T) discontinuous (Lei et al., 2019). On the other hand, neural networks lack proficiency in approximating such discontinuous functions. For multi-modal input laws, an estimated transformation approximating only the continuous branches of the target OT map results in mode collapse during translation (Lei et al., 2019). As such, the error associated with mode collapse remains convoluted in* $\mathcal{E}_2$*. Fragmenting the realized estimation loss into finer components to address mode collapse may be taken up as potential future work.*

Let us now shift our focus towards the residual task a cycle-consistent I2I translator needs to execute.

3.4.4 Cycle Consistency Analysis

The cyclic loss, as given in (3.2), measures the expected discrepancy between the input data and its reconstructed counterpart. However, the density estimation approach we follow allows us to reframe the objective as a divergence between distributions, given as

$$\mathcal{L}_{cyc}(\mu, \nu, F, G) = \|\mu - (G \circ F)_{\#}\mu\|_1 + \|\nu - (F \circ G)_{\#}\nu\|_1.$$

For two distributions P, Q; $\|P - Q\|_1 = 2\|P - Q\|_{TV} = \int |\rho_P - \rho_Q| d\lambda$, given that $\frac{dP}{d\lambda} = \rho_P$ and $\frac{dQ}{d\lambda} = \rho_Q$. This formulation provides a stronger notion of the loss. The first result of this section discovers the relationship between translation and reconstruction.

Lemma 3.2. *For* $G \in \mathscr{F}(\mathcal{Y}, \mathcal{P}(\mathcal{X}))$ *and* $F \in \mathscr{F}(\mathcal{X}, \mathcal{P}(\mathcal{Y}))$,

$$\mathcal{L}_{cyc}(\mu, \nu, F, G) \leq 2\Big\{ \|\mu - G_{\#}\nu\|_1 + \|\nu - F_{\#}\mu\|_1 \Big\}.$$

The extent to which a cycle-consistent translator can be inaccurate is determined by its performance in the simultaneous generations. This is a rather desired outcome. However,

the indication of much intrigue that this result gives is that in case the translations are 'successful' in both directions, cycle-consistency can be achieved. One key feature of the input distributions that becomes crucial henceforth is their smoothness. As a consequence of Corollary 3.1, we infer $\phi_{\#}\hat{\nu}_{n_2} \to \mu$ weakly. A similarly constructed forward translator ψ may also ensure that $\psi_{\#}\hat{\mu}_{n_1} \to \nu$ weakly. Based on such guarantees, we make the following assumptions about the regularity of the transported laws.

Assumption 3.2. *The translated distributions $\phi_{\#}\nu$ and $\psi_{\#}\mu$ possess corresponding densities given by $p_{\phi_{\#}\nu} \in \mathcal{W}_L^{m_x,p'}(\Omega_x)$ and $p_{\psi_{\#}\mu} \in \mathcal{W}_L^{m_y,q'}(\Omega_y)$; $p', q' \in [1,\infty)$.*

Before presenting upcoming theoretical results, let us recall the Regularly Invariant kernels (Definition 2.8), which play a key role henceforth. Here, we assume the invariance property to hold under the absolute difference between the arguments, i.e., $\mathcal{O}(|v-u|)$.

We mention that the total variation metric can also be expressed as a transportation distance, the underlying cost function being $c(x,y) = 1_{x \neq y}$. However, as Chae and Walker (2020) points out, the topologies that the TV and Wasserstein distances generate are hardly comparable. For Sobolev densities, TV often fails to appreciate the nuances that 'smoothness' brings along. A method to alleviate such difficulties lies in regular kernels. Minute deviations between smooth functions can be apprehended in greater detail when convoluted with such kernels. Inevitably, regularly invariant kernels become the cornerstone of our next result. The proof, placed in the Appendix, highlights its contribution.

Theorem 3.3. *Given the metric $c \equiv L_1$, there exists a constant $M > 0$ dependent on m_x, such that*

$$\left\|p_\mu - p_{\phi_{\#}\nu}\right\|_1 \leq M\left[\left\|D^{m_x}p_\mu\right\|_p + \left\|D^{m_x}p_{\phi_{\#}\nu}\right\|_{p'}\right]^{\frac{1}{m_x+1}}\left[d_{\mathscr{L}_c^1}(\mu,\phi_{\#}\nu)\right]^{\frac{m_x}{m_x+1}}.$$

Remark 3.5. *This result is a multivariate generalization of Theorem 2.1 in Chae and Walker (2020).*

Note that a similar conclusion can also be drawn for the loss, indicating the difference between the target and generated density in case of forward translation. That is to say,

$$\left\|p_\nu - p_{\psi_{\#}\mu}\right\|_1 \leq M'\left[\left\|D^{m_y}p_\nu\right\|_q + \left\|D^{m_y}p_{\psi_{\#}\mu}\right\|_{q'}\right]^{\frac{1}{m_y+1}}\left[d_{\mathscr{L}_c^1}(\nu,\psi_{\#}\mu)\right]^{\frac{m_y}{m_y+1}}, \tag{3.4}$$

where M' is a constant depending on m_y. Likewise, any pair of translators (G,F) that preserve the smoothness of input densities onto generated ones satisfy Theorem 3.3 and (3.4). The collective evidence from these two results suggests that a sufficient condition for achieving cycle-consistency is the arbitrary closeness between real and translated Sobolev-smooth densities, in both domains, under the 1-Wasserstein metric. Moreover, we already know

that $\frac{2}{B_x} d_{\mathscr{L}_c^1}(\mu, \phi_{\#}\nu) \leq \|\mu - \phi_{\#}\nu\|_1$ (Gibbs and Su, 2002). As such, establishing translation consistency under the critic $\mathscr{L}_c^1$ is equivalent to attaining cycle-consistency.

Remark 3.6. *The observation, however, does not hold in general. While empirically the deployment of cycle-consistency loss seems necessary, without the smoothness assumptions on the translated laws, one may construct counterexamples. For example, given* $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, *consider the distributions* $\mu = \nu \overset{d}{=} \mathcal{U}[0,1]$. *Observe that taking both* F *and* G *as identities satisfy* $F_{\#}\mu = \nu$ *and* $G_{\#}\nu = \mu$ *a.e. However, as Chakraborty and Bartlett (2025) suggests, for* $F^n(x) = \frac{1}{n}\sum_{i=0}^{n-1} i\mathbb{1}\left(x \in \left(\frac{i}{n}, \frac{i+1}{n}\right]\right)$ *and* $G^n(x) = x - x\sum_{j=0}^{n-1} \mathbb{1}(x = j/n)$, *though* $F^n_{\#}\mu \xrightarrow{d} \nu$ *and* $G^n_{\#}\nu \xrightarrow{d} \mu$ *hold, we have* $(G^n \circ F^n)_{\#}\mu \overset{d}{\not\to} \mu$.

Now, let us focus on the sample version of the cyclic loss, given as $\mathcal{L}_{cyc}(\hat{\mu}_{n_1}, \hat{\nu}_{n_2}, F, G)$. The inability of translation maps to approximate optimal transports up to arbitrarily high accuracy affects cycle-consistency as well. Noisy outputs from a backward generation process should not ideally recover, even under a 'perfect' forward translator. Meanwhile, a 'perfectly' translated forward image will be distorted due to such imperfect backward generators. If the effects due to the departure of translator maps from their 'ideal' benchmarks get multiplied, we may observe severe corruption in reconstruction quality. Much to our relief, the next result assures that the effects of ill-posedness amplify only as a sum.

Proposition 3.1. *Denote* $B_y \left\| \Gamma'_{n_2} - \widehat{(F_{\#}\mu)}_{n_1} \right\|_{TV} = \Lambda'_{(n_1,n_2)}$, *given that* $\Gamma'_{n_2} = \operatorname{argmin}_{\tau \in \mathcal{P}(\mathcal{Y})} \|\tau - \hat{\nu}_{n_2}\|_{TV}$, $B_y = diam(\Omega_y)$ *with respect to the metric* c'. *Then*

$$\mathcal{L}_{cyc}(\hat{\mu}_{n_1}, \hat{\nu}_{n_2}, F, G) - 4\left\{\frac{\Lambda_{(n_1,n_2)}}{B_x} + \frac{\Lambda'_{(n_1,n_2)}}{B_y}\right\} \leq \mathcal{E}_1^* + \mathcal{E}_2^*,$$

where $\mathcal{E}_1^* := 4\left\{\|\hat{\mu}_{n_1} - \mu\|_{TV} + \|\hat{\nu}_{n_2} - \nu\|_{TV}\right\}$ *(Cumulative statistical approximation error),*
$\mathcal{E}_2^* := 4\left\{\left\|\widehat{(F_{\#}\mu)}_{n_1} - F_{\#}\hat{\mu}_{n_1}\right\|_{TV} + \left\|\widehat{(G_{\#}\nu)}_{n_2} - G_{\#}\hat{\nu}_{n_2}\right\|_{TV}\right\}$ *(Total information dissipated).*

It is expected of the pair of maps that commit zero translation error (e.g., (ϕ, ψ), asymptotically) to belong to the 'kernel' of a cycle-consistent network. In other words, the realized cyclic loss should also lie near zero. To showcase the idea of reconstruction consistency, let us concentrate on the term:

$$\hat{\mathcal{L}}_{cyc}(\hat{\mu}_{n_1}, \hat{\nu}_{n_2}, \psi, \phi) = \|\mu - (\phi \circ \psi)_{\#}\hat{\mu}_{n_1}\|_1 + \|\nu - (\psi \circ \phi)_{\#}\hat{\nu}_{n_2}\|_1.$$

Since the smoothness of underlying distributions is paramount in our analysis, usage of regularly invariant kernel density estimates $(\tilde{\mu}_{n_1}, \tilde{\nu}_{n_2})$ instead may lead to improved approximation. Based on the same set of observations we build $\hat{p}_{\mu,n_1}(x) = \frac{d\tilde{\mu}_{n_1}}{dx} = \frac{1}{n_1 h^d}\sum_{i=1}^{n_1} K\left(\frac{x}{h}, \frac{x_i}{h}\right)$, $x \in \Omega_x$ where $h \equiv h(n_1)$. Similarly define $\hat{p}_{\nu,n_2}$.

Theorem 3.4. *For the pair of forward-backward maps (ψ, ϕ), as constructed in Theorem 3.1*

$$\mathbb{E}\big[\hat{\mathcal{L}}_{cyc}(\tilde{\mu}_{n_1}, \tilde{\nu}_{n_2}, \psi, \phi)\big] \precsim \max\Big\{{n_1}^{-\frac{m_x}{(d\vee 2)m_x+d}}, {n_2}^{-\frac{m_y}{(k\vee 2)m_y+k}}\Big\}.$$

The eventual nullification of the average reconstruction loss is a desirable outcome. However, the concentration of random empirical losses around such an aggregate bears more significance. On that note, we present the concluding result that embodies our idea of reconstruction consistency.

Corollary 3.2 (Regeneration consistency)**.** *As $\min(n_1, n_2) \to \infty$, we observe $(\phi\circ\psi)_{\#}\tilde{\mu}_{n_1} \to \mu$ and $(\psi\circ\phi)_{\#}\tilde{\nu}_{n_2} \to \nu$, both in total variation.*

Remark 3.7. *While the usage of 'smoother' estimates produce faster convergence rates, usual empirical distributions $(\hat{\mu}_{n_1}, \hat{\nu}_{n_2})$ also lead to an outcome similar to Corollary 3.2, given that the VC dimensions of both $\mathcal{Y}(\mathcal{P}(\mathcal{X}))$ and $\mathcal{Y}(\mathcal{P}(\mathcal{Y}))$ are finite.*

3.5 Discussion

This chapter establishes statistical translation and regeneration guarantees of cycle-consistent networks. In the process, we recommend precise recipes to build translator maps to achieve such consistency. At its time of publication, it was the first endeavour of its kind in this context. We prove that deep ReLU-based translators, being fine approximators of Lipschitz functions, asymptotically behave like IPTs. We theoretically show that for Sobolev-smooth input data, deployment of the 1-Wasserstein distance and L^1 in the cyclic loss are equivalent. This substantiates the conclusion Zhu et al. (2017) had reached for CycleGAN. A key highlight of our analysis is the absence of any restrictions on the data dimensions. We also discuss the ramifications of ill-posedness during translation and the impact it leaves on the regeneration. The decomposition of the translation and cyclic errors in the process, based on independent sources of variation, is also new in this setting.

Our analytical approach has since paved the way for further scrutiny of cycle-consistent networks. One aspect that yet lies unexplored is their robustness to outliers in the data, especially since such networks are found to be prone to self-attack. In case the target mapping is many-to-one (e.g., photos to semantic labels), the realized translators tend to hide information as a noisy component in the translated law, imperceptible to discriminators. Though effective defense mechanisms against self-attack (adversarial training with noise, and using guess discriminators) have been proposed (Bashkirova et al., 2019), deterministic bounds on the permissible departure of maps from their theoretical references remain absent.

Observe that we can always establish upper bounds in the spirit of Theorem 2.6 for cycle-consistent models. However, a more responsible line of questioning seeks a way to robustify the translations. To that end, we mention a remark made by Zhang et al. (2022). They

observe that optimizing (3.1) is rather equivalent to solving the Bidirectional Gromov-Monge (GM) divergence, under a relaxed isometry requirement, given as

$$\inf_{\substack{F\in\mathscr{F}(\mathcal{X},\mathcal{P}(\mathcal{Y}))\\ G\in\mathscr{F}(\mathcal{Y},\mathcal{P}(\mathcal{X}))}}\sup_{\substack{D_X\in\mathscr{L}_X\\ D_Y\in\mathscr{L}_Y}}\Big\{\Delta_1(\mu,\nu,F,G)+\lambda_1\mathcal{L}_{D_X}(\mu,\nu,G)+\lambda_2\mathcal{L}_{D_Y}(\nu,\mu,F)\Big\},$$

where $\Delta_p(\mu,\nu,F,G)\coloneqq\Delta^{(p)}_{\mathcal{X}}(\mu,F)+\Delta^{(p)}_{\mathcal{Y}}(\nu,G)+\Delta^{(p)}_{\mathcal{X},\mathcal{Y}}(\mu,\nu,F,G)$, such that for $p\geq 1$

$$\begin{aligned}
\Delta^{(p)}_{\mathcal{X}}(\mu,F)&\coloneqq\left(\mathbb{E}\left[\left|d_{\mathcal{X}}(X,X')-d_{\mathcal{Y}}(F(X),F(X'))\right|^p\right]\right)^{\frac{1}{p}},\\
\Delta^{(p)}_{\mathcal{Y}}(\nu,G)&\coloneqq\left(\mathbb{E}\left[\left|d_{\mathcal{X}}(G(Y),G(Y'))-d_{\mathcal{Y}}(Y,Y')\right|^p\right]\right)^{\frac{1}{p}},\ \text{and}\\
\Delta^{(p)}_{\mathcal{X},\mathcal{Y}}(\mu,\nu,F,G)&\coloneqq\left(\mathbb{E}\left[\left|d_{\mathcal{X}}(X,G(Y))-d_{\mathcal{Y}}(F(X),Y)\right|^p\right]\right)^{\frac{1}{p}}.
\end{aligned}$$

As such, finding a robust cross-domain translator boils down in principle to robustifying GM-type solvers. In Chapter 4, taking inspiration from the same, we propose a unified solution to the contamination problem in generative models with diverse spaces involved. Our methods consolidate both Chapter 3 and Chapter 2 since WAEs can also be extended to GM-type AE architectures (Nakagawa et al., 2023).

3.6 Appendix: Proofs

3.6.1 Proof of Theorem 3.1

Let us begin by fragmenting the translation error as follows,

$$d_{\mathscr{L}^1_c}(\hat{\mu}_{n_1},\phi_{\#}\hat{\nu}_{n_2})\leq d_{\mathscr{L}^1_c}(\hat{\mu}_{n_1},\phi_{\#}\nu)+d_{\mathscr{L}^1_c}(\phi_{\#}\nu,\phi_{\#}\hat{\nu}_{n_2}). \tag{3.5}$$

Before moving forward, denote the set of discrete probability measures based on at most $n\in\mathbb{N}^+$ points in $\mathbb{R}^d$ by, $\mathcal{P}_n(d)\coloneqq\left\{\sum_{i=1}^n a_i\delta_{x_i}:a_i\geq 0,\sum_{i=1}^n a_i=1,\{x_i\}_{i=1}^n\in\mathbb{R}^d\right\}$. The following lemma allows us to show that the first term on the right-hand side of (3.5) can be made arbitrarily small.

Lemma 3.3 (Yang et al. (2022)). *Let p be an absolutely continuous univariate distribution and $\pi\in\mathcal{P}_n(d)$. There exists $\phi\in\Phi(W,L)^d_1$ with $W\geq 7d+1$ and $L\geq 2$, such that whenever $n\leq\frac{W-d-1}{2}\lfloor\frac{W-d-1}{6d}\rfloor\lfloor\frac{L}{2}\rfloor+2$*

$$d_{\mathscr{L}^1_c}(\pi,\phi_{\#}p)\leq\epsilon,\ \ \textit{given}\,\epsilon>0.$$

Observe that, $\hat{\mu}_{n_1}\in\mathcal{P}_{n_1}(d)$. Now, choose $\phi\in\Phi(W,L)^d_k$ such that the first layer deploys an additional linear map A that projects ν to a one-dimensional absolutely continuous distribution first. This can always be done due to the absolute continuity of ν itself. For the

resultant map ϕ, we notice $L \geq 3$, and the specifications of W, n_1 remain as directed by Lemma 3.3. As a result, $d_{\mathscr{L}_c^1}(\hat{\mu}_{n_1}, \phi_{\#}\nu) \leq \epsilon$ for arbitrary $\epsilon > 0$.

The second term in (3.5) is the portion of the density estimation error from the base domain that translates onto the target space. To get control over such a discrepancy, we exploit the regularity of the transformation ϕ. Observe that the activation function $\sigma \equiv$ ReLU is 1-Lipschitz. The transformation carrying signal y from i^{th} layer to the next is of the form $A_i(y) = M_i y + b_i$; $M_i \in \mathbb{R}^{N_{i+1} \times N_i}$ and $b_i \in \mathbb{R}^{N_{i+1}}$, $i = 0, ..., L$. The matrix M_i can be constructed such that $\|M_i\|_p = \sup_{\|y\|_p=1} \|M_i y\|_p \leq k_i$, for some constant $k_i > 0$. For cases $p = 2$ or ∞, Anil et al. (2019) present exact techniques to ensure $\|M_i\|_p = 1$. Under such a framework, A_i's become k_i-Lipschitz transforms. Since Lipschitz functions are closed under composition, we can expect ϕ to behave similarly, with a constant k^* dependent on $\{k_0, k_1, ..., k_L\}$.

However, deep ReLU networks are much more expressive and are capable of approximating a vast array of smooth functions. Let us denote the class of L_G-Lipschitz functions mapping $(\Omega_y, c') \to (\Omega_x, c)$ as G_{Lip}, where $L_G > 0$. The following two results encapsulate our idea precisely.

Lemma 3.4. *For $\alpha, \beta \in \mathcal{P}(\mathcal{Y})$, $d_{\mathscr{L}_c^1}(\phi_{\#}\alpha, \phi_{\#}\beta) \leq 2 \inf_{g \in G_{Lip}} \|\phi - g\|_\infty + L_G \, d_{\mathscr{L}_{c'}^1}(\alpha, \beta)$.*

Proof of Lemma 3.4. Let us begin by specifying the class of discriminators $\mathscr{L}_X \equiv \mathscr{L}_c^1$. Now, given $\alpha, \beta \in \mathcal{P}(\mathcal{Y})$

$$d_{\mathscr{L}_X}(\phi_{\#}\alpha, \phi_{\#}\beta) = \sup_{l \in \mathscr{L}_X} \big[\mathbb{E}_{\phi_{\#}\alpha} l - \mathbb{E}_{\phi_{\#}\beta} l\big] = \sup_{l \in \mathscr{L}_X} \big[\mathbb{E}_\alpha(l \circ \phi) - \mathbb{E}_\beta(l \circ \phi)\big].$$

Due to the definition of supremum, for any $\epsilon > 0$ $\exists$ $l_\epsilon \in \mathscr{L}_X$ for which

$$\begin{aligned} d_{\mathscr{L}_X}(\phi_{\#}\alpha, \phi_{\#}\beta) &\leq \mathbb{E}_\alpha(l_\epsilon \circ \phi) - \mathbb{E}_\beta(l_\epsilon \circ \phi) + \epsilon \\ &= \inf_{g \in l_\epsilon \circ G_{Lip}} \Big\{\mathbb{E}_\alpha|(l_\epsilon \circ \phi) - g| - \mathbb{E}_\beta|(l_\epsilon \circ \phi) - g| + \mathbb{E}_\alpha(g) - \mathbb{E}_\beta(g)\Big\} + \epsilon \\ &\leq 2 \inf_{g' \in G_{Lip}} \Big\|\phi - g'\Big\|_\infty + \Big\{ \sup_{l \in \mathscr{L}_X} \big[\mathbb{E}_\alpha(l \circ g^*) - \mathbb{E}_\beta(l \circ g^*)\big]\Big\} + \epsilon, \ \forall \, g^* \in G_{Lip}. \end{aligned}$$

Here, $l_\epsilon \circ G_{Lip} := \{l_\epsilon \circ f : f \in G_{Lip}\}$. Now,

$$\begin{aligned} \sup_{l \in \mathscr{L}_X} \big[\mathbb{E}_\alpha(l \circ g^*) - \mathbb{E}_\beta(l \circ g^*)\big] &= \inf_{\gamma \in \Gamma(\alpha,\beta)} \int c\big(g^*(x), g^*(y)\big) d\gamma(x, y) \\ &\leq L_G \inf_{\gamma \in \Gamma(\alpha,\beta)} \int c'(x, y) d\gamma(x, y), \end{aligned} \tag{3.6}$$

where (3.6) is due to the fact that $g^* \in G_{Lip}$. As such,

$$d_{\mathscr{L}_c^1}(\phi_{\#}\alpha, \phi_{\#}\beta) \leq 2 \inf_{g' \in G_{Lip}} \Big\|\phi - g'\Big\|_\infty + L_G \, d_{\mathscr{L}_{c'}^1}(\alpha, \beta).$$

□

Lemma 3.5 (Shen et al. (2019)). *Let $g \in G_{Lip}$. Also, ϕ is the ReLU NN-induced function as given in Lemma 3.3, having width $\mathcal{O}(W)$ and depth $\mathcal{O}(L)$. Then*

$$\|\phi - g\|_\infty \leq \mathcal{O}(C_1 W^{-\frac{2}{k}} L^{-\frac{2}{k}}),$$

where $C_1 > 0$ is a constant, dependent on $L_G, \sqrt{k}$, and $diam(\Omega_y)$.

Using both lemmas, we obtain $d_{\mathscr{L}^1_c}(\hat{\mu}_{n_1}, \phi_\# \hat{\nu}_{n_2}) \leq \epsilon + L_G\, d_{\mathscr{L}^1_{c'}}(\nu, \hat{\nu}_{n_2}) + \mathcal{O}(C_1 W^{-\frac{2}{k}} L^{-\frac{2}{k}})$. The sole task remaining is to upper-bound the statistical estimation error in the base space. To that end, by applying Corollary 2.1 of Liang (2018) for $k \geq 2$, we get $\mathbb{E}_\nu[d_{\mathscr{L}^1_{c'}}(\nu, \hat{\nu}_{n_2})] \leq \mathcal{O}((k^2 n_2)^{-\frac{1}{k}})$.

3.6.2 Proof of Corollary 3.1

We have already noticed $\mathbb{E}_\nu[d_{\mathscr{L}^1_{c'}}(\nu, \hat{\nu}_{n_2})] \leq \mathcal{O}((k^2 n_2)^{-\frac{1}{k}})$, $k \geq 2$. Since the distance $d_{\mathscr{L}^1_{c'}}(.,.)$ satisfies the bounded difference inequality, the application of McDiarmid's inequality leads to

$$\mathbb{P}\Big(d_{\mathscr{L}^1_{c'}}(\nu, \hat{\nu}_{n_2}) \leq \mathcal{O}((k^2 n_2)^{-\frac{1}{k}}) + t\Big) \geq 1 - \exp\Big\{-\frac{2n_2 t^2}{{B_y}^2}\Big\}, \tag{3.7}$$

where $B_y = diam(\Omega_y)$ with respect to the metric c'. We point out that (3.7) is a generalized version of Proposition 20 in Weed and Bach (2019). Now, Theorem 3.1 tells us,

$$d_{\mathscr{L}^1_c}(\hat{\mu}_{n_1}, \phi_\# \hat{\nu}_{n_2}) \leq \epsilon + L_G\, d_{\mathscr{L}^1_{c'}}(\nu, \hat{\nu}_{n_2}) + \mathcal{O}(C_1 W^{-\frac{2}{k}} L^{-\frac{2}{k}}),$$

given $\epsilon > 0$ and $n_1 \leq \frac{W-d-1}{2} \lfloor \frac{W-d-1}{6d} \rfloor \lfloor \frac{L}{2} \rfloor + 2$. Combining these two results, we get

$$\mathbb{P}\Big(d_{\mathscr{L}^1_c}(\hat{\mu}_{n_1}, \phi_\# \hat{\nu}_{n_2}) \leq \mathcal{O}((k^2 n_2)^{-\frac{1}{k}}) + \frac{(1+L_G)B_y}{\sqrt{2}} {n_2}^{-\frac{1}{2}} \sqrt{\ln\big(\frac{1}{\delta}\big)} + \mathcal{O}(C_1 W^{-\frac{2}{k}} L^{-\frac{2}{k}})\Big) \geq 1 - \delta,$$

by taking $\delta = \exp\Big\{-\frac{2n_2 t^2}{{B_y}^2}\Big\}$. The statement also holds if we replace the two sample sizes n_1, n_2 with $\min(n_1, n_2)$. In such a case, the Borel-Cantelli lemma implies that $d_{\mathscr{L}^1_c}(\hat{\mu}_{n_1}, \phi_\# \hat{\nu}_{n_2}) \longrightarrow 0$ almost surely (under $\mathbb{P}$), provided d, k remain fixed.

Remark 3.8. *We draw the attention of the reader to a particular consequence of this result. Observe that the width (W) and depth (L) of the translator network are intrinsically related to the sample size (n_1) from the target law. In case $\min(n_1, n_2) \longrightarrow \infty$, W also follows suit, given that L remains constant. As such, our ideal backward translator, achieving generation consistency, is a finite sample approximation of an infinitely wide ReLU network. Maps induced by such an infinitely wide network converge in distribution to a Gaussian process*

de G. Matthews et al. (2018). This determines the large sample property of ϕ. Finding out the exact statistical properties of such a process in a parametric setup might be taken up as future work.

Remark 3.9. *For any $n_1 \in \mathbb{N}^+$, $d_{\mathscr{L}_c^1}(\mu, \phi_{\#}\hat{\nu}_{n_2}) \leq d_{\mathscr{L}_c^1}(\mu, \hat{\mu}_{n_1}) + d_{\mathscr{L}_c^1}(\hat{\mu}_{n_1}, \phi_{\#}\hat{\nu}_{n_2})$. We have already seen that the second term on the right-hand side of the inequality vanishes eventually [Corollary 1]. Moreover, similar to (3.7)*

$$\mathbb{P}\Big(d_{\mathscr{L}_c^1}(\mu, \hat{\mu}_{n_1}) \leq \mathcal{O}((d^2 n_1)^{-\frac{1}{d}}) + t\Big) \geq 1 - \exp\Big\{-\frac{2n_1 t^2}{{B_x}^2}\Big\}.$$

As a result, $d_{\mathscr{L}_c^1}(\mu, \hat{\mu}_{n_1}) \xrightarrow{a.s.} 0$ (using Borel-Cantelli lemma). Hence, it can be concluded that $\phi_{\#}\hat{\nu}_{n_2}$ converges weakly to μ in $\mathcal{P}(\mathcal{X})$ [Theorem 6.9 in Villani (2009)].

3.6.3 Proof of Theorem 3.2

Let us carry out the decomposition of the realized backward translation error, similar to that in Theorem 3.1.

$$d_{\mathcal{W}_1^{m,\infty}}(\hat{\mu}_{n_1}, \phi_{\#}\hat{\nu}_{n_2}) \leq d_{\mathcal{W}_1^{m,\infty}}(\hat{\mu}_{n_1}, \phi_{\#}\nu) + d_{\mathcal{W}_1^{m,\infty}}(\phi_{\#}\nu, \phi_{\#}\hat{\nu}_{n_2}).$$

Observe that $\mathcal{W}_1^{m,\infty} \subset \mathcal{W}_1^{1,\infty}$, for any positive integer m. Also, the class $\mathcal{W}_1^{1,\infty}$ is a dense subset of 1-Lipschitz functions on $\mathcal{X}$. As such, $d_{\mathcal{W}_1^{m,\infty}}(\hat{\mu}_{n_1}, \phi_{\#}\nu) \leq d_{\mathscr{L}_c^1}(\hat{\mu}_{n_1}, \phi_{\#}\nu) \leq \epsilon$, where $\epsilon > 0$ (as in the proof of Theorem 3.1).

The remaining approximation error can similarly be upper bound using the same technique. However, it would be far from tight. Let us recall the class of Hölder functions that eventually help in the pursuit of sharper bounds.

Definition 3.3. *For $s \in \mathbb{R}_{>0}$, with $\lfloor s \rfloor$ indicating the largest integer strictly smaller than s, the Hölder space of order s is defined as*

$$\mathcal{C}_L^s(\mathbb{R}^d) = \Big\{f \in C_u(\mathbb{R}^d) : \|f\|_{\mathcal{C}^s} \equiv \|f\|_{\mathcal{W}^{\lfloor s \rfloor}} + \sum_{|\alpha|=\lfloor s \rfloor} \sup_{\substack{x \neq y \\ x,y \in \mathbb{R}^d}} \frac{|D^\alpha f(x) - D^\alpha f(y)|}{|x-y|^{s-\lfloor s \rfloor}} < L\Big\}.$$

Now, similar to the proof of Lemma 3.4, for any $\epsilon^{'} > 0 \; \exists \; l_{\epsilon^{'}} \in \mathcal{W}_1^{m,\infty}$ such that

$$\begin{aligned} d_{\mathcal{W}_1^{m,\infty}}(\phi_{\#}\alpha, \phi_{\#}\beta) &\leq \mathbb{E}_\alpha(l_{\epsilon^{'}} \circ \phi) - \mathbb{E}_\beta(l_{\epsilon^{'}} \circ \phi) + \epsilon^{'}, \quad \text{where } \alpha, \beta \in \mathcal{P}(\mathcal{Y}) \\ &= \inf_{g \in l_{\epsilon^{'}} \circ G_{Lip}} \Big\{\mathbb{E}_\alpha\big|(l_{\epsilon^{'}} \circ \phi) - g\big| - \mathbb{E}_\beta\big|(l_{\epsilon^{'}} \circ \phi) - g\big| + \mathbb{E}_\alpha(g) - \mathbb{E}_\beta(g)\Big\} + \epsilon^{'} \\ &\leq 2 \inf_{g^{'} \in G_{Lip}} \Big\|\phi - g^{'}\Big\|_\infty + \Big\{\sup_{l \in \mathcal{W}_1^{m,\infty}} \big[\mathbb{E}_\alpha(l \circ g^*) - \mathbb{E}_\beta(l \circ g^*)\big]\Big\} + \epsilon^{'}, \; \forall \, g^* \in G_{Lip}. \end{aligned} \tag{3.8}$$

The first term in (3.8) is obtained due to the Lipschitz property of $l_{\epsilon'}$. Here,

$$\sup_{l\in\mathcal{W}_1^{m,\infty}}\big[\mathbb{E}_\alpha(l\circ g^*)-\mathbb{E}_\beta(l\circ g^*)\big]=d_{\mathcal{W}_1^{m,\infty}}(g^*_\#\alpha,g^*_\#\beta)\le d_{\mathcal{C}_r^m}(g^*_\#\alpha,g^*_\#\beta) \tag{3.9}$$

$$=\sup_{l\in\mathcal{C}_r^m\circ g^*}\Big\{\mathbb{E}_{x\sim\alpha}[l(x)]-\mathbb{E}_{x\sim\beta}[l(x)]\Big\}. \tag{3.10}$$

Inequality (3.9) is based on the observation that there exists $r>0$ for which $\mathcal{W}_1^{m,\infty}\subset\mathcal{C}_r^m$ Schreuder (2020). Given any $f\in\mathcal{C}_r^m$ and $g^*\in G_{Lip}$,

$$\begin{aligned}\|f\circ g^*\|_\infty=\big\{\sup|f(g^*(y))|:y\in\mathbb{R}^k\big\}&=\big\{\sup|f(x)|:x=g^*(y)\in\mathbb{R}^d,y\in\mathbb{R}^k\big\}\\&\le\big\{\sup|f(x)|:x\in\mathbb{R}^d\big\}=\|f\|_\infty.\end{aligned}$$

Moreover, for $x,y\in\mathbb{R}^k$, $x\ne y$

$$\begin{aligned}\frac{|D^\alpha f(g^*(x))-D^\alpha f(g^*(y))|}{|x-y|^{s-\lfloor s\rfloor}}&=\frac{|D^\alpha f(g^*(x))-D^\alpha f(g^*(y))|}{|g^*(x)-g^*(y)|^{s-\lfloor s\rfloor}}\Big\{\frac{|g^*(x)-g^*(y)|}{|x-y|}\Big\}^{s-\lfloor s\rfloor}\\&\le\frac{|D^\alpha f(x^*)-D^\alpha f(y^*)|}{|x^*-y^*|^{s-\lfloor s\rfloor}}(L_G)^{s-\lfloor s\rfloor},\end{aligned}$$

assuming $x^*\ne y^*\in\mathbb{R}^d$. Here, we choose both the metrics c,c' to be L_1 in their respective spaces. This convention conforms to the rest of the discussion as well.

Also, for $1\le|s|\le m$ we have

$$D^s(f\circ g^*)(x)=s!\sum_{1\le|i|\le|s|}\frac{(D^if)(g^*(x))}{i!}P_{s,i}(g^*;x),$$

where $P_{s,i}(g^*;x)$ is a homogeneous polynomial of degree $|i|$. Schreuder *et al.* [Lemma 7.2 in Schreuder et al. (2021)] show that $|D^s(f\circ g^*)(x)|<C$, where $C>0$ is a constant. This implies that there exists $r^*>0$ for which $f\circ g^*\in\mathcal{C}_{r^*}^m(\mathbb{R}^k)$. As such, we may upper bound (3.10) by replacing the supremum over $\mathcal{C}_r^m(\mathbb{R}^d)\circ g^*$ by the same over $\mathcal{C}_{r^*}^m(\mathbb{R}^k)$.

Hence, for $\epsilon>0$

$$d_{\mathcal{W}_1^{m,\infty}}(\hat\mu_{n_1},\phi_\#\hat\nu_{n_2})\le 2\inf_{g'\in G_{Lip}}\Big\|\phi-g'\Big\|_\infty+d_{\mathcal{C}_{r^*}^m}(\nu,\hat\nu_{n_2})+\epsilon.$$

The expected approximation error in the base domain can be put under a deterministic upper bound given by $\mathbb{E}_\nu\big[d_{\mathcal{C}_{r^*}^m}(\nu,\hat\nu_{n_2})\big]\precsim {n_2}^{-\frac{m}{k}}+\frac{\log n_2}{\sqrt{n_2}}$ [Lemma 2.8 in Huang et al. (2021)]. As such, we get $\mathbb{E}\big[d_{\mathcal{W}_1^{m,\infty}}(\hat\mu_{n_1},\phi_\#\hat\nu_{n_2})\big]\le\mathcal{O}\big({n_2}^{-\frac{m}{k}}+\frac{\log n_2}{\sqrt{n_2}}\big)+\mathcal{O}(\sqrt{k}L_GB_yW^{-\frac{2}{k}}L^{-\frac{2}{k}})$.

3.6.4 Proof of Lemma 3.1

Our characterization of the critics allow $\mathscr{L}_X$ to be $\mathscr{L}_c^1$ or $\mathcal{W}_1^{m,\infty}$. Under this setup, for any backward translator G

$$\begin{aligned} d_{\mathscr{L}_X}(\hat{\mu}_{n_1}, G_{\#}\hat{\nu}_{n_2}) &\leq d_{\mathscr{L}_X}(\hat{\mu}_{n_1}, \widehat{(G_{\#}\nu)}_{n_2}) + d_{\mathscr{L}_X}\big(\widehat{(G_{\#}\nu)}_{n_2}, G_{\#}\hat{\nu}_{n_2}\big) \\ &\leq B_x \left\| \hat{\mu}_{n_1} - \widehat{(G_{\#}\nu)}_{n_2} \right\|_{TV} + \mathcal{E}_3 \\ &\leq B_x \left\| \hat{\mu}_{n_1} - \Gamma_{n_1} \right\|_{TV} + \Lambda_{(n_1,n_2)} + \mathcal{E}_3, \end{aligned} \tag{3.11}$$

where $\Gamma_{n_1} = \text{argmin}_{\tau \in \mathcal{P}(\mathcal{X})} \left\| \tau - \hat{\mu}_{n_1} \right\|_{TV}$. It is often called the *Empirical Yatracos Minimizer* Devroye and Lugosi (2001). Observe that $\left\| \hat{\mu}_{n_1} - \Gamma_{n_1} \right\|_{TV} \leq \left\| \hat{\mu}_{n_1} - \mu \right\|_{TV}$. Now, in case the OT map T exists such that $T_{\#}\nu = \mu$, we get $\left\| \hat{\mu}_{n_1} - \Gamma_{n_1} \right\|_{TV} \leq \mathcal{E}_1$.

Remark 3.10. *The information loss (in the right-hand side of (3.11)) can be taken care of by deploying an IPT as the translator. As such, it is the term $d_{\mathscr{L}_X}(\hat{\mu}_{n_1}, \widehat{(G_{\#}\nu)}_{n_2})$ that mainly contributes to the upper bound. We had built the empirical distribution $\hat{\mu}_{n_1}$ based on $\{X_i\}_{i=1}^{n_1} \overset{i.i.d.}{\sim} \mu$. Similarly, let $\widehat{(G_{\#}\nu)}_{n_2}$ be based on $\{Y_i\}_{i=1}^{n_2} \overset{i.i.d.}{\sim} G_{\#}\nu$. We may write*

$$d_{\mathscr{L}_X}(\hat{\mu}_{n_1}, \widehat{(G_{\#}\nu)}_{n_2}) = \sup_{f \in \mathscr{L}_X} \left| \sum_{i=1}^{N} W_i f(Z_i) \right|, \tag{3.12}$$

where $N = n_1 + n_2$; $W_i = \frac{1}{n_1}$ when $Z_i = X_i$, $i = 1, ..., n_1$ and $W_{n_1+j} = -\frac{1}{n_2}$ when $Z_{n_1+j} = Y_j$, $j = 1, ..., n_2$. Under this framework, the solution to (3.12) can be achieved by solving a linear program, given that $\mathscr{L}_X \equiv \mathscr{L}_c^1$ [Theorem 2.1 in Sriperumbudur et al. (2012)]. This provides a pathway to get hold of the realized approximation error, making the upper bound deterministic.

3.6.5 Proof of Lemma 3.2

Given translator maps $G \in \mathscr{F}(\mathcal{Y}, \mathcal{P}(\mathcal{X}))$ and $F \in \mathscr{F}(\mathcal{X}, \mathcal{P}(\mathcal{Y}))$, the cyclic loss in the space $\mathcal{X}$ can be broken down as the following:

$$\left\| \mu - (G \circ F)_{\#}\mu \right\|_1 \leq \left\| \mu - G_{\#}\nu \right\|_1 + \left\| G_{\#}\nu - (G \circ F)_{\#}\mu \right\|_1,$$

where

$$\begin{aligned} \left\| G_{\#}\nu - (G \circ F)_{\#}\mu \right\|_1 = \left\| G_{\#}\nu - G_{\#}(F_{\#}\mu) \right\|_1 &= 2 \sup_{\omega \subseteq \sigma(\mathcal{X})} \left| G_{\#}\nu(\omega) - G_{\#}(F_{\#}\mu)(\omega) \right| \\ &= 2 \sup_{\omega \subseteq \sigma(\mathcal{X})} \left| \nu(G^{-1}(\omega)) - F_{\#}\mu(G^{-1}(\omega)) \right| \\ &\leq 2 \sup_{\omega' \subseteq \sigma(\mathcal{Y})} \left| \nu(\omega') - F_{\#}\mu(\omega') \right| = \left\| \nu - F_{\#}\mu \right\|_1. \end{aligned}$$

The inequality holds by taking supremum over all measurable sets belonging to the Borel σ-algebra on $\mathcal{Y}$ instead of the particular path directed by G^{-1}. As such

$$\|\mu - (G \circ F)_{\#}\mu\|_1 \leq \|\mu - G_{\#}\nu\|_1 + \|\nu - F_{\#}\mu\|_1.$$

Similarly, $\|\nu - (F \circ G)_{\#}\nu\|_1 \leq \|\nu - F_{\#}\mu\|_1 + \|\mu - G_{\#}\nu\|_1$. Hence the proof.

3.6.6 Proof of Theorem 3.3

Given a measurable function $f : \mathbb{R}^d \to \mathbb{R}$, let us define its *convolution* with the kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as the following:

$$K_h(f) = \int_{\mathbb{R}^d} K_h(., y) f(y) dy = \frac{1}{h^d} \int_{\mathbb{R}^d} K(\frac{.}{h}, \frac{y}{h}) f(y) dy,$$

where $\frac{y}{h} = (\frac{y_1}{h}, ..., \frac{y_d}{h})'$, $h > 0$. We begin by taking K to be regularly invariant. Now,

$$\begin{aligned} \left\|p_\mu - p_{\phi_{\#}\nu}\right\|_1 &\leq \|p_\mu - K_h(p_\mu)\|_1 + \left\|K_h(p_\mu) - K_h(p_{\phi_{\#}\nu})\right\|_1 + \left\|K_h(p_{\phi_{\#}\nu}) - p_{\phi_{\#}\nu}\right\|_1 \\ &\leq J\|p_\mu - K_h(p_\mu)\|_p + \left\|K_h(p_\mu) - K_h(p_{\phi_{\#}\nu})\right\|_1 + J\left\|K_h(p_{\phi_{\#}\nu}) - p_{\phi_{\#}\nu}\right\|_{p'}, \end{aligned} \tag{3.13}$$

where $J > 0$. The existence of such a constant, and hence the inequality (3.13), is ensured by the fact $\|f\|_1 \leq J\|f\|_p$, $p \geq 1$ since we have $\lambda(\Omega_x) < \infty$. Also, there exists a constant l depending upon m_x and K, such that $\|K_h(p_\mu) - p_\mu\|_p \leq l\|D^{m_x} p_\mu\|_p h^{m_x}$ [Proposition 4.3.33 in Giné and Nickl (2021)]. As such, we get hold of a constant $J^* = Jl$ for which

$$\left\|p_\mu - p_{\phi_{\#}\nu}\right\|_1 \leq J^* \left\{\|D^{m_x} p_\mu\|_p + \left\|D^{m_x} p_{\phi_{\#}\nu}\right\|_{p'}\right\} h^{m_x} + \left\|K_h(p_\mu) - K_h(p_{\phi_{\#}\nu})\right\|_1$$

(by Assumption 3.2). Observe that,

$$K_h(p_\mu)(x) - K_h(p_{\phi_{\#}\nu})(x) = \frac{1}{h^d} \int \left\{K(\frac{x}{h}, \frac{y}{h}) - K(\frac{x}{h}, \frac{z}{h})\right\} d\kappa(y, z),$$

where κ is a coupling between μ and $\phi_{\#}\nu$. Hence,

$$\begin{aligned} \left\|K_h(p_\mu) - K_h(p_{\phi_{\#}\nu})\right\|_1 &\leq \int \left\{\frac{1}{h^d} \int \left|K(\frac{x}{h}, \frac{y}{h}) - K(\frac{x}{h}, \frac{z}{h})\right| dx\right\} d\kappa(y, z) \qquad (3.14) \\ &= \int \left\{\frac{\int \left|K(x', \frac{y}{h}) - K(x', \frac{z}{h})\right| dx'}{|y - z|}\right\} |y - z| d\kappa(y, z) \\ &\leq \frac{M^*}{h} \int |y - z| d\kappa(y, z), \qquad (3.15) \end{aligned}$$

where M^* is a positive constant. The step (3.14) is due to Jensen's inequality, whereas (3.15) exploits the invariance of K. Since the inequality holds for all possible measure couples κ, we conclude

$$\left\|K_h(p_\mu) - K_h(p_{\phi_\#\nu})\right\|_1 \leq \frac{M^*}{h} W_c^1(\mu, \phi_\#\nu),$$

given that $c \equiv L_1$. A similar inference can be drawn for a general class of metrics c by altering the specification of the same in the definition of invariance. Now, choose

$$h = \left\{ \frac{W_c^1(\mu, \phi_\#\nu)}{\|D^{m_x} p_\mu\|_p + \left\|D^{m_x} p_{\phi_\#\nu}\right\|_{p'}} \right\}^{\frac{1}{m_x+1}}.$$

Finally, we obtain

$$\left\|p_\mu - p_{\phi_\#\nu}\right\|_1 \leq M \left[\|D^{m_x} p_\mu\|_p + \left\|D^{m_x} p_{\phi_\#\nu}\right\|_{p'} \right]^{\frac{1}{m_x+1}} \left[W_c^1(\mu, \phi_\#\nu)\right]^{\frac{m_x}{m_x+1}},$$

where $M = 2(J^* \vee M^*)$.

3.6.7 Proof of Proposition 3.1

Using Lemma 3.2,

$$\begin{aligned}\mathcal{L}_{cyc}(\hat{\mu}_{n_1}, \hat{\nu}_{n_2}, F, G) &= \|\hat{\mu}_{n_1} - (G \circ F)_\#\hat{\mu}_{n_1}\|_1 + \|\hat{\nu}_{n_2} - (F \circ G)_\#\hat{\nu}_{n_2}\|_1 \\ &\leq 4\big\{\|\hat{\mu}_{n_1} - G_\#\hat{\nu}_{n_2}\|_{TV} + \|\hat{\nu}_{n_2} - F_\#\hat{\mu}_{n_1}\|_{TV}\big\}.\end{aligned}$$

Now, a similar decomposition of the translation errors under the TV metric, as in the proof of Lemma 3.1, results in the following:

$$\begin{aligned}\|\hat{\mu}_{n_1} - G_\#\hat{\nu}_{n_2}\|_{TV} &\leq \|\hat{\mu}_{n_1} - \Gamma_{n_1}\|_{TV} + \left\|\Gamma_{n_1} - \widehat{(G_\#\nu)}_{n_2}\right\|_{TV} + \left\|\widehat{(G_\#\nu)}_{n_2} - G_\#\hat{\nu}_{n_2}\right\|_{TV} \\ &\leq \|\hat{\mu}_{n_1} - \mu\|_{TV} + \frac{\Lambda_{(n_1,n_2)}}{B_x} + \left\|\widehat{(G_\#\nu)}_{n_2} - G_\#\hat{\nu}_{n_2}\right\|_{TV}.\end{aligned}$$

Similarly, given that $\Gamma'_{n_2} = \text{argmin}_{\tau \in \mathcal{P}(\mathcal{Y})} \|\tau - \hat{\nu}_{n_2}\|_{TV}$

$$\|\hat{\nu}_{n_2} - F_\#\hat{\mu}_{n_1}\|_{TV} \leq \|\hat{\nu}_{n_2} - \nu\|_{TV} + \frac{\Lambda'_{(n_1,n_2)}}{B_y} + \left\|\widehat{(F_\#\mu)}_{n_1} - F_\#\hat{\mu}_{n_1}\right\|_{TV}.$$

3.6.8 Proof of Theorem 3.4

Let $\phi \in \Phi(W, L)_k^d$, as specified in Theorem 3.1. Also, let $\psi \in \Phi(W', L')_d^k$ be a forward translator that achieves consistency. Observe that

$$\begin{aligned}\hat{\mathcal{L}}_{cyc}(\tilde{\mu}_{n_1}, \tilde{\nu}_{n_2}, \psi, \phi) &\leq \|\tilde{\mu}_{n_1} - \mu\|_1 + \|\tilde{\nu}_{n_2} - \nu\|_1 + \mathcal{L}_{cyc}(\mu, \nu, \psi, \phi) \\ &\leq \|\tilde{\mu}_{n_1} - \mu\|_1 + \|\tilde{\nu}_{n_2} - \nu\|_1 + 2\Big\{\|\mu - \phi_{\#}\nu\|_1 + \|\nu - \psi_{\#}\mu\|_1\Big\}. \qquad (3.16)\end{aligned}$$

For $1 \leq p, q < \infty$, we know that

$$\mathbb{E}\left[\|\hat{p}_{\mu,n_1} - p_\mu\|_p\right] \precsim {n_1}^{-\frac{m_x}{2m_x+d}},$$

[Theorem 6.1 in Cleanthous et al. (2019)]. Similarly, for the estimation error in $\mathcal{Y}$, $\mathbb{E}\left[\|\hat{p}_{\nu,n_2} - p_\nu\|_q\right] \precsim {n_2}^{-\frac{m_y}{2m_y+k}}$. Moreover, Theorem 3.3 implies that

$$\Big\{\|p_\mu - p_{\phi_{\#}\nu}\|_1\Big\}^{\frac{m_x+1}{m_x}} \leq R\, d_{\mathscr{L}_c^1}(\mu, \phi_{\#}\nu) \leq R\left\{d_{\mathscr{L}_c^1}(\mu, \hat{\mu}_{n_1}) + d_{\mathscr{L}_c^1}(\hat{\mu}_{n_1}, \phi_{\#}\nu)\right\}, \qquad (3.17)$$

where $R = M^{\frac{m_x+1}{m_x}}\left[\|D^{m_x}p_\mu\|_p + \|D^{m_x}p_{\phi_{\#}\nu}\|_{p'}\right]^{\frac{1}{m_x}}$, and $\hat{\mu}_{n_1}$ is an usual empirical measure corresponding to μ. The term $d_{\mathscr{L}_c^1}(\hat{\mu}_{n_1}, \phi_{\#}\nu)$ can be made arbitrarily small due to the construction of ϕ [Lemma 3.3]. Also, we have already seen that $\mathbb{E}\left[d_{\mathscr{L}_c^1}(\mu, \hat{\mu}_{n_1})\right] \precsim {n_1}^{-\frac{1}{d}}$.

As such,

$$\mathbb{E}\left[\|\tilde{\mu}_{n_1} - \mu\|_1 + 2\|\mu - \phi_{\#}\nu\|_1\right] \leq \mathcal{O}\left({n_1}^{-\frac{m_x}{(d\vee 2)m_x+d}}\right),$$

by applying Jensen's inequality to (3.17). This bound, together with a similar result corresponding to its forward counterpart, will imply

$$\mathbb{E}\left[\hat{\mathcal{L}}_{cyc}(\tilde{\mu}_{n_1}, \tilde{\nu}_{n_2}, \psi, \phi)\right] \precsim \max\left\{{n_1}^{-\frac{m_x}{(d\vee 2)m_x+d}}, {n_2}^{-\frac{m_y}{(k\vee 2)m_y+k}}\right\}.$$

3.6.9 Proof of Corollary 3.2

We point out that, $K(x, y)$ can be taken in particular as $\tilde{K}(|x - y|)$, where $\tilde{K} : \mathbb{R}^d \to \mathbb{R}$ identically follows the traits of K. Under such a kernel function,

$$\left\|\mathbb{E}[\hat{p}_{\mu,n_1}] - p_\mu\right\|_1 \leq l^* h^{m_x},$$

for some constant $l^* > 0$ (Giné and Nickl, 2021). Now, given an $\epsilon \leq \frac{2}{3}$, concentration inequalities on kernel density estimates tell us: there exist constants $E_1, E_2 > 0$ such that

$$\mathbb{P}\Big(\left\|\hat{p}_{\mu,n_1} - \mathbb{E}[\hat{p}_{\mu,n_1}]\right\|_\infty > \epsilon\Big) \leq E_1\Big(\frac{\sqrt{d}B_x}{h^{d+1}\epsilon}\Big)^d \exp\Big\{\big(-E_2 n_1 \epsilon^2 h^d\big)\Big\}.$$

The exact value of $E_2 = \frac{3}{28\tilde{K}(0)}$ can be obtained based on the convention that $\tilde{K}(.)$ achieves its modal value at 0. Such a centering can always be done. Hence,

$$\mathbb{P}\Big(\|\hat{p}_{\mu,n_1} - p_\mu\|_1 > \epsilon + l^* h^{m_x}\Big) \leq E_1 \Big(\frac{\sqrt{d}B_x}{h^{d+1}\epsilon}\Big)^d \exp\Big\{\big(-E_2 n_1 \epsilon^2 h^d\big)\Big\}. \tag{3.18}$$

By applying Borel-Cantelli lemma one can show that $\|\hat{p}_{\mu,n_1} - p_\mu\|_1 \longrightarrow 0$ almost surely, under suitable choice of $h \equiv h(n_1, m_x, d)$. (3.18) inspires a similar concentration for the estimate $\hat{p}_{\nu,n_2}$ around p_ν, under L^1. As such, by taking the corresponding bandwidth $h^{'} \equiv h^{'}(n_2, m_y, k)$, it can also be said that $\|\hat{p}_{\nu,n_2} - p_\nu\|_1 \longrightarrow 0$ almost surely. To unify the two processes, one may assess the convergence based on $n = \min\{n_1, n_2\}$. Putting these results back in (3.16), along with (3.17), we conclude

$$\hat{\mathcal{L}}_{cyc}(\tilde{\mu}_{n_1}, \tilde{\nu}_{n_2}, \psi, \phi) \longrightarrow 0, \text{ almost surely.}$$

In other words, $(\phi \circ \psi)_{\#}\tilde{\mu}_{n_1} \to \mu$ and $(\psi \circ \phi)_{\#}\tilde{\nu}_{n_2} \to \nu$, both in total variation.

Identity loss

Let us first rewrite the identity loss in terms of the underlying measures. Based on the notations in our framework,

$$\mathcal{L}_{id}(\mu, \nu, F, G) = \|\mu - F_{\#}\mu\|_1 + \|\nu - G_{\#}\nu\|_1.$$

Observe that the distributions must be equivariate to conform to this loss. Moreover,

$$\|\mu - \nu\|_1 - \|F_{\#}\mu - \nu\|_1 \leq \|\mu - F_{\#}\mu\|_1. \tag{3.19}$$

If the forward translated law $F_{\#}\mu$ is Sobolev-smooth of order m_y (Assumption 3.2), Theorem 3.3 asserts the existence of a constant $R^{'} > 0$ such that $\big\|p_\nu - p_{F_{\#}\mu}\big\|_1 \leq R^{'}\big[d_{\mathscr{L}^1_{c'}}(\nu, F_{\#}\mu)\big]^{\frac{m_y}{m_y+1}}$. In case F is also translation consistent, the second term on the left-hand side of (3.19) vanishes. A similar conclusion can be drawn for the quantity $\|\nu - G_{\#}\nu\|_1$ as well. As such, the cumulative identity loss from both domains cannot be minimized beyond the intrinsic discrepancy between the input distributions.

Chapter 4

Robustifying Cross-Domain Generative Models

Summary

The Gromov-Wasserstein (GW) distance is an effective measure of alignment between distributions supported on distinct ambient spaces. Calculating essentially the mutual departure from isometry, it has found vast usage in domain translation and network analysis. It has long been shown to be vulnerable to contamination in the underlying measures. All efforts to introduce robustness in GW have been inspired by similar optimal transport (OT) techniques, which predominantly advocate partial mass transport or unbalancing. In contrast, the cross-domain alignment problem, being fundamentally different from OT, demands specific solutions to tackle diverse applications and contamination regimes. Deriving from robust statistics, we discuss three contextually novel techniques to robustify GW and its variants. For each method, we explore metric properties and robustness guarantees along with their co-dependencies and individual relations with the GW distance. For a comprehensive view, we empirically validate their superior resilience to contamination under real machine learning tasks against state-of-the-art methods.

4.1 Introduction

Aligning unalike objects (images, networks, point clouds, etc.) based on their geometry remains the crux of machine learning challenges such as style transfer, graph correspondence, and shape matching. The first hint of a statistical measure of discrepancy between two such distinct distributions came in the form of Gromov-Wasserstein distance (Mémoli, 2011), quickly finding continual application in data alignment (Demetci et al., 2022), clustering (Chowdhury and Needham, 2021; Gong et al., 2022), and dimensionality reduction (Clark et al., 2024). Emerging as an L^p-relaxation of the Gromov-Hausdorff distance, it calculates the minimal distortion between replicates from distributions μ and ν, themselves defined on

spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively. In other words,

$$\inf_{\pi} \left\| d_X(x, x') - d_Y(y, y') \right\|_{L^p(\pi \otimes \pi)},$$

where π denotes a coupling between distributions (μ, ν) and the spaces are endowed with the respective metrics d_X and d_Y, $p \geq 1$. We note that the metric space $(\mathcal{X}, d_X)$ coupled with the measure μ defines a *metric measure* (mm) space. Resembling the Kantorovich formulation in OT, it immediately inspires a Monge-like upper bound to the distance (namely, Gromov-Monge (GM)), given by

$$\inf_{\phi} \left\| d_X(x, x') - d_Y(\phi(x), \phi(x')) \right\|_{L^p(\mu \otimes \mu)},$$

where the infimum is instead over measure preserving maps $\phi : \text{supp}(\mu) \to \text{supp}(\nu)$. In both cases, the underlying cost, measuring the extent of departure from strong isometry, differentiates the problem from mass transportation. It is rather the susceptibility to contamination that unites alignment and OT. The value of GW can be arbitrarily perturbed only by implanting an arbitrarily 'outlying' observation. However, the defense against such outliers in the context of alignment turns out to be much more nuanced compared to OT (see Section 4.4). Its diverse applications, coupled with the objective of aligning geometries, demand unique solutions in different contexts. The very formulation of GW also hints towards several avenues to search for a robust formulation. On the other hand, existing approaches to robustify GW and its progenies all draw insight from similar techniques in OT. While relaxing the optimization following *partial* (Chapel et al., 2020) or *unbalanced* (Séjourné et al., 2021) OT fosters capable solutions, it is perhaps unfounded to expect them to serve every context. For example, relieving the set of feasible couplings from meeting the marginal constraints also takes away metric properties. Moreover, despite showing that image-to-image (I2I) translation architectures such as CycleGAN (Cycle-consistent Generative Adversarial Networks) are indeed special cases of GM-like distances (Zhang et al., 2022), current literature does not provide a pathway to accurate generation under contaminated source data. As remedies, this chapter analyzes three principal means of robustifying the cross-domain alignment problem. This way, besides suggesting solutions to the problems discussed in earlier chapters, we address the larger landscape of contamination. We refer the reader to Section 4.4 for a detailed outline of the discussion.

Contributions: The key takeaways of our discussion are as follows.

- The first method introduces penalization to large distortions while calculating GW in the spirit of Tukey and Huber. In context, it gives rise to relaxed GW distances that preserve topologies and usual metric properties (Proposition 4.1). We show that GW, under Tukey's penalization, becomes robust to Huber contamination (Theorem 4.1) and promotes resilience to underlying distributions (Corollary 4.1). Provably, it extends the

Robust OT (ROBOT) to distributions supported on distinct mm spaces. We provide algorithms to calculate the Tukey and Huber GW distances, which in applications such as shape matching exhibit superior performance compared to existing techniques, under contamination. We also suggest data-dependent parameter tuning schemes that produce precise levels of robustness.

- Offering a finer control over extreme pairwise distances from either space, the second method rather deploys relaxed metrics that preserve topology. The resultant *locally* robust distance, surrogate to GW, becomes a lower bound to the first formulation (lemma 4.1). We prove that solving the same boils down to calculating an OT between truncated observations from μ and ν (Theorem 4.2). We also show that the notion can be generalized to define robust distances over probabilistic mm spaces. This eventually leads to a framework that offers denoising capability to Image translation models, assuming a shared latent space.

- The third approach regularizes the optimization based on 'clean' proxy distributions to achieve robust measure-preserving maps. We show its connection to robust OT formulations (lemma 4.2) and the sample complexity that such optimizations demand under contamination. The resultant optimization generalizes the notion of partial alignment, as plans corresponding to the latter can be shown to be an amenable candidate of ours. Based on the same, we propose RRGM, a novel image-to-image translation architecture that exhibits superior denoising capacity while generating handwritten digit images under contamination.

4.2 Background

Recovering unperturbed transport plans under contamination poses a significant challenge in cross-domain alignment. In most treatments based on GW (and Sturm's GW (Sturm, 2006)), the *unbalanced* relaxation to the class of underlying couplings is used to ensure robustness (De Ponti and Mondino, 2022; Séjourné et al., 2021). As a result, the 'denoised' solutions in both spaces become merely positive Radon measures. In case the marginal constraints are imposed using the TV norm (instead of Csiszár or ϕ-divergence), the idea boils down to transporting only a fraction of the mass under the distributions (Bai et al., 2024; Chapel et al., 2020). UCOOT's (Tran et al., 2023) robust formulation to deter Huber contamination utilizes a similar relaxation additionally on the feature spaces of the domains. While such a mass-trimming approach penalizes outliers, the resultant distance suffers significant deviations from its balanced counterpart (Nguyen et al., 2023). Moreover, the alignment problem, fundamentally different from mass transportation, raises more unanswered questions. For example, in most image-to-image translation problems, only one domain runs the risk of

contamination. Unbalancing turns out to be ill-posed to handle such a semi-constrained robustification (Le et al., 2021). Also, the landscape of contamination models stretches way beyond that of Huber's, which the current unbalanced techniques are solely equipped to deal with. The most recent technique offering robustness in GW alignment (Kong et al., 2024) reinforces unbalancing, based on *inlying* surrogate distributions over graphs. This, essentially being an upper bound to UGW, carries all the aforementioned issues. As such, a detailed exploration of robust alignment between distinct domains subject to diverse underlying tasks remains overdue.

Notations: We reiterate some notational conventions from earlier chapters to improve readability. Given a Polish space $\mathcal{X}$, we denote by $\mathcal{P}(\mathcal{X})$ and $\mathcal{M}(\mathcal{X})$ the set of Borel probability measures and signed Radon measures defined on it, respectively. For $p \in [1, \infty)$, measures $\rho \in \mathcal{P}(\mathcal{X})$ with finite p-th absolute moment, $M_p(\rho) \coloneqq \int \|x\|^p \rho(dx) < \infty$ form the space $\mathcal{P}_p(\mathcal{X})$. The Total Variation (TV) norm of $\rho \in \mathcal{M}(\mathcal{X})$ is denoted as $\|\rho\|_{\mathrm{TV}} \coloneqq \frac{1}{2}|\rho|(\mathcal{X})$. The space of measurable functions $f : \mathcal{X} \to \mathbb{R}$ satisfying $\|f\|_{L^p(\rho)} \coloneqq (\int |f|^p d\rho)^{1/p} < \infty$ is denoted by $L^p(\rho)$. The pushforward of $\rho \in \mathcal{P}(\mathcal{X})$ by a measurable map f is defined as $f_{\#}\mu = \mu(f^{-1})$. We define the uniform norm as $\|f\|_\infty \coloneqq \sup_{x \in \mathcal{X}} |f(x)|$. The notation $\circledcirc$ denotes the tensor-matrix multiplication, whereas $\odot$ and $\oslash$ signify element-wise product and division in matrices, respectively. The notation used for the Frobenius norm is $\|\cdot\|_{\mathrm{F}}$. Given $a, b \in \mathbb{R}$, we write $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. The uniform ε-covering number of a class of functions $\mathcal{F}$, based on n points $\{x_i\}_{i=1}^n$, with respect to (w.r.t.) the metric $d(f, f') \coloneqq \max_{i \in [n]} |f(x_i) - f'(x_i)|$ is denoted as $N_\infty(\varepsilon, \mathcal{F}, n)$. We also write inequalities, suppressing absolute constants, as $\lesssim$ and $\gtrsim$. In case $a \lesssim b$, we equivalently write $a = O(b)$. Given that the previous relation holds for a polylogarithmic function of b (i.e., $a = O(b \log^{O(1)} b)$), we write $a = \tilde{O}(b)$. If there exists a (strong) isometry between the spaces $\mathcal{X}$ and $\mathcal{Y}$, we write $\mathcal{X} \cong \mathcal{Y}$.

4.3 Preliminaries

Before introducing our robust formulations, we review the basics of transportation and alignment between metric measure spaces.

Optimal Transport and Entropic Regularization: Given a Polish space $\mathcal{X}$ endowed with a metric $d(\cdot, \cdot)$, the OT problem between $\mu, \nu \in \mathcal{P}(\mathcal{X})$ is defined as

$$\mathrm{OT}_c(\mu, \nu) \coloneqq \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} c(x, y) d\pi(x, y), \tag{4.1}$$

where $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is the lower semi-continuous transportation cost and $\Pi(\mu, \nu) = \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) : \pi(\cdot \times \mathcal{X}) = \mu, \pi(\mathcal{X} \times \cdot) = \nu\}$ is the set of couplings between μ and ν. We note that (4.1) is the Kantorovich formulation and can be shown to possess a minimizer. Given

$c(x,y) = d(x,y)^p$, $p \geq 1$ it defines the p-Wasserstein metric $W_p(\mu,\nu) \coloneqq [\mathrm{OT}_c(\mu,\nu)]^{1/p}$ on the space $\mathcal{P}_p(\mathcal{X})$ and metrizes weak convergence (Villani, 2009). OT is a typical linear program, and it is an entropic regularization that makes it strictly convex. Given a parameter $\varepsilon > 0$, reinforcing the marginal constraint under the Kullback-Leibler (KL) divergence yields the primal Entropic OT (EOT) problem:

$$\mathrm{EOT}_c^{\varepsilon}(\mu,\nu) \coloneqq \mathrm{OT}_c(\mu,\nu) + \varepsilon d_{\mathrm{KL}}(\pi|\mu\otimes\nu). \tag{4.2}$$

Unlike its unregularized counterpart, the convergence rate corresponding to the empirical EOT cost (towards the population limit) becomes devoid of $\dim(\mathcal{X})$ (Mena and Niles-Weed, 2019). Entropic regularization also enables computing δ-approximate estimates of the transport cost in $\tilde{O}(n^2/\delta)$ time (Blanchet et al., 2024). Despite computational and theoretical prowess, observe that the EOT cost (also OT) and corresponding potentials can be arbitrarily perturbed if either μ or ν (or both) is perturbed the slightest in TV.

The Gromov-Wasserstein distance: As mentioned before, we call the triplet $(\mathcal{X}, d, \mu)$ a metric measure space, where μ has full support, i.e. $\mathrm{supp}(\mu) = \mathcal{X}$. While it is technically convenient to define GW as an extension of OT between two distinct mm spaces, we differentiate them based on their origins in mass transportation and object alignment. Given two Polish mm spaces $(\mathcal{X}, d_X, \mu)$ and $(\mathcal{Y}, d_Y, \nu)$, the GW distance in all its generality is defined as

$$d_{\mathrm{GW}}(\mu,\nu) \coloneqq \Big(\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}} [\Lambda(d_X(x,x'), d_Y(y,y'))]^p d\pi\otimes\pi\Big)^{\frac{1}{p}}, \tag{4.3}$$

where $\Lambda : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ is a pseudometric measuring the extent of distortion, $1 \leq p < \infty$. We also sparingly write $d_{\mathrm{GW}}(X,Y)$. Observe that (4.3) is essentially the L^p-relaxation of the Gromov-Hausdorff distance (Mémoli (2011), Section 4.1), an operation similar to what leads to the Kantorovich-Rubinstein formulation in OT (W_p). Now, considering $\Lambda = \Lambda_q(a,b) \coloneqq \frac{1}{2}|a^q - b^q|^{1/q}$, $q < \infty$ one can recover the (p,q)-GW distance (Arya et al., 2024), which induces a metric over the class of strongly isomorphic[1] mm spaces with finite p-diameter, i.e. $\int_{\mathcal{X}\times\mathcal{X}}[d_X(x,x')]^p\mu_X(dx)\mu_X(dx') < \infty$. Different choices of d_X, d_Y also lead to interesting variants of the GW distance, e.g., considering $d_X = \langle\cdot,\cdot\rangle$ (with $p=2, q=1$) and $\|\cdot - \cdot\|$ (with $p=4, q=2$) makes the corresponding distances invariant to orthogonal transformations and translations, respectively. Bauer et al. (2024) proposes $\mathcal{Z}$-GW distances by further generalizing $d_X : \mathcal{X}\times\mathcal{X} \to \mathcal{Z}$ (also d_Y) as network kernels, given any complete and separable metric space $\mathcal{Z}$. Despite enjoying structural maneuverability, unlike OT, GW distances pose a quadratic assignment problem (QAP) and are, in general, NP-hard to compute. While it is still feasible to determine the exact value of $(4,2)$-GW between spheres (Arya et al.,

[1] $(\mathcal{X}, d_X, \mu)$ and $(\mathcal{Y}, d_Y, \nu)$ are said to be *strongly isomorphic* if there exists a measure preserving isometry $\phi : \mathcal{X} \to \mathcal{Y}$ (i.e. $\phi_{\#}\mu = \nu$ and $d_Y(\phi(x), \phi(x')) = d_X(x,x')$) which is also a bijection.

2024), given samples from arbitrary distributions, one must resort to entropic regularization to ensure computational tractability (Rioux et al., 2023; Scetbon et al., 2022). Following the setup in (4.3), the Entropic GW (EGW) distance is defined as

$$\mathrm{EGW}^{\varepsilon}(\mu,\nu) \coloneqq d_{\mathrm{GW}}(\mu,\nu) + \varepsilon d_{\mathrm{KL}}(\pi|\mu\otimes\nu). \tag{4.4}$$

This becomes particularly useful in case both the mm spaces are Euclidean with $(\mu,\nu) \in \mathcal{P}_4(\mathcal{X}) \times \mathcal{P}_4(\mathcal{Y})$, as it ties the underlying $(4,2)$-EGW[2] optimization to EOT with an altered cost. However, the issue regarding uncontrolled perturbation under contamination still persists.

4.4 Robustifying Gromov-Wasserstein

Formulating a mechanism that forestalls the effects of contamination in GW is more elusive compared to OT. Firstly, there is the context of the underlying optimization itself. In OT, the treatment ensuring robustness differs based on the task at hand. For example, cases that prioritize a divergence (e.g., generative models requiring a robust loss) usually call for a robust surrogate to W_p only. As a result, relaxations such as unbalancing or mass truncation (equivalently, addition) are often appropriate (Nietert et al., 2022, 2023). The goal in such cases lies mainly to recover $W_p(\mu,\nu)$ based on a robust proxy $W_p^{\epsilon}(\hat{\mu}_n,\hat{\nu}_n)$, i.e. $|W_p(\mu,\nu) - W_p^{\epsilon}(\hat{\mu}_n,\hat{\nu}_n)| \to 0$ in probability, where $\epsilon > 0$ denotes the radius of robustness. This can also be achieved by defining a margin on the extent of allowable perturbation while choosing the surrogate (Raghvendra et al., 2024). While such formulations preserve sample complexity, the resultant transport plans (π_{ϵ}^*) do not carry robust marginals that are also necessarily probability distributions. This becomes crucial when one is also interested in finding a robust measure-preserving map $(T_{\epsilon} : \mathrm{supp}(\mu) \to \mathrm{supp}(\nu))$ between the two distributions in the sense of Monge. A surrogate loss ignoring the marginal constraints is bound to result in a map whose deviation from the oracle (T^*) has a non-vanishing lower bound (i.e. there exists $\tau_{\epsilon} > 0$ such that $\|T_{\epsilon} - T^*\| \gtrsim \tau_{\epsilon}$). In this regard, Balaji et al. (2020); Le et al. (2021) (ROT) maintains a balanced transport by optimizing over proxy distributions instead. While statistical properties of the resulting plans remain unexamined, KL-enforced regularization makes them tractable with comparable efficacy ($\tilde{O}(n^2/\delta)$, where $\delta > 0$ is the error margin and n is the sample size.).

Due to its role in alignment (e.g., shape matching) and the involvement of two distinct mm spaces, one needs to be more cautious in approaching the GW problem using similar techniques. Observe that, $d_{\mathrm{GW}}(\mu,\nu)$ calculates the optimal p-distortion of a coupling between μ and ν (i.e. $||\Lambda(d_X(x,x'),d_Y(y,y'))||_{L^p(\pi\otimes\pi)}$). As such, it may become *extremely*

[2]Essentially the square root of the $(4,2)$-EGW distance. A more convenient way of realizing it is to assume $\Lambda_q(a,b) \coloneqq \frac{1}{2}|a^q - b^q|$ instead, under which the parameters become $p = 2, q = 2$ (Rioux et al., 2023).

fragile (Blumberg et al. (2014), Proposition 4.3) and sustain uncontrolled fluctuation if a single observation from either space is perturbed heavily. Unbalancing readily limits mass allocation to such outliers, resulting in a robust surrogate to d_{GW}. However, unlike OT, it risks sacrificing geometric information contained solely in pairwise distances. This creates significant misalignment between the resultant plan and the isometric benchmark. The latter technique of penalizing the distributions (μ and ν) themselves based on robust proxies also needs additional consideration. For example, when only one of them is contaminated (semi-constrained), the focus must lie on robustifying the pairwise distances, which is not the same as making the law robust. The problem is further confounded if near-isometric robust Monge maps (Dumont et al., 2024)(bidirectional, in case of Reverse Gromov-Monge (Hur et al., 2024)) are sought.

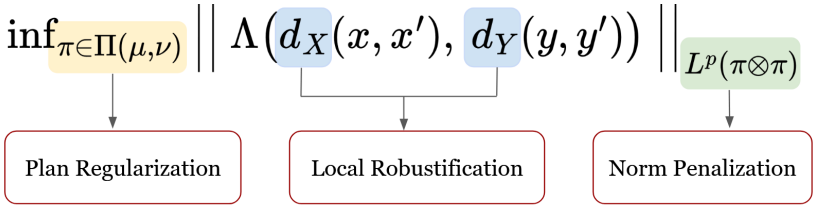


Figure 4.1: Three disjoint approaches leading to outlier-robustness of different degrees in Gromov-Wasserstein formulations. The forthcoming discussion follows the course: Section 4.4.1 (■), Section 4.4.2 (■), and Section 4.4.3 (■).

On top of these existing ambiguities, the participating mm spaces might suffer different types of contamination with varying intensity. As such, the spectrum of contamination models (Wasserstein, Huber's ϵ-contamination, etc.) needs to be kept in mind. Based on the varied demands of cross-domain alignment problems, we identify *three* solutions to the robustification problem in GW. The *primary* and most immediate way is to arrest the extreme distortions, given pairwise observations $(x, y), (x', y')$. In the process, we introduce Tukey's and Huber's relaxation into the GW metric (Section 4.4.1). The *second* solution stems from robust surrogates to the metrics d_X, d_Y that limit fluctuations at their nascency while calculating pairwise distances (Section 4.4.2). Based on the nature of the mm spaces, this approach may propose both structural and optimization-based robustness. In search of robust translation maps, the *third* method advocates relaxing the optimization itself by regularizing the set of plans $\Pi(\mu, \nu)$ (Section 4.4.3). Besides introducing them, the following discussion demonstrates that these three seemingly distinct solutions are indeed related. The idea that binds them is the very key to penalization: reallocating mass from outliers to inliers. By introducing more effective ways to reallocate, we observe intriguing relations to robust OT formulations and duality emerging.

4.4.1 Norm Penalization: Towards Huber's Gromov-Wasserstein

The most recognized contamination model in robust statistics (Huber, 1964) assumes the existence of an arbitrary distribution $\mu_c \in \mathcal{P}(\mathcal{X})$ from which outliers originate. Under the same, it is equivalent to tossing a coin with $\epsilon \in (0,1)$ probability in favor of μ_c during each independent draw from μ. In case μ_c admits a density with heavy tails, it implies vastly outlying observations in the sample (see e.g., Figure 4.2(c)). A similar output may occur if the moments corresponding to μ and μ_c differ significantly. In a shape-matching context, this amounts to pronounced disfiguration of shapes. Now, let us recall the definition (4.3),

$$d_{\mathrm{GW}}(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \frac{1}{2}\|\Lambda_{X,Y}\|_{L^p(\pi\otimes\pi)},$$

where $\Lambda_{X,Y} = 2\Lambda_1(d_X, d_Y)$ in particular. In a sample problem, the l_p norm calculating the departure from isometry is sensitive to unusually large (or small) observations. We employ relaxed l_p norms to curb the effects of such outliers in distortion. Considering the computational complexity of the GW formulation, we first invoke the most intuitive way of implementing a relaxation, namely Tukey's relaxation (Clarkson et al., 2019).

Definition 4.1 (Tukey loss function)**.** *Given a threshold $\tau \geq 0$, the p-Tukey loss function for $p \in [1,\infty)$ is defined as*

$$\mathcal{T}_p(x) \coloneqq \begin{cases} |x|^p & \text{if } |x| \leq \tau \\ \tau^p & \text{otherwise.} \end{cases}$$

Observe that, it is polynomially bounded above[3] with degree p. It also induces the corresponding 'norm', $\|f\|_{\mathcal{T}_p(\mu)} = \left(\int_{\mathcal{X}} \mathcal{T}_p(f(x))d\mu(x)\right)^{1/p}$, given $f \in L^p(\mu)$. Though dependent on the parameter τ, it enables one to define

$$\|\Lambda_{X,Y}\|_{\mathcal{T}_p(\pi\otimes\pi)} \coloneqq \left(\int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}} \mathcal{T}_p\big|d_X(x,x') - d_Y(y,y')\big| d\pi\otimes\pi(x,y,x',y')\right)^{\frac{1}{p}}. \tag{4.5}$$

We call the quantity $d_{\mathrm{TGW}}(\mu,\nu) \coloneqq \inf_{\pi\in\Pi(\mu,\nu)} \frac{1}{2}\|\Lambda_{X,Y}\|_{\mathcal{T}_p(\pi\otimes\pi)}$, *Tukey's GW* (TGW, or specifically (p,τ)-TGW). Clearly, it is a lower bound to the corresponding $d_{\mathrm{GW}}(\mu,\nu)$ and $d_{\mathrm{TGW}} \to d_{\mathrm{GW}}$ as $\tau \to \infty$. It also carries some major properties of the original GW distance (Mémoli (2011), Theorem 5.1). The non-negativity and symmetry are obvious, and given $\mathcal{X} \cong \mathcal{Y}$, it becomes 0. Conversely, given a threshold $\tau > 0$, $d_{\mathrm{TGW}}(\mu,\nu) = 0$ implies that the mm spaces $\mathcal{X}$ and $\mathcal{Y}$ are isometric, for $p \in [1,\infty)$. Here, we only define the loss for $p < \infty$ since given $\tau > 1$, the essential norm at $p = \infty$ spoils the thresholding and eventually the robustness. Our goal also lies in avoiding large deviations between TGW from its perturbed GW benchmark, caused solely due to large thresholds. As such, it is not in our interest to

[3] An increasing function $f : \mathbb{R}_+ \to \mathbb{R}_+$ is said to be polynomially bounded above with degree p if $\frac{f(b)}{f(a)} \lesssim \left(\frac{b}{a}\right)^p$.

check the continuity of $\|\cdot\|_{\mathcal{T}_p(\pi)}$ w.r.t. the weak convergence of π. Hence, a feasible coupling $\pi \in \Pi(\mu, \nu)$ that realizes the infimum only exists for $p \in [1, \infty)$ (Mémoli (2011), Corollary 10.1). Proving the same only requires modifying the first part of the proof of Corollary 10.1 as follows: Define $f : [0, M] \to \mathbb{R}_+$ as $t \mapsto t^p$ if $t \leq \tau$, and τ^p if $\tau < t \leq M$, which also becomes Lipschitz with constant pM^{p-1}. Observe that, $\tau \geq M = \text{diam}(\mathcal{X}) \vee \text{diam}(\mathcal{Y})$ implies the exact function as in Mémoli (2011). Also, the triangle inequality exists owing to the same property for $\|\cdot\|_{\mathcal{T}_p}$.

Proposition 4.1 (Metric properties)**.** *Let $(\mathcal{X}, d_X, \mu_X)$, $(\mathcal{Y}, d_Y, \mu_Y)$ and $(\mathcal{Z}, d_Z, \mu_Z)$ denote arbitrary Polish mm spaces. Then,*

(i) (Triangle inequality)

$$d_{TGW}(X, Y) \leq d_{TGW}(X, Z) + d_{TGW}(Z, Y).$$

(ii) Given $\gamma \in \Pi(\mu_X, \mu_Y)$, for all $0 \leq p \leq p' < \infty$

$$\|\Lambda_{X,Y}\|_{\mathcal{T}_p(\gamma \otimes \gamma)} \leq \|\Lambda_{X,Y}\|_{\mathcal{T}_{p'}(\gamma \otimes \gamma)}.$$

(iii) For $\tau' > \tau \geq 0$, we have (p, τ)-TGW $\leq (p, \tau')$-TGW.

The properties altogether make d_{TGW} a pseudometric over the collection of isomorphism classes of mm spaces. It also limits the corruption due to Wasserstein and Huber's ϵ-contamination. For simplicity, let us assume only one of the distributions is contaminated, say μ. As such, one now has observations from $\mu' = (1-\epsilon)\mu + \epsilon\mu_c$ instead, where $\mu_c \in \mathcal{P}_p(\mathcal{X})$. Under this setup, the following result gives the extent to which the population-level loss can propagate.

Theorem 4.1. *Given that the two distributions μ, ν belong to the same mm space (i.e. they are namely $(\mathcal{X}, d_X, \mu)$ and $(\mathcal{X}, d_X, \nu)$), if μ suffers Huber's ϵ-contamination, we have*

$$d_{TGW}(\mu', \nu) \leq \tau\epsilon^{\frac{1}{p}} + W_{\mathcal{T}_p}(\mu, \nu),$$

where $W_{\mathcal{T}_p} \coloneqq \inf_\Pi \|d_X\|_{\mathcal{T}_p}$ is the OT_{d_X} distance under the p-Tukey norm.

The result also implies that TGW can pose as a provably robust estimate to GW, i.e., $d_{\text{TGW}}(\mu', \nu) - d_{\text{GW}}(\mu, \nu) \leq \tau\epsilon^{1/p}$, if the distributions are supported on the same space. This observation is instrumental in robust shape-matching and generation problems. Theorem 4.1 also has interesting consequences under specific assumptions on the contamination model. For example, if there exists $k \geq 1$ such that $W_p(\mu, \mu_c) = kW_p(\mu, \nu)$ (Balaji et al., 2020), we have

$$d_{\text{TGW}}(\mu', \nu) \leq (1 + k\epsilon^{\frac{1}{p}})W_p(\mu, \nu),$$

due to the trivial upper bound on $W_{\mathcal{T}_p}$ by the p-Wasserstein distance. On the other hand, *Wasserstein contamination* (Zhu et al., 2022) only assumes that $W_p(\mu', \mu) < \epsilon$, regardless of the nature of the adversary. This typically occurs if all observations are perturbed by small degrees. In such a case, the inequality, as in Proposition 4.1, adapts to give the following upper bound $< \tau \wedge \epsilon + W_{p,\lambda}(\mu, \nu)$.

Remark 4.1 (Resilience under TGW)**.** *Theorem 4.1 (see proof, Appendix) also enables one to discuss the 'resilience' of a distribution μ under d_{TGW}. $\mu \in \mathcal{P}(\mathcal{X})$ is said to be (ρ, ε)-resilient w.r.t. the divergence d if $\forall \tilde{\mu} \in \mathcal{P}(\mathcal{X})$ such that $\tilde{\mu} \leq \frac{1}{1-\varepsilon}\mu$, we have $d(\mu, \tilde{\mu}) \leq \rho$, where $0 \leq \varepsilon < 1$ and $\rho \geq 0$. It boils down to checking the maximum change in $W_{\mathcal{T}_p}$ if a ε-fraction of mass under μ is deleted and renormalized to form $\tilde{\mu} \leq \frac{1}{1-\varepsilon}\mu$. Nietert et al. (2023) show that $|\mathbb{E}_{\tilde{\mu}}[d_X(Y, x_0)^p] - \mathbb{E}_{\mu}[d_X(Z, x_0)^p]| \leq \rho$ implies resilience under W_p, given $x_0 \in \mathcal{X}$ (Lemma 11). In the process, it is sufficient to assume that μ has a finite p-moment. Observe that,*

$$
\begin{aligned}
&|\mathbb{E}_{\tilde{\mu}}[\mathcal{T}_p(d_X(Y, x_0))] - \mathbb{E}_{\mu}[\mathcal{T}_p(d_X(Z, x_0))]| \\
&\leq \mathbb{E}_{Y\sim\tilde{\mu}, Z\sim\mu}|\mathcal{T}_p(d_X(Y, x_0)) - \mathcal{T}_p(d_X(Z, x_0))| \\
&\leq \mathbb{E}_{Y\sim\tilde{\mu}, Z\sim\mu}|d_X(Y, x_0)^p - d_X(Z, x_0)^p| \\
&\leq \sqrt{Var_{\tilde{\mu}}[d_X(Y, x_0)^p]} + \sqrt{Var_{\mu}[d_X(Z, x_0)^p]} + |\mathbb{E}_{\tilde{\mu}}[d_X(Y, x_0)^p] - \mathbb{E}_{\mu}[d_X(Z, x_0)^p]|,
\end{aligned}
$$

where the first inequality is due to Jensen's inequality. As such, since the variances are finite, the mean resilience also implies the same under $\mathcal{T}_p$. However, in case $\tilde{\mu}$ results in a vastly distinct variance, the associated resilience bound on $W_{\mathcal{T}_p}$ becomes weak.

Corollary 4.1 (Nietert et al. (2023))**.** *Given $Z \sim \mu$ and $x_0 \in \mathcal{X}$, let $\mathbb{E}_{\mu}[d_X(Z, x_0)^p] \leq \sigma^p$ for some $\sigma \geq 0$. If $\mathcal{T}_p(d_X(Z, x_0))$ is (ρ, ε)-resilient in mean, then μ is $\Big(2\Big((\rho^{\frac{1}{p}} + \varepsilon^{\frac{1}{p}}(\sigma \wedge \tau)) \wedge \varepsilon^{\frac{1}{p}}\tau\Big), \varepsilon\Big)$-resilient w.r.t. $W_{\mathcal{T}_p}$.*

Observe that the bound is non-trivial only when $\sigma \leq \tau$. While a user-defined τ ensures resilience for distributions having thicker tails, the result hints towards distributions (μ) that imply sharper resilience bounds. One immediate example is the class of sub-Gaussian distributions.

Remark 4.2 (Lower bound to ROBOT)**.** *TGW has a surprising relation to existing robust efforts in OT, following from Theorem 4.1. Given a transportation cost $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, Mukherjee et al. (2021) (formulation 2) define the λ-ROBOT distance between $\mu, \nu \in \mathcal{P}(\mathcal{X})$ as $OT_{c_\lambda}(\mu, \nu)$, where $c_\lambda(x, y) := l_{2\lambda}(c(x, y)) = \min\{c(x, y), 2\lambda\}$. Specifically for $c = d_X$, we write $W_{1,2\lambda}(\mu, \nu)$. It metrizes the underlying class of distributions and was shown earlier to lead to faster cost computation in tasks such as image retrieval (Pele and Werman, 2009). On the other hand, if $p = 1$, the Tukey loss boils down to the exact functional $l_\lambda(x) = \min\{x, \tau\}$,*

$x > 0$. As such, for any coupling π, by assuming $\tau = 2\lambda$ we have

$$\|\min\{2\Lambda_1, \tau\}\|_{L^1(\pi\otimes\pi)} \leq \big\|\min\{d_X(x,y) + d_X(x',y'), \tau\}\big\|_{L^1(\pi\otimes\pi)} \leq 2\|\min\{d_X(x,y), \tau\}\|_{L^1(\pi)}.$$

Taking infimum over $\Pi(\mu,\nu)$ we conclude $(1, 2\lambda)$-$d_{TGW} \leq W_{1,2\lambda}$. Observe that such a truncation gives an immediate solution to make WAE robust against outliers.

Remark 4.3 (Concentration)**.** *In reality, often outlying observations find their way into the pool of samples, which is unlike drawing i.i.d. replicates from a contaminated distribution $\mu' = (1-\epsilon)\mu + \epsilon\mu_c$. Rather, in a set of samples $\{x_i\}_{i=1}^m$, we are left with $|\mathcal{I}|$ i.i.d. observations from μ and the rest, $|\mathcal{O}| \coloneqq n - |\mathcal{I}|$ drawn independently from adversaries. If the outliers also follow μ_c identically and $|\mathcal{O}| = m\epsilon$, it becomes equivalent to Huber's contamination regime in an empirical setup. Lecué and Lerasle (2020) call this the $\mathcal{O} \cup \mathcal{I}$ framework. This is crucial since it allows one to comment on the concentration of the empirical d_{TGW}. In such a setup, given samples $\{(x_i, y_j)\}^{m,n}$, we get for $p = 2$*

$$\big|d^2_{TGW}(\hat{\mu}_m, \hat{\nu}_n) - d^2_{GW}(\hat{\mu}^{\mathcal{I}}_m, \hat{\nu}^{\mathcal{I}}_n)\big| \leq \left|\frac{|\mathcal{I}^X||\mathcal{I}^Y|}{mn} - 1\right| M^4 + \tau\Big(\frac{|\mathcal{O}^X|}{m} + \frac{|\mathcal{O}^Y|}{n} - \frac{|\mathcal{O}^X||\mathcal{O}^Y|}{mn}\Big),$$

where $\hat{\mu}_m, \hat{\nu}_n$ are the usual empirical distributions based on $\{(x_i, y_j)\}^{m,n}$, and $\hat{\mu}^{\mathcal{I}}_m \coloneqq \big|\mathcal{I}^X\big|^{-1} \sum_{i\in\mathcal{I}^X} \delta_{x_i}$ is the same based on inliers. The same goes for $\hat{\nu}^{\mathcal{I}}_n$ in the other space. Moreover, $\mathcal{O}^X$ and $\mathcal{O}^Y$ denote the set of outliers. As such, given $\big|\mathcal{O}^X\big| \vee \big|\mathcal{O}^Y\big| = o(m \wedge n)$, obtaining TGW may be done alternatively by calculating GW solely based on the inliers. The associated error becomes arbitrarily small. Moreover,

$$\big|\mathbb{E}\big[d^2_{TGW}(\hat{\mu}_m, \hat{\nu}_n)\big] - d^2_{GW}(\mu,\nu)\big| \leq \mathbb{E}\big|d^2_{TGW}(\hat{\mu}_m, \hat{\nu}_n) - d^2_{GW}(\hat{\mu}^{\mathcal{I}}_m, \hat{\nu}^{\mathcal{I}}_n)\big| + \mathcal{E}_{\mathcal{I}}, \tag{4.6}$$

where, (4.6) utilizes Jensen's inequality and the estimation error $\mathcal{E}_{\mathcal{I}}$ is bounded from above as the following due to Zhang et al. (2024), Theorem 3

$$\begin{aligned}\mathcal{E}_{\mathcal{I}} &\coloneqq \big|\mathbb{E}[d^2_{GW}(\hat{\mu}^{\mathcal{I}}_m, \hat{\nu}^{\mathcal{I}}_n)] - d^2_{GW}(\mu,\nu)\big| \\ &\lesssim \frac{M^4}{\sqrt{|\mathcal{I}^X| \wedge |\mathcal{I}^Y|}} + (1 + M^4) \bigvee_{\mathcal{I}^X, \mathcal{I}^Y} |\mathcal{I}|^{-\frac{2}{(d\wedge d')\vee 4}} (\log|\mathcal{I}|)^{\mathbb{1}\{d\wedge d' = 4\}}.\end{aligned}$$

As such, in the $\mathcal{O} \cup \mathcal{I}$ framework, LRGW estimates the squared GW distance between inlying distributions asymptotically unbiasedly, and hence robustly.

The theoretical richness of TGW gives us a solid foundation to search for better approximations using data-dependent thresholding. It is also quite intuitive that a misspecified $\tau > 0$ may lead to heavier penalization than required, generating a large deviation from GW. In practice, even minute fine-tuning errors lead to a significant loss in tail information. A

smoother thresholding may achieve a nearer robust approximation without sacrificing favorable properties. This leads us to the Huber 'norm'.

Definition 4.2 (Huber loss function)**.** *The Huber loss with threshold $\tau > 0$ is defined as*

$$\mathcal{H}(x) \coloneqq \begin{cases} x^2/2\tau & \text{if } |x| \leq \tau \\ |x| - \tau/2 & \text{otherwise,} \end{cases}$$

which induces the corresponding norm, $\|f\|_{\mathcal{H}(\mu)} = \left(\int_{\mathcal{X}} \mathcal{H}(f(x))d\mu(x)\right)^{1/2}$.

Observe that, $\mathcal{H}(\cdot)$ is continuously differentiable and given $\tau \simeq 0$, closely approximates the l_1 loss. The robust penalization also becomes data-dependent, making it essential in robust M-estimation and regression (Loh, 2017). Following our previous formulation, for $(\mu, \nu) \in \mathcal{P}_2(\mathcal{X}) \times \mathcal{P}_2(\mathcal{Y})$, we define $d_{\text{HGW}}(\mu, \nu) \coloneqq \inf_{\pi \in \Pi(\mu,\nu)} \frac{1}{2}\|\Lambda_{X,Y}\|_{\mathcal{H}(\pi \otimes \pi)}$, namely the *Huber's GW* (HGW). Addressing the robustness of GW for $p = 2$ in particular, $\|\Lambda_{X,Y}\|_{\mathcal{H}}$ does not admit a monotonic property. However, based on the fact that $\mathcal{H}^{\frac{1}{2}}$ is subadditive (Clarkson and Woodruff, 2014), one can recover a Minkowski-type inequality as in TGW. Moreover, HGW poses as a robust estimate of the corresponding GW value as it follows a property similar to Theorem 4.1. The result involves defining the Huber version of the modified Wasserstein 'distance' $W_{\mathcal{H}} \coloneqq \inf_{\Pi} \|d_X\|_{\mathcal{H}}$. Consequently, it promotes resilience to a distribution under it upon mass truncation.

To empirically demonstrate HGW's robustness, we devote the rest of the section to building an algorithm to solve a sample HGW. Given m and $n \in \mathbb{N}_+$ i.i.d. samples from μ and ν respectively, let us denote the two pairwise distance matrics (based on d_X and d_Y) as $C^X \in \mathbb{R}_+^{m \times m}$ and $C^Y \in \mathbb{R}_+^{n \times n}$. Then, d^2_{HGW} boils down to solving the familiar non-convex optimization (Peyré et al., 2016)

$$\min_{\pi \in \Pi(\hat{\mu}_m, \hat{\nu}_n)} \sum_{i,i',j,j'} \mathcal{H}(C^X_{ii'} - C^Y_{jj'})\pi_{ij}\pi_{i'j'} = \min_{\pi \in \Pi(\hat{\mu}_m, \hat{\nu}_n)} \langle \mathcal{H}(C^X - C^Y) \circledcirc \pi, \pi \rangle, \tag{4.7}$$

where $\hat{\mu}_m, \hat{\nu}_n$ are empirical distributions or rather simplexes and $\Pi \in \mathbb{R}_+^{m \times n}$. To adapt to the Sinkhorn scaling framework (Cuturi, 2013), we additionally impose an entropic regularization $d_{\text{KL}}(\pi)$ to (4.7), which at the k-th iteration calculates $d_{\text{KL}}(\pi|\pi^k) = \langle \pi, \log \pi - \log \pi^k \rangle$. The resultant Huber's EGW formulation follows the Algorithm 1. It is immediately beneficial for computing a robust loss, compared to Unbalanced GW (UGW) or Partial GW (PGW), since it results in marginal distributions and computationally scales with the EGW ($O(m^2n^2)$) exactly.

While we present a simple working algorithm[4], HEGW also adapts to lower-complexity approximations. In Algorithm 1, computing the cost $\mathcal{C}(\pi)$ alone incurs the high complexity

[4]This allows seamless integration into existing libraries: `https://pythonot.github.io/`

Algorithm 1 Huber's Entropic Gromov-Wasserstein

Input: Initialised distributions p, q, regularization parameter ε, number of inner and outer iterations N_2, N_1.
Output: HEGW

Compute pairwise distance matrices C^X, C^Y
Initialise $\pi^{(0)} = pq^T$, $a^{(0)} = 1_m$, and $b^{(0)} = 1_n$
for $i \in \{0, 1, \cdots, N_1 - 1\}$ **do**
  $\mathscr{C}(\pi^{(i)}) \leftarrow \mathcal{H}(C^X - C^Y) \circledcirc \pi^{(i)}$ ▷ Compute cost matrix
  $K^{(i)} \leftarrow \exp\Big\{\big(-\frac{\mathscr{C}(\pi^{(i)})}{\varepsilon}\big)\Big\} \odot \pi^{(i)}$ ▷ Compute kernel
  for $j \in \{0, 1, \cdots, N_2 - 1\}$ **do**
    $\{a^{(j+1)}, b^{(j+1)}\} \leftarrow \{p \oslash (K^{(i)} b^{(j)}), q \oslash ({K^{(i)}}^T a^{(j+1)})\}$ ▷ Sinkhorn scaling
  end for
  $\pi^{(i+1)} \leftarrow \mathrm{diag}(a^{(N_2)}) K^{(i)} \mathrm{diag}(b^{(N_2)})$
end for
Return $\big\langle \mathscr{C}(\pi^{(N_1)}), \pi^{(N_1)} \big\rangle$

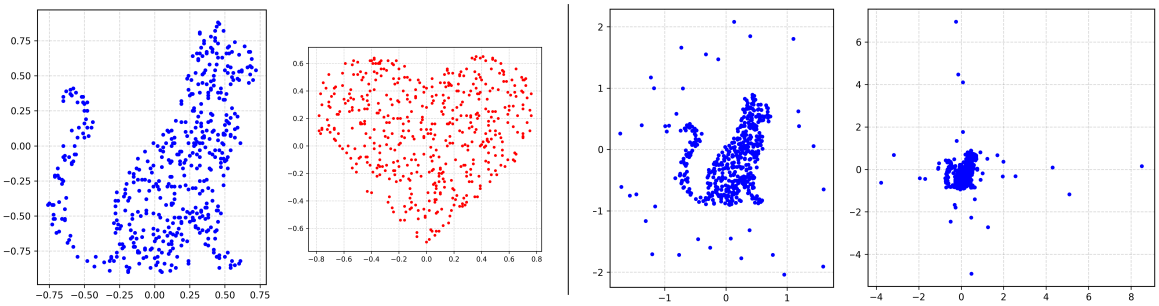
$O(m^2 n^2)$. However, we can write $\mathcal{H}(a-b) = f_1(a) + f_2(b) - h_1(a)h_2(b)$, where given $|a-b| \leq \tau$, $f_1(a) = a^2/2\tau, f_2(b) = b^2/2\tau, h_1(a) = a/\tau, h_2(b) = b$ and if $a - b > \tau$, we have $f_1(a) = a, f_2(b) = -b, h_1(a) = \tau, h_2(b) = 1/2$. As such, the cost computation can be eased down to $O(m^2 n + mn^2)$ (Peyré et al. (2016), Proposition 1). This is the best complexity achievable if μ, ν are not sliced first or no additional constraint on the matrices satisfying lower ranks is imposed. However, if along with the robust penalization, we identify a set of indices $\mathcal{S} = \{(i, j)\}$ with $|\mathcal{S}| = s$, such that

$$\tilde{\mathcal{H}}(C^X_{ii'} - C^Y_{jj'}) = \begin{cases} \mathcal{H}(C^X_{ii'} - C^Y_{jj'}) & \text{if } (i', j') \in \mathcal{S} \\ 0 & otherwise, \end{cases}$$

where $(i, j) \in \mathcal{S}$, the modified HEGW problem can be solved with accompanying complexity $O(mn + s^2)$ (Li et al., 2023). While this results in a truncated plan (π), it effectively thwarts outliers in graphs.

Experiment: Shape Matching with Outliers

We deploy HGW for robust 2D shape matching based on point cloud data (Mroueh and Rigotti, 2020). Observe that the underlying mm spaces are essentially $(\mathbb{R}^2, \|\cdot\|_2)$, endowed with measures $\hat{\mu}_m$ and $\hat{\nu}_n$ respectively. Given two shapes (e.g. cat and heart), we identify one as the target and the other as the source. The contamination regime we follow is the following: for $\alpha \in (0, 1)$, we randomly sample $m\alpha$ observations from the source point cloud and replace them with replicates from an adversary μ_c (e.g., standard bi-variate Cauchy). For comparison, we use the vanilla GW, FGW (Vayer et al., 2020), PGW (Chapel et al., 2020),

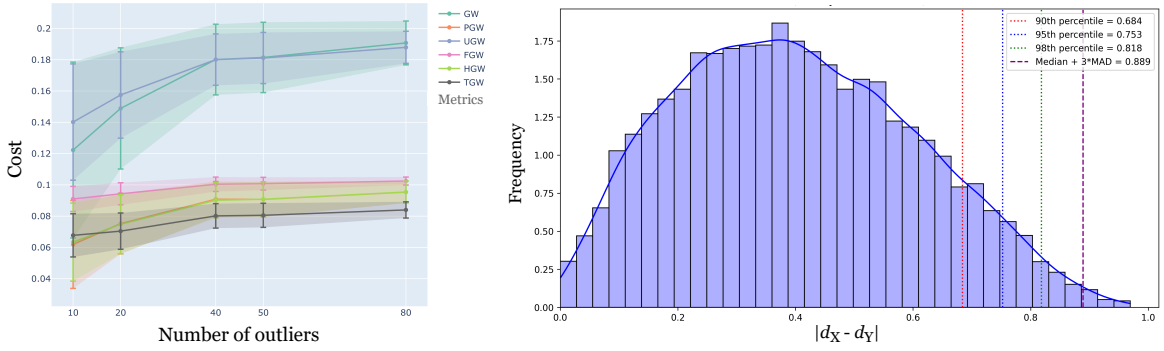


(a) Source (cat) and target (heart) shapes (b) Gaussian outliers (c) Cauchy outliers

Figure 4.2: (a) Point clouds ($m = n = 500$) corresponding to shapes of cat (source) and heart (target). Contaminated source with 20 outliers drawn independently from a standard (b) bivariate Gaussian and (c) bivariate Cauchy.

and UGW (Séjourné et al., 2021) as baselines under $p = 2$. For the unbalanced methods, we allocate unit mass to each point, and for the rest, we normalize. Each method, at each level of contamination, is repeated 100 times to cover for any variation generated due to optimization. The parameters for each distance are selected based on the recommendations by the respective authors, e.g. in the case of FGW, the mixing coefficient of GW and the OT is kept at 0.5. The regularization parameter for PGW is taken as 0.001. We find that only such small values, chosen judiciously, can strike a balance between adequate penalization and a low enough value of the corresponding loss. In TGW (and HGW), the method's accuracy hinges on the selection of τ. Very low values lead to over-penalization and large deviations from the actual robust benchmark. In our study, we devise a data-driven scheme for selecting τ. Given observations $\{(x_i, y_j)\}^{m,n} \sim \mu \otimes \nu$, we scrutinize the distribution of sample distortion values (say, $J_{X,Y}$) $\mid \|x_i - x_{i'}\|^2 - \|y_j - y_{j'}\|^2 \mid$. An immediate estimate for a threshold that trims outlying $J_{X,Y}$ values is a higher percentile, e.g., $98\%, 95\%$, which we use as a reference. Our choice of an appropriate τ becomes $\tilde{m} + 3\tilde{\sigma}$, where $\tilde{m}$ and $\tilde{\sigma}$ are the median and mean absolute deviation about median of $J_{X,Y}$. Ideally, for a standard folded Normal distribution, the value turns out ≈ 2.04 (see Appendix). The method enables a dynamic parameter selection that adjusts according to the proportion of outliers. We present a detailed discussion in Appendix.

Based on the proposed scheme, the value of τ is chosen dynamically at each level α for HGW. As a reference, we use the 95-percentile of $J_{X,Y}$ in TGW. The immediate observation is that TGW remains the most stable under increasing contamination. On the other hand, HGW exhibits performance comparable to that of partial mass allocation, as in PGW. Since the threshold only penalizes extreme values of $J_{X,Y}$, pairwise distances between outliers that are similar to those between inliers contribute to the overall loss. This implies a minute increase in HGW. The effect is much pronounced in Gaussian outliers (see Appendix). The stability of PGW is intuitive since its optimal plan ignores the outliers altogether. Remark-



(a) (b)

Figure 4.3: (a) Average loss values under increasing proportion of bi-variate Cauchy outliers $(0.02, 0.04, 0.08, 0.1, 0.16)$ in the source domain. (b) Empirical distribution of deviations between pairwise distances under 80 Cauchy outliers. Realized 95-percentile and $\tilde{m} + 3\tilde{\sigma}$ are 0.753 and 0.889 respectively. The stability of TGW (and HGW) under increasing corruption corroborates the concentration (Remark 4.3).

ably, HGW simulates the same effect without altering the plan. FGW (with a mixing ratio 0.5) shows elevated fluctuation as its OT component is also vulnerable to contamination. The instability of UGW might stem from mass-splitting and the partial dependence of its plan on outliers (as also pointed out by Bai et al. (2024)). However, despite relying on a full plan connecting outliers to points in the target shape, HGW results in robust estimates of the corresponding loss. In Appendix, we show the results corresponding to Gaussian contamination and compare the effects due to varying τ.

4.4.2 Local Robustification

Techniques that make the distortion $\Lambda(d_X(x, x'), d_Y(y, y'))$, given $(x, y), (x', y') \in \mu \otimes \nu$, robust to outliers, offer a robust estimate to the extent of deviation from isometry. It is equivalent to selecting a different measurement to compare the two local geometries, however perturbed; for example, TGW. In other words, the penalization applies over the discrepancy in pairwise distances, rather than $d_X(x, x')$ and $d_Y(y, y')$ themselves. This may suffer from inefficiency in information retention, as given only an outlying observation $x' \in \mathcal{X}$, it depreciates the contribution of a 'clean' $d_Y(y, y')$. An earlier-stage robustification of the distances solves this issue. In search of nearer GW approximates, we turn to robust surrogates of d_X and d_Y.

For typical choices of metric spaces (for example, $(\mathbb{R}^d, \|\cdot\|)$), it is quite straightforward to retain metric properties of d_X under thresholding (e.g., Tukey) or Winsorization. For example, the *truncated* surrogate $l_\lambda(d_X) = \min\{d_X, \lambda\}$, $\lambda \geq 0$ satisfies non-negativity, symmetry and the triangle inequality. Based on the fact that $\mathcal{T}_2(a - b) \geq |l_\lambda(a) - l_\lambda(b)|^2$, for $a, b \geq 0$, it immediately improves the corresponding robust GW formulation, evoking a lower bound to Tukey-type distances.

Lemma 4.1. *Given $(\mu,\nu)\in\mathcal{P}_4(\mathbb{R}^d)\times\mathcal{P}_4(\mathbb{R}^{d'})$, the $(2,\lambda)$-TGW between them satisfies*

$$d_{TGW}(\mu,\nu)\geq\frac{1}{2}\Big(\inf_{\pi\in\Pi(\mu,\nu)}\int\int\Big|l_\lambda\Big(\big\|x-x'\big\|^2\Big)-l_\lambda\Big(\big\|y-y'\big\|^2\Big)\Big|^2d\pi\otimes\pi\Big)^{\frac{1}{2}}.$$

We call such formulations, as in the lower bound, *locally robust GW* ((p,λ)-LRGW, in general). Infima of the corresponding optimization are always realized at relaxed optimal couplings (see Appendix). The modification gives greater control over the extent of robustification based on distinct choices of $\lambda,\lambda'\geq 0$ for the two mm spaces. Based on its formulation, LRGW is particularly adept at providing robustness in the face of *geometric* outliers. In other words, if there are outliers in a pool of samples, lying at a distance with identifiable separation from the ambient points, LRGW proves to be significantly more effective compared to conventional methods. LRGW also follows most metric properties of GW, particularly non-negativity, symmetry, and triangle inequality. The proofs become similar to showing the same for GW under altered mm spaces of the form $(\mathcal{X},l_\lambda(d_X),\mu)$, $\lambda>0$. For completeness, we mention some properties of LRGW, including its dependence on the threshold λ.

Proposition 4.2 (Properties)**.** *Given Polish mm spaces $(\mathcal{X},d_X,\mu)$, $(\mathcal{Y},d_Y,\nu)$; and $p\in[0,\infty]$*

(i) for $\lambda'>\lambda\geq 0$, we have (p,λ)-LRGW$(\mu,\nu)\leq(p,\lambda')$-LRGW$(\mu,\nu)$. In fact,

$$(p,\lambda')\text{-LRGW}(\mu,\nu)-(p,\lambda)\text{-LRGW}(\mu,\nu)\leq\inf_{\pi\in\Pi^\lambda(\mu,\nu)}\big\|l_{\lambda'}(\bar{l}_\lambda(d_X))-l_{\lambda'}(\bar{l}_\lambda(d_Y))\big\|_{L^p(\pi\times\pi)},$$

where $\bar{l}_\lambda(x)=\max\{x,\lambda\}$ and Π^λ is the set of couplings optimal for (p,λ)-LRGW.

(ii) (p,∞)-LRGW$(\mu,\nu)=p$-GW(μ,ν).

(iii) If $p\geq q\geq 1$, then (p,λ)-LRGW$(\mu,\nu)\leq(\lambda\wedge 2M)^{1-\frac{q}{p}}[(q,\lambda)\text{-LRGW}(\mu,\nu)]^{\frac{q}{p}}$.

Observe that, Proposition 4.2*(ii)* rather holds for any $\lambda\geq M=\operatorname{diam}(\mathcal{X})\vee\operatorname{diam}(\mathcal{Y})$, given $\operatorname{diam}(\mathcal{X})=\max_{x,x'\in\mathcal{X}}d_X(x,x')$, which however may become arbitrarily large in the presence of outliers in a sample problem. As a consequence of Proposition 4.2*(ii)*, $\mathcal{X}\cong\mathcal{Y}$ implies that the associated LRGW nullifies. However, the converse does not hold necessarily since, $l_\lambda(d_X(x,x'))=l_\lambda(d_Y(y,y'))$ a.s. does not imply $d_X(x,x')=d_Y(y,y')$ a.s. While this formulation sacrifices non-degeneracy, it preserves geometric sensitivity under appropriately tuned λ. It delineates an estimated support of the distribution of distances based on inlying observations[5]. Though finer than TGW, such a filtration allows distances corresponding to outlying samples within λ-radius to each other to pass through. This hints towards addressing contamination due to distribution shifts and mass reallocation as a result. The proof of Proposition 4.2 involves showing that the infimum of the underlying optimization is always

[5]Given $x\in\mathcal{X}$, define the map $u_x^\lambda:\mathcal{X}\to[0,\lambda)$ by $u_x^\lambda(x')=l_\lambda(d_X(x,x'))$. Then, $u_{x\#}^\lambda\mu\in\mathcal{P}([0,\lambda))$ denotes the distribution of distances supported on the trimmed interval.

realized at a relaxed optimal coupling. In other words, given a threshold λ, there exists $\pi^\lambda \in \Pi(\mu,\nu)$ such that $d_{\text{LRGW}}(\mu,\nu) = \|l_\lambda(d_X) - l_\lambda(d_Y)\|_{L^p(\pi^\lambda \otimes \pi^\lambda)}$. Now, let us consider two couplings $\pi^\lambda, \pi^{\lambda'}$ that are optimal for the LRGW problem under thresholds λ and λ' respectively, where $\lambda' > \lambda$ and $p = 1$. Letting $\lambda' - \lambda \leq \delta$ for some $\delta > 0$, we have

$$\begin{aligned}
\big|\lambda'\text{-LRGW} - \lambda\text{-LRGW}\big| &= \left| \int |l_{\lambda'}(d_X) - l_{\lambda'}(d_Y)| d\pi^{\lambda'} \otimes \pi^{\lambda'} - \int |l_\lambda(d_X) - l_\lambda(d_Y)| d\pi^\lambda \otimes \pi^\lambda \right| \\
&\leq \left| \int |l_{\lambda'}(d_X) - l_{\lambda'}(d_Y)| d\pi^{\lambda} \otimes \pi^{\lambda} - \int |l_\lambda(d_X) - l_\lambda(d_Y)| d\pi^\lambda \otimes \pi^\lambda \right| \qquad (4.8) \\
&= \left| \int \big(|l_{\lambda'}(d_X) - l_{\lambda'}(d_Y)| - |l_\lambda(d_X) - l_\lambda(d_Y)| \big) d\pi^\lambda \otimes \pi^\lambda \right| \\
&\leq \lambda' - \lambda \leq \delta.
\end{aligned}$$

Since the choice of δ is arbitrary, LRGW is continuous in λ, given $p = 1$. The inequality (4.8) is due to π^λ not necessarily being optimal (suboptimal) for λ'-LRGW. This implies that $\lim_{\lambda\to\infty} LRGW = GW$. As such, the degeneracy LRGW suffers is *removable*.

Let us now look at some invariants (Chowdhury and Mémoli, 2019) of LRGW— appearing as lower bounds— onto the real line.

Proposition 4.3 (Hierarchy of lower bounds)**.** *Let us denote by $C : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ the cost matrix $C(x,y) \coloneqq W_p(u_x^\lambda(\cdot)_\# \mu, u_y^\lambda(\cdot)_\# \nu)$. Then, for $p \in [1,\infty]$, the following holds*

$$\begin{aligned}
d_{LRGW}(\mu,\nu) &\geq \inf_{\pi\in\Pi(\mu,\nu)} \left\| \mathcal{E}^\lambda_{p,\mathcal{X},\mathcal{Y}} \right\|_{L^p(\pi)} = \inf_{\pi\in\Pi(\mu,\nu)} \|C\|_{L^p(\pi)} \qquad \text{(III)} \\
&\geq \big| size_p^\lambda(\mathcal{X}) - size_p^\lambda(\mathcal{Y}) \big|, \qquad \text{(I)}
\end{aligned}$$

where $\mathcal{E}^\lambda_{p,\mathcal{X},\mathcal{Y}} : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}_+$ denotes the 'truncated outgoing joint eccentricity function', defined as $\mathcal{E}^\lambda_{p,\mathcal{X},\mathcal{Y}}(x,y) = \inf_{\pi\in\Pi(\mu,\nu)} \|l_\lambda[d_X(x,\cdot)] - l_\lambda[d_Y(y,\cdot)]\|_{L^p(\pi)}$ and $size_p^\lambda(\mathcal{X}) \coloneqq \|l_\lambda(d_X)\|_{L^p(\mu\otimes\mu)}$ is the pth root of the p-diameter under the altered metric. Moreover,

$$d_{LRGW}(\mu,\nu) \geq W_p\big(l_\lambda(d_X)_\#(\mu\otimes\mu), l_\lambda(d_Y)_\#(\nu\otimes\nu)\big). \qquad \text{(II)}$$

The lower bounds can be deemed linearizations of the LRGW problem into $\mathbb{R}$. Besides providing a simpler explanation for the locally robust alignment, they can also be computed by solving an OT. For example, $\mathcal{E}^\lambda_{p,\mathcal{X},\mathcal{Y}}(x,y)$, for each $x, y \sim \mu\otimes\nu$ is solvable using a linear program (Sriperumbudur et al., 2012), in turn, implying the same for the lower bound in (III). In fact, it assumes a closed form. On the other hand, size_p^λ is a scalar quantification of the global information present in an mm space under contamination. Being particularly easy to compute, the bound (I) presents a low-level view into the extent of misalignment between them. The bound in inequality (II) can be interpreted as the optimal transportation cost

between the pairwise distance distributions based on $\mathcal{X}$ and $\mathcal{Y}$.

Later in the chapter, we discuss the origin of local penalization (l_λ) in a generalized setup. To motivate, we mention that similar costs are often used to devise robust OT-based divergences metrizing $\mathcal{P}(\mathcal{X})$ (see Remark 4.2). Remarkably, impartially trimmed-W_p due to Czado and Munk (1998) becomes equivalent to trimming the underlying univariate measures (Alvarez-Esteban et al., 2008). In this line, our next result shows that variational representations of LRGW formulations link the alignment problem to certain robust OT costs, relying on trimmed observations instead. Given two mm spaces $(\mathbb{R}^d_{\geq 0}, \|\cdot\|, \mu)$ and $(\mathbb{R}^{d'}_{\geq 0}, \|\cdot\|, \nu)$, let us consider the locally robust inner product GW distance (Mémoli, 2011)

$$d_{\text{LRIGW}}(\mu,\nu;\lambda) \coloneqq \Big(\inf_{\pi\in\Pi(\mu,\nu)} \int\int \big|l_\lambda(\langle x,x'\rangle) - l_\lambda(\langle y,y'\rangle)\big|^2 d\pi\otimes\pi(x,y,x',y')\Big)^{\frac{1}{2}}.$$

Notice that such subspaces remain Polish equipped with the inherited metric (Fristedt and Gray (2013), Chapter 18.1). We may equivalently choose $[0,1]^d$, which is sufficient for the IGW formulation. In both cases, the corresponding problem boils down to assessing the alignment between distributions corresponding to non-negative multivariate random variables. Here, the robustification translates to limiting extreme angular deviation. At $\lambda\to\infty$, the cost circles back to IGW. To state the result, let us first decompose the squared LRIGW cost in the following way:

$$d^2_{\text{LRIGW}}(\mu,\nu;\lambda) = F_1 + F_2,$$

where

$$F_1(\mu,\nu;\lambda) = \int \big|l_\lambda(\langle x,x'\rangle)\big|^2 d\mu\otimes\mu(x,x') + \int \big|l_\lambda(\langle y,y'\rangle)\big|^2 d\nu\otimes\nu(y,y'),$$
$$F_2(\mu,\nu;\lambda) = \inf_{\pi\in\Pi(\mu,\nu)} -2\int l_\lambda(\langle x,x'\rangle)\, l_\lambda(\langle y,y'\rangle) d\pi\otimes\pi.$$

Theorem 4.2 (Locally robust IGW duality)**.** *Given $(\mu,\nu)\in\mathcal{P}_4(\mathbb{R}^d_{\geq 0})\times\mathcal{P}_4(\mathbb{R}^{d'}_{\geq 0})$, define $M^\lambda_{\mu,\nu} \coloneqq \sqrt{M_2(\mu;\lambda)M_2(\nu;\lambda)}$, where for any distribution ρ, $M_2(\rho;\lambda) = \int \|l_\lambda(x)\|^2 d\rho(x)$, $\lambda\geq 0$. Then, there exists an upper bound to F_2, say $\bar{F}_2$, satisfying*

$$\bar{F}_2(\mu,\nu;(d\vee d')\lambda^2) = \inf_{\boldsymbol{A}\in\mathcal{D}_{M^\lambda_{\mu,\nu}}} 8\|\boldsymbol{A}\|^2_F + OT_{c^\lambda_{\boldsymbol{A}}}(\mu,\nu),$$

where $\mathcal{D}_{M^\lambda_{\mu,\nu}} \coloneqq [0, M^\lambda_{\mu,\nu}/2]^{d\times d'}$ and $c^\lambda_{\boldsymbol{A}} : (x,y)\in\mathbb{R}^d_{\geq 0}\times\mathbb{R}^{d'}_{\geq 0} \mapsto -8 l_\lambda(x)^T \boldsymbol{A} l_\lambda(y)$ denotes the cost function deployed under $OT_{c^\lambda_{\boldsymbol{A}}}$.

While it is intuitive that a sample-level robustification is sufficient for LR, Theorem 4.2 additionally provides the deterministic extent to which the associated cost may propagate. This reassures the claim that LRGW turns out to be effective in a contamination regime

where outliers are significantly remote from unperturbed observations under the ambient metric. The complexity related to computing such a bound shrinks down to that of an OT. To observe that, by denoting $(d \vee d')\lambda^2 = \tilde{\lambda}$, let us write

$$\bar{F}_2(\mu, \nu; \tilde{\lambda}) = \inf_{\mathbf{A} \in \mathcal{D}_{M^{\lambda}_{\mu,\nu}}} 8\|\mathbf{A}\|_F^2 + \mathrm{OT}_{c^{\lambda}_{\mathbf{A}}}(\mu, \nu) = \inf_{\mathbf{A} \in \mathcal{D}_{M^{\lambda}_{\mu,\nu}}} U^{\mu,\nu}_{\lambda}(\mathbf{A}).$$

Now, given any other $\tilde{\mu} \in \mathcal{P}_4(\mathbb{R}^d_{\geq 0})$ and $\tilde{\nu} \in \mathcal{P}_4(\mathbb{R}^{d'}_{\geq 0})$, Theorem 4.2 ensures the existence of $\mathbf{A}, \tilde{\mathbf{A}} \in \mathcal{D}_{M^{\lambda}_{\mu,\nu}}$ such that $\bar{F}_2(\mu, \nu; \tilde{\lambda}) = U^{\mu,\nu}_{\lambda}(\mathbf{A})$ and $\bar{F}_2(\tilde{\mu}, \tilde{\nu}; \tilde{\lambda}) = U^{\mu,\nu}_{\lambda}(\tilde{\mathbf{A}})$. As such, by optimality

$$\begin{aligned}
\left|\bar{F}_2(\mu, \nu; \tilde{\lambda}) - \bar{F}_2(\tilde{\mu}, \tilde{\nu}; \tilde{\lambda})\right| &\leq \left|U^{\mu,\nu}_{\lambda}(\mathbf{A}) - U^{\tilde{\mu},\tilde{\nu}}_{\lambda}(\mathbf{A})\right| + \left|U^{\mu,\nu}_{\lambda}(\tilde{\mathbf{A}}) - U^{\tilde{\mu},\tilde{\nu}}_{\lambda}(\tilde{\mathbf{A}})\right| \\
&= \left|\mathrm{OT}_{c^{\lambda}_{\mathbf{A}}}(\mu, \nu) - \mathrm{OT}_{c^{\lambda}_{\mathbf{A}}}(\tilde{\mu}, \tilde{\nu})\right| + \left|\mathrm{OT}_{c^{\lambda}_{\tilde{\mathbf{A}}}}(\mu, \nu) - \mathrm{OT}_{c^{\lambda}_{\tilde{\mathbf{A}}}}(\tilde{\mu}, \tilde{\nu})\right| \\
&\leq 2 \sup_{\mathbf{A} \in \mathcal{D}_{M^{\lambda}_{\mu,\nu}}} \left|\mathrm{OT}_{c^{\lambda}_{\mathbf{A}}}(\mu, \nu) - \mathrm{OT}_{c^{\lambda}_{\mathbf{A}}}(\tilde{\mu}, \tilde{\nu})\right|. \qquad (4.9)
\end{aligned}$$

Since $\mathrm{OT}_{c^{\lambda}_{\mathbf{A}}}$ essentially calculates the transportation cost between truncated observations from μ and ν — which offers finer control over extreme values — LR turns out to achieve arbitrary accuracy in finding a robust surrogate to the GW cost. By plugging in the empirical distributions $\tilde{\mu} = \hat{\mu}_n$ and $\tilde{\nu} = \hat{\nu}_n$ in the stability bound (4.9), one can also comment on the sample complexity of $\bar{F}_2$. First, observe that $M^{\lambda}_{\mu,\nu} \lesssim \lambda^2$ and based on the truncation, the measurable cost $c^{\lambda}_{\mathbf{A}}$ is absolutely bounded. This narrows down the search for dual potentials $(\phi : \mathbb{R}^d_{\geq 0} \to \mathbb{R})$ corresponding to $\mathrm{OT}_{c^{\lambda}_{\mathbf{A}}}$ to the class $\mathcal{F}_{\lambda} \coloneqq \bigcup_{\mathbf{A} \in \mathcal{D}_{M^{\lambda}_{\mu,\nu}}} \mathcal{F}_{\mathbf{A},\lambda}$ such that

$$\mathcal{F}_{\mathbf{A},\lambda} \coloneqq \left\{\phi \mid \exists\, \psi : \mathbb{R}^{d'}_{\geq 0} \to \mathbb{R} \ni \phi = \psi^c; \|\phi\|_{\infty}, \|\psi\|_{\infty} \lesssim \lambda\right\},$$

where ψ^c (the c-transform of $\psi : \mathbb{R}^{d'}_{\geq 0} \to \mathbb{R}$ w.r.t. $c^{\lambda}_{\mathbf{A}}$) is given by $\psi^c = \inf_y c^{\lambda}_{\mathbf{A}}(\cdot, y) - \psi(y)$. As such, we can further upper bound (4.9) to obtain

$$\left|\bar{F}_2(\mu, \nu; \tilde{\lambda}) - \bar{F}_2(\hat{\mu}_n, \hat{\nu}_n; \tilde{\lambda})\right| \leq 4 \sup_{\phi \in \mathcal{F}_{\lambda}} \left|\int \phi \, d[\mu - \hat{\mu}_n]\right| + 4 \sup_{\psi \in \mathcal{F}^c_{\lambda}} \left|\int \psi \, d[\nu - \hat{\nu}_n]\right|,$$

where $\mathcal{F}^c_{\lambda} \coloneqq \bigcup_{\mathbf{A} \in \mathcal{D}_{M^{\lambda}_{\mu,\nu}}} \mathcal{F}^c_{\mathbf{A},\lambda}$ (Groppe and Hundrieser, 2023). This reduces the problem to controlling the two empirical processes over $\mathcal{F}_{\lambda}$ and $\mathcal{F}^c_{\lambda}$. The involvement of raw empirical measures $(\hat{\mu}_n, \hat{\nu}_n)$, susceptible to outliers, makes further upper bounding the individual errors in terms of entropy only feasible in an $\mathcal{O} \cup \mathcal{I}$ framework. We identify this as a potential future work since it does not follow directly by adopting the approach of Ma et al. (2023) into the framework of Zhang et al. (2024) [Theorem 3]. However, the trivial upper bound (see Remark 4.2) along with properties such as resilience (Corollary 4.1) and robust estimation

of corresponding GW (Theorem 4.1) hold as a result of lemma 4.1. Nonetheless, it is not apparent how a penalization as $c_{\mathbf{A}}^{\lambda}$ addresses shifts in mass allocation due to contamination. In this line, to motivate our upcoming formulation, let us first unify the underlying spaces under the general framework of *probabilistic mm spaces*.

Probabilistic metric spaces $(\mathcal{X}, p_X)$ are generalizations of typical metric spaces based on the deterministic 'distance' p_X following a modified triangle inequality

$$T\{p_X(x,x')[0,s], p_X(x',x'')[0,t]\} \leq p_X(x,x'')[0,s+t],$$

where $T : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ (Bauer et al., 2024), for $x, x', x'' \in \mathcal{X}$ and $s, t \geq 0$. In case p_X is replaced by the distribution function, specific choices of T equate them to Menger spaces (e.g., taking $T = \min$ or $\max$) or Wald spaces ($T = *$, convolution) (Schweizer et al., 1960). Defining a measure on this collection $\mathcal{X}$ completes the triplet $(\mathcal{X}, p_X, \mu)$, which we call a probabilistic mm (pmm) space. In our setup, we particularly choose the pair $(\mathcal{X} = \bar{\mathcal{P}}_p(X), W_p)$, where $\bar{\mathcal{P}}_p \subset \mathcal{P}_p$ is the collection of probability measures with full support on $X \subseteq \mathbb{R}^d$ having finite p-moments, $p \geq 1$. This reinforces the notion of generalization based on the fact that given $x, x' \in X$, we have $W_1(\delta_x, \delta_{x'}) = \mathrm{OT}_{d_X}(\delta_x, \delta_{x'}) = d_X(x, x')$. The idea can similarly be extended to an alignment problem between distinct pmm spaces, which brings us back to GW. Observe that, considering a distance $\mathrm{OT}_{l_\lambda(d_X)}$ recovers our earlier LR formulation, as in lemma 4.1. We may alternatively choose the Lévy-Prokhorov (LP) metric $(\hat{\rho})$ to construct our pmm space since it ensures that $(\mathcal{X}, \hat{\rho})$ remains Polish, given that (X, d_X) is Polish (Fristedt and Gray (2013), Chapter 18.7).

Before discussing further generalizations LR motivates, let us demonstrate its efficacy to provide a robust solution. In particular, we put LRGW to the test in finding robust interpolations between the shapes (namely, cat and heart as in Fig. 4.2).

Experiment: Fixed-Support Barycenters

In the context of alignment, it boils down to solving for the Fréchet mean between general mm spaces $\{(\mathcal{X}_i, d_{X_i}, \mu_i)\}_{i=1}^k$, $k \geq 2$ given as

$$\underset{\nu}{\operatorname{argmin}} \sum_{i=1}^{k} \rho_i d_{\mathrm{LRGW}}(\mu_i, \nu), \tag{4.10}$$

where $\rho_i \geq 0$ with $\sum_{i=1}^k \rho_i = 1$. Since our setup demands an interpolated shape as the solution[6], the barycenter problem (4.10) only involves optimizing over a suitable measure supported on $\mathbb{R}^2$. On the other hand, since $k = 2$, a free-support solution to (4.10) leads to the optimal mm space $(\mathcal{X} \times \mathcal{Y}, \rho_1 l_\lambda(d_X) + \rho_2 l_\lambda(d_Y))$, where the associated measure corresponds to a

[6]This inherently specifies the intermediate metric space as $(\mathbb{R}^2, l_\lambda \|\cdot\|_2)$.

geodesic in the GW space (Sturm, 2023). Recall that, unlike partial alignment techniques (Bai et al., 2024), LRGW maintains a full plan. As such, solving (4.10) becomes equivalent to finding a GW barycenter due to Peyré et al. (2016) under truncated metrics. In other words, the optimization boils down to $\operatorname{argmin}_{C,\pi_i} \sum_{i=1}^{k} \rho_i \langle |C_\lambda^{X_i} - C| \circledcirc \pi_i, \pi_i \rangle$, where $\pi_i \in \Pi(\mu_i, \nu)$ and $C_\lambda^{X_i}$ denotes the matrix of truncated pairwise distances. The simultaneous minimization over C and $\{\pi_i\}_1^k$ follows a block-coordinate descent, where the latter uses a Sinkhorn-based projected gradient descent to reach the optima. For details regarding the parameter setting, we refer the reader to the code repository. Updations corresponding to C have a closed form under the L_2 distortion, which eventually produces the visualization of the barycenters upon multidimensional scaling (MDS) (Flamary et al., 2021).

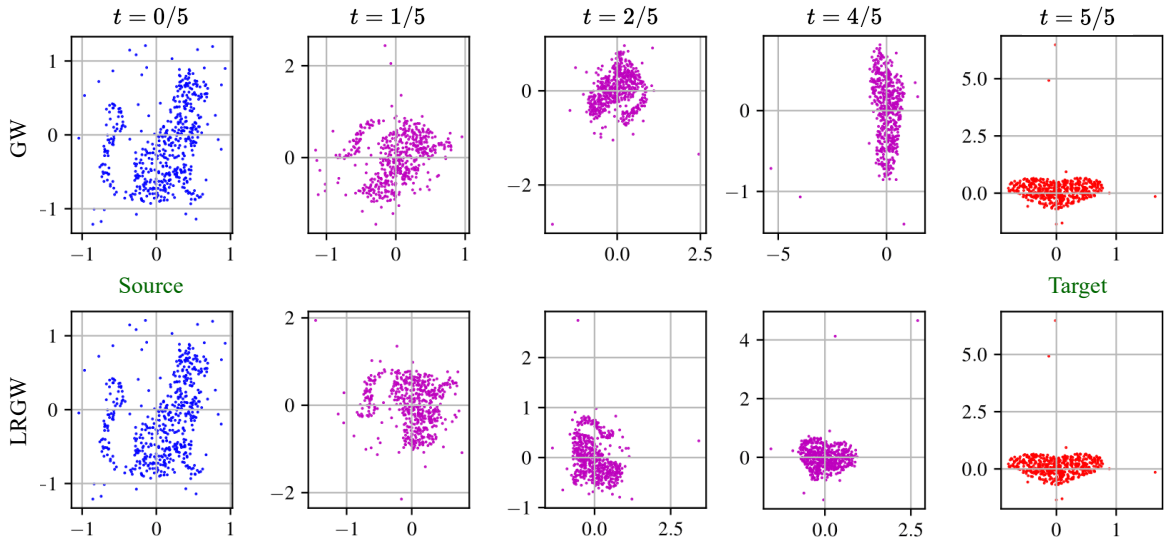


Figure 4.4: GW and LRGW barycenters between source (cat) and target (heart) datasets at levels $t = 0/5, 1/5, \cdots, 5/5$. The source contains 10% Gaussian outliers, whereas the target is contaminated with 10% Cauchy. The λ values for both domains are taken as respective 98-percentiles, i.e., 1.752 and 1.414. Even under far-lying Cauchy noise, LRGW barycenters recover original structures.

The contamination regime in this case remains similar to that in the shape-matching experiment. From Fig. 4.4, we observe that the barycentric shapes corresponding to LRGW (under Cauchy outliers in the target) suffer significantly less perturbation compared to GW. For example, at $t = \frac{4}{5}$, LRGW recovers the geometric integrity of the corresponding noise-free interpolation, where $t = \rho_1$. Similar traits can be observed under Gaussian contamination in both domains (see Appendix). This is even more challenging since Gaussian noise crowds the local neighborhood, triggering pronounced distortion of shapes. The improvement in barycenters using LRGW is remarkable, as it does not redistribute weights assigned to outliers, and the underlying optimization involves a full plan. Note that the arbitrary orientation of the barycenters is due to MDS, which does not affect alignment.

Remark 4.4 (Localization leading to pmm spaces)**.** *Though Dirac measures are the easiest choice to show that individual x's are represented in the pmm space, it is only a special case of a localized measure. Given $\alpha \in \bar{\mathcal{P}}(X)$, a localized measure $m_\alpha^L(x) \in \mathcal{P}(X)$ is tasked with preserving information about the neighborhood of $x \in X$ under the localization operator L*[7]*. Given the choice of the metric as W_p in the pmm space, it implies that $W_p(m_\alpha^L(x), m_\alpha^L(x')) = d_\alpha^L(x, x')$: a generalization over the metric d_X. This is the precise reason we name our proposal of a robust GW based on robust d_X 'local robustification'.*

Given two of such pmm spaces $(\mathcal{X}, W_p, \mu)$ and $(\mathcal{Y}, W_p, \nu)$, the GW distance ($\mathcal{Z}$-GW according to Bauer et al. (2024), Section 3.2.5) between them turns out as

$$d_{\mathrm{GW}}(\mu, \nu) \coloneqq \Big(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} \int_{\mathcal{X}\times\mathcal{Y}} \big| W_p(x, x') - W_p(y, y') \big|^{p'} d\pi \otimes \pi \Big)^{\frac{1}{p'}}. \tag{4.11}$$

For simplicity, we will always assume $p = p'$. However, it is not straightforward to imagine the ambient measures μ, ν and the feasible couplings Π they form. Let us look at an example that puts the problem in context.

Example 4.1 (Gromov-Wasserstein between mixture of Gaussians)**.** *Consider the mm spaces $(\mathcal{N}(\mathbb{R}^d), W_2, \mu)$ and $(\mathcal{N}(\mathbb{R}^{d'}), W_2, \nu)$, where $\mathcal{N}(\mathbb{R}^d)$ is the space of d-variate Gaussian distributions. Observe that*

(I) since Gaussians are exactly identifiable based on their mean and covariance matrix, a finite Gaussian mixture $\in \mathcal{G}(\mathbb{R}^d) \coloneqq \bigcup_{k\geq 0} \mathcal{G}_k(\mathbb{R}^d)$[8] *can be deemed a discrete probability distribution on $\mathcal{N}(\mathbb{R}^d)$.*

(II) Endowed with W_2, $\mathcal{N}(\mathbb{R}^d)$ becomes Polish.

(III) Given $\alpha_i = \mathcal{N}(m_i, \Sigma_i)$, $i = \{1, 0\}$ due to Dowson and Landau (1982)

$$W_2^2(\alpha_0, \alpha_1) = \|m_0 - m_1\|^2 + tr\Big[\Sigma_0 + \Sigma_1 - 2(\Sigma_0^{\frac{1}{2}} \Sigma_1 \Sigma_0^{\frac{1}{2}})^{\frac{1}{2}}\Big].$$

Hence, for any $\alpha = \sum_{i=1}^k a_i \alpha_i \in \mathcal{G}_k(\mathbb{R}^d)$ and $\beta = \sum_{j=1}^l b_j \beta_j \in \mathcal{G}_l(\mathbb{R}^{d'})$, there uniquely exist $\mu = \sum_{i=1}^k a_i \delta_{\alpha_i} \in \mathcal{P}(\mathcal{N}(\mathbb{R}^d))$ and $\nu = \sum_{j=1}^l b_j \delta_{\beta_j} \in \mathcal{P}(\mathcal{N}(\mathbb{R}^{d'}))$ respectively, where $a \coloneqq \{a_i\}_{i=1}^k \in \Delta_k$ and $b \coloneqq \{b_j\}_{j=1}^l \in \Delta_l$ (using (I)). (III) implies that, under such a μ, $(\mathcal{N}(\mathbb{R}^d), W_2, \mu)$ has finite 2-diameter, i.e. $\int_{\mathcal{N}(\mathbb{R}^d)\times\mathcal{N}(\mathbb{R}^d)} W_2^2(x, x') d\mu(x) d\mu(x') < \infty$ (same holds for $(\mathcal{N}(\mathbb{R}^{d'}), W_2, \nu)$). As such, the corresponding $d_{GW}(\mu, \nu)$ (as in (4.11), given $p = 2$) exists and follows all metric properties of GW. Salmona et al. (2024) show that solving the

[7] L maps $\bar{\mathcal{P}}(X)$ to Markov kernels over X (Memoli et al., 2019).
[8] $\mathcal{G}_k(\mathbb{R}^d) \coloneqq$ set of Gaussian mixtures with $\leq k$ components.

problem (4.11) essentially boils down to

$$\Big(\inf_{\pi\in\Pi(a,b)}\sum_{i,j,s,t}|W_2(\alpha_i,\alpha_s)-W_2(\beta_j,\beta_t)|^2\pi_{i,j}\pi_{s,t}\Big)^{\frac{1}{2}}, \tag{4.12}$$

where $\Pi(a,b)$ is a subset of the simplex $\Delta_{k\times l}$ with marginals a and b. This example can be further extended to general distribution classes $\subset\bar{\mathcal{P}}_p(\mathbb{R}^d)$ owing to the fact that $\mathcal{G}(\mathbb{R}^d)$ is dense in $\bar{\mathcal{P}}_p(\mathbb{R}^d)$ for W_p, as long as they are complete and separable.

Evidently, observations from the distributions μ and ν can suffer arbitrary corruptions as before. In cases such as Example 1, one or more contaminated individual Gaussian components from either space may contribute to such corruption. To remedy ϵ-contamination in the components, we replace W_p with the smallest cost achieved by 'optimally' removing ϵ-mass from them. This extends our LR framework to alignment models concerning pmm spaces, endowed with W_p. Given $\alpha,\beta\in\mathcal{P}(X)$, it is defined as

$$W_p^\epsilon(\alpha,\beta)=\inf_{\substack{\alpha',\beta'\in\mathcal{P}(X):\\ \alpha\in\mathcal{B}_\epsilon(\alpha'),\beta\in\mathcal{B}_\epsilon(\beta')}}W_p(\alpha',\beta'), \tag{4.13}$$

where $\mathcal{B}_\epsilon(\alpha):=\{(1-\epsilon)\alpha+\epsilon\gamma:\gamma\in\mathcal{P}(X)\}$ denotes the ϵ-Huber ball centered at α (Nietert et al., 2022). In this context, $\epsilon\in[0,1]$ signifies the radius of robustness, which when chosen distinctly for the two distributions generalizes the notion (i.e. $W_p^{\epsilon,\epsilon'}$). Remarkably, the dual formulation of $\mathrm{OT}_{l_\lambda(d_X)}$ can be derived as a special case of that of (4.13) (see Appendix). It also ties the threshold leading to truncation to the underlying optimization. Based on (4.13), we define the locally robust GW distance between pmm spaces (4.11) as

$$d_{\mathrm{LRGW}}(\mu,\nu;\epsilon):=\Big(\inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}}\big|W_p^\epsilon(x,x')-W_p(y,y')\big|^p d\pi\otimes\pi\Big)^{\frac{1}{p}}, \tag{4.14}$$

which also serves as a robust proxy to MGW_2 (Salmona et al., 2024) or the $\mathcal{Z}$-GW distance. While a generalization is imminent, the asymmetric robustness (only on space $\mathcal{X}$) in (4.14) is specifically useful in cross-domain generative tasks, e.g., unpaired image-to-image translation.

Proposition 4.4 (Dependence on robustness radius)**.** *For any $p\in[1,\infty)$ and $0\le\epsilon\le\epsilon'\le 1$ we have*

(i) $d_{LRGW}(\mu,\nu;0)=d_{GW}(\mu,\nu)$,

(ii) $|d_{LRGW}(\mu,\nu;\epsilon)-d_{LRGW}(\mu,\nu;\epsilon')|\lesssim\left(\frac{\epsilon'-\epsilon}{1-\epsilon}\right)^{\frac{1}{p}}$.

Remark 4.5 (Local robustness based on Lévy-Prokhorov metric)**.** *The LP distance between $\alpha,\beta\in\mathcal{P}(X)$ is defined as*

$$\hat{\rho}(\alpha,\beta):=\inf\{\epsilon>0:\beta(A)\le\alpha(A_\epsilon)+\epsilon,\forall\ \textit{Borel}\ A\subseteq X\},$$

where $A_\epsilon = \{x \in X : d_X(x, A) \leq \epsilon\}$ is the closed ϵ-ball around A. It inherently carries mass allocation robustness, allowing free movement of an ϵ-fraction mass between α and β. To observe the same, let us write LP in its alternative characterization due to Strassen's theorem (Villani (2021), Section 1.4)

$$\inf_{\pi \in \Pi(\alpha,\beta)} \Big\{ \inf\{\epsilon > 0 : \pi(\{(x, y) : d_X(x, y) > \epsilon\}) \leq \epsilon\} \Big\}. \tag{4.15}$$

As such, it follows that $\big|\hat{\rho}(\alpha, \beta) - \inf\{\epsilon > 0 : W_\infty^{1-\epsilon}(\alpha, \beta) \leq \epsilon\}\big|$ becomes arbitrarily small, given X has unit diameter (Raghvendra et al., 2024). As a result, the feasibility of LR formulations equivalent to (4.14) based on LP metrics is guaranteed. We show that local robustification using truncation, i.e., $l_\lambda(d_X)$, also extends to pmm spaces under the LP metric. It becomes evident due to the relation between $W_{p,\lambda}$ (ROBOT) and the modified LP.

Proposition 4.5. *Define $\hat{\rho}_\lambda(\alpha, \beta) = \inf_{\pi \in \Pi(\alpha,\beta)} \Big\{ \inf\{\epsilon > 0 : \pi(\{l_\lambda(d_X(x, y)) > \epsilon\}) \leq \epsilon\} \Big\}$, for $\lambda > 0$. Then*

$$\frac{1}{1+\lambda} W_{1,\lambda} \leq \hat{\rho}_\lambda \leq \sqrt{W_{1,\lambda}}.$$

4.4.2.1 Sturm's GW and robust image-to-image translation

Instilling intrinsic robustness to outliers in a measurable map (induced by neural networks) $\mathcal{X} \mapsto \mathcal{Y}$, learned based on contaminated data requires additional regularization. While introducing trimming methods (as in HGW or LRGW) in a GM setup may lead to denoised translations, the approximation capability of the maps thus produced remains shrouded. In this section, we rather connect natural upper bounds of GW to losses that fuel existing I2I models. This way, we ensure robustness guarantees without compromising complexity in an I2I translator.

In our pursuit, let us first define Sturm's GW distance (Sturm, 2006) between the altered spaces $(\mathcal{X}, l_\lambda(d_X), \mu)$ and $(\mathcal{Y}, l_\lambda(d_Y), \nu)$ as $\inf_{\tilde{d},\pi} ||\tilde{d}||_{L^p(\pi)}$. Here, the infimum is over $\pi \in \Pi(\mu, \nu)$ and $\tilde{d} \in \mathscr{D}(l_\lambda(d_X), l_\lambda(d_Y))$, the set of *metric couplings*[9]. Observe that, for $\lambda > 0$, $\mathscr{D}^\lambda := l_\lambda(\mathscr{D}(d_X, d_Y)) \subset \mathscr{D}(l_\lambda(d_X), l_\lambda(d_Y))$. As such,

$$\inf_{\tilde{d} \in \mathscr{D}(l_\lambda(d_X), l_\lambda(d_Y)), \pi} ||\tilde{d}||_{L^p(\pi)} \leq \inf_{\tilde{d} \in \mathscr{D}^\lambda, \pi} ||\tilde{d}||_{L^p(\pi)} =: d_{RSGW}(\mu, \nu), \tag{4.16}$$

which we call the (p, λ)-*Robust Sturm's GW*, $p \in [1, \infty)$. The distance essentially embodies a locally robust formulation based on the couplings between d_X and d_Y. We refer to the lower bound of (4.16) as the *lower* RSGW. Complementing the relationship between GW and Sturm's GW, the respective robust formulations follow a similar inequality.

[9] $\mathscr{D}(d_X, d_Y) :=$ the set of metrics on $\mathcal{X} \sqcup \mathcal{Y}$ that extend d_X and d_Y (Sturm, 2006).

Theorem 4.3 (Upper bound to LRGW). *Given $p \in [1, \infty)$ and $\lambda \geq 0$,*

$$(p, \lambda)\text{-}LRGW \leq 2(p, \lambda)\text{-}lRSGW.$$

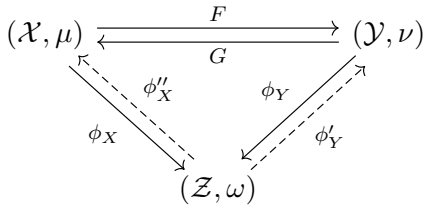
Also, for $\delta \in (0, \frac{1}{2}]$, whenever (p, λ)-$LRGW \leq \delta^5$, we have (p, λ)-$lRSGW \lesssim \lambda(4\lambda + \delta)^{\frac{1}{p}}$.

The immediate benefit of such a result is that minimizing a realized loss of the RSGW-type arbitrarily in an I2I setup establishes a near-isometric relation between the two image spaces. Intuitively, this should produce robust translations that preserve geometry. To invoke the notion of an actual architecture, we recall the equivalent formulation of Sturm's GW. In our setup,

$$d_{RSGW}(\mu, \nu) = \inf_{d \in \mathscr{D}(d_X, d_Y), \pi} \|l_\lambda(d)\|_{L^p(\pi)} = \inf_{\mathcal{Z}, \phi_X, \phi_Y} W_{p,\lambda}(\phi_{X\#}\mu, \phi_{Y\#}\nu), \tag{4.17}$$

where the infimum is over all isometric embeddings $\phi_X : \mathcal{X} \to \mathcal{Z}$ and $\phi_Y : \mathcal{Y} \to \mathcal{Z}$ into a *latent space* $\mathcal{Z}$, endowed with the metric d (Sturm (2006), lemma 3.3). We deliberately bring on the term 'latent space' to emphasize the connection to I2I architectures. The formulation also makes it sufficient to embed observations from both spaces into an optimal $\mathcal{Z}$ prior to truncation. Observe that if instead of $W_{p,\lambda}$, we deploy $\hat{\rho}_\lambda$ based on the metric d in (4.17), we obtain the robust Gromov-Prokhorov (RGP) metric (Blumberg et al. (2014), Section 2.5). By definition, RGP $< \epsilon$ implies the existence of a metric space $\mathcal{Z}$ with embeddings ϕ_X, ϕ_Y into it, that satisfy $\hat{\rho}_\lambda(\phi_{X\#}\mu, \phi_{Y\#}\nu) < \epsilon$.

Equivalence of losses. With the foundation in place, we explore the similarity between the loss (4.17) and that of I2I translation models such as UNIT (Liu et al., 2017a) and GcGAN (Fu et al., 2019). We choose the two models based on their sustained relevance in the domain. However, the equivalence about to be shown can be extended to models that recognize the role of a latent space or deploy a cycle-consistency (CC) loss, such as DistanceGAN (Benaim and Wolf, 2017), StarGAN (Choi et al., 2018), or MUNIT (Huang et al., 2018). The cornerstone of successful I2I learning is inarguably the CC loss. In the population regime, it can be expressed as $W_1(\mu, G \circ F_{\#}\mu)$ for the space $\mathcal{X}$, where F, G are optimized over measure-preserving (transport) maps parametrized using neural networks (NN). It becomes equivalent to optimizing the commonly used L^1 norm if μ possesses a Hölder smooth density (Chakrabarty and Das, 2022).



(4.18)

Now, recognizing the existence of a shared latent space, we may construct $G = \phi''_X \circ \phi_Y$ and $F = \phi'_Y \circ \phi_X$, where $\phi'_Y : \mathcal{Z} \to \mathcal{Y}$ is the left-inverse of ϕ_Y, and $\phi''_X : \mathcal{Z} \to \mathcal{X}$ is the right-inverse of ϕ_X. We can assume them to be full functional inverses, as the same applies to isomorphic embeddings, in which case CC is achieved a.s. However, the maps ϕ''_X, ϕ'_Y may not be measure-preserving in general. Therefore,

$$
\begin{aligned}
&W_1(\mu, G \circ F_{\#}\mu) \\
&= \inf_{\pi \in \Pi(\mu, F_{\#}\mu)} \int d_X(x, \phi''_X \circ \phi_Y(y))\, d\pi(x, y) \\
&= \inf_{\pi \in \Pi(\mu, F_{\#}\mu)} \int d\Big(\phi_X(x), (\phi_X \circ \phi''_X) \circ \phi_Y(y)\Big)\, d\pi(x, y) \qquad (4.19) \\
&= \inf_{\pi \in \Pi(\phi_{X\#}\mu, (\phi_Y \circ \phi'_Y) \circ \phi_{X\#}\mu)} \int d(x, y)\, d\pi(x, y) \\
&= \inf_{\pi \in \Pi(\omega, (\phi_Y \circ \phi'_Y)_{\#}\omega)} \int d(x, y)\, d\pi(x, y) \\
&= W_1(\omega, \phi_Y \circ \phi'_{Y\#}\omega), \qquad (4.20)
\end{aligned}
$$

where $d \in \mathscr{D}(d_X, d_Y)$[10]. We list out some immediate observations from the upper derivation. Firstly, constructing such a chaining $(\mathcal{X} \leftarrow \mathcal{Z} \leftarrow \mathcal{Y})$ reduces the problem of achieving CC in $\mathcal{X}$ to that of ensuring accurate autoencoding of ω based on the contextual latent law ν (4.20). The same choice of F, G also guarantees CC in $\mathcal{Y}$ a.s. Observe that, for any F satisfying $F_{\#}\mu = \nu$, given an optimal $\mathcal{Z}$ and the pair of embeddings into it, (4.19) equates to SGW. As such, SGW is an upper bound to the optimal CC loss $\inf_{F,G} W_1(\mu, G \circ F_{\#}\mu)$ when G follows our construction optimally, which is rather intuitive. It becomes much simpler if $\mu, \nu \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$, in which the construction can be made uniquely (see Appendix).

The second common loss component between UNIT and GcGAN is the constraint that ensures $F_{\#}\mu = \nu$ and $G_{\#}\nu = \mu$. Typically, the imposition is done using a GAN or WGAN objective. In our framework, a WGAN loss under 1-Lipschitz critics turns out as

$$
W_1(\mu, G_{\#}\nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int d_X(x, G(y))\, d\pi(x, y) = W_1(\phi_{X\#}\mu, \phi_{Y\#}\nu),
$$

which, again, at an optimal latent space equals SGW. Similarly, the loss $W_1(F_{\#}\mu, \nu)$ boils down to solving (4.20). As such, it is sufficient to optimize SGW between μ and ν subject to the autoencoding constraint (4.20) to solve the UNIT problem.

The only additional term GcGAN employs is namely the geometric-consistency (GC) loss.

[10] To avoid complications, we do not differentiate the two W_1 metrics in terms of notations, which are indeed calculated based on d_X and d respectively.

In the population regime, a $\mathcal{X} \xrightarrow{F} \mathcal{Y}$ translation model incurs a GC loss

$$W_1(F \circ s_{X\#}\mu, s_Y \circ F_{\#}\mu),$$

where s_X and s_Y are automorphisms in $\mathcal{X}$ and $\mathcal{Y}$ respectively, e.g. rotation. Based on our construction, considering $s_X = \phi''_X \circ \phi_X$ and $s_Y = \phi'_Y \circ \phi_Y = \mathrm{Id}_Y$ meets the constraint. Combining all the above observations gives the clear impression that effectively choosing a latent space $\mathcal{Z}$ — in turn, enabling appropriate construction of F and G — implies consistent I2I translation in UNIT and GcGAN. Remarkably, all the results hold exactly under the altered metric $W_{1,\lambda}$ (also, W_1^ϵ) since the constructions remain same for $l_\lambda(d_X)$ and $l_\lambda(d_Y)$ (4.17). As such, robustifying UNIT or GcGAN only requires updating their dependence on SGW to one with RSGW.

Experiment: Style transfer with noise

The first experiment we conduct tests the denoising capability of a robust GcGAN deploying (4.13) during I2I style transfer. Despite an overhaul in the optimization, we call our proposed model 'robust GcGAN' for simplicity. Notably, this is the first outlier-robust cross-domain generative model to our knowledge. Based on the dataset 'Apples-Oranges' (Zhu et al., 2017), the underlying task is to translate the visual style of oranges onto apples that are contaminated. Unlike the Huber setup here, standard Gaussian noise is added to the RGB channels of each target sample (apple). The mixing intensity α is kept at 0.2. We present a detailed discussion on the experimental setup in Appendix. As discussed in the previous section, it is sufficient to optimize the RSGW loss for a suitable $\mathcal{Z}$, which in this case is the image space itself. As a regularizer, we add the GC loss, taking s_X, s_Y as 90° clockwise rotations in their respective spaces. Model architecture and the choice of the Lagrangian parameter remain similar to that directed by the GcGAN authors. For comparison, the

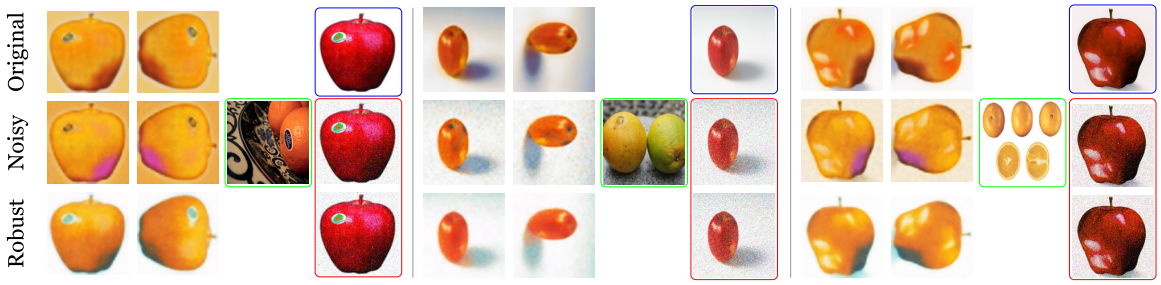


Figure 4.5: Style transfer performance of robust GcGAN under contamination ($\alpha = 0.3$). Images encircled in 'blue' represent clean target samples, in 'red' are noisy versions of them, and the ones in 'green' act as sources of the style to be transferred. At $\epsilon = 0.5$, the robust translations (third row) maintain sharpness and prevent artifacts from appearing, improving the FID score to 152.65 (compared to 154.74 in the noiseless setting: first row).

experiment contains three phases. As in the first row of Figure 4.5, we generate samples using the original GcGAN (without any modifications) on clean observations (control). The second row shows the degradation in translation once noise is added. Finally, applying our robust formulation at $\epsilon = 0.5$ we observe a significant improvement in images, both qualitative and quantitative. We present our parameter selection scheme in Appendix, in the form of an ablation study.

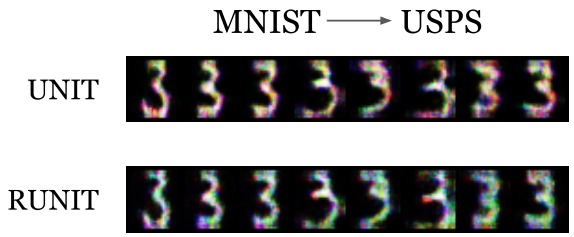


Figure 4.6: Unpaired translation under contamination ($\alpha = 0.4$) using robust UNIT. At $\epsilon = 0.5$, RUNIT recovers the visual quality of generated USPS samples (FID = 262.48, compared to 304.39 in case of UNIT under pixel noise).

The second experiment is of a different spirit in terms of the dissimilarity between dimensions of the two spaces, namely handwritten digit datasets MNIST (28×28) and USPS (16×16). Our goal lies in checking the robust domain translation capacity of a UNIT architecture reinforced with RSGW. Unlike style transfer, here, samples from MNIST (base distribution) are subjected to Gaussian noise. Keeping the robustness radius at 0.5, the robust UNIT model produces USPS samples with improved FID score (see Figure 4.6) compared to vanilla UNIT. Besides the expected denoising, the heightened image quality is a result of employing WGAN instead of vanilla GAN regularization.

4.4.3 Plan Robustification

All existing efforts to make GW robust to outliers rely on penalizing the plan π based on partial mass transport. Relieving the constraint from aligning all points enables the optimization to filter out outliers as only their contribution to the total mass is ignored. In a GW setup, unbalancing (Séjourné et al., 2021; Tran et al., 2023) in principle achieves the same by relaxing $\Pi(\mu, \nu)$ to $\mathcal{M}_+(\mathcal{X} \times \mathcal{Y})$, i.e. the set of positive Radon measures. However, it does not restrict the amount of mass to be pruned in its imposition of marginal constraints under quadratic ϕ-divergences. Moreover, it fails to guarantee an optimal plan in the exact sense. While a rescaling afterward produces a joint probability distribution, it only redistributes the leftover mass to inlying points from both spaces uniformly. Using the TV metric instead connects the problem to PGW (Bai et al., 2024). It readily carries out the redistribution, which ties the idea to our previous truncation methods (TGW and LRGW). While the redistribution pathway in TGW is not uniform, it is directed to the points, which through their

interactions with the other space, result in $\tau > 0$ distortion. In LR, points almost $\lambda > 0$ apart receive the mass.

Though imposed on top of an unbalanced GW between surrogates of μ and ν, Kong et al. (2024) is the only method to date that employs a direct penalization to robustify in the spirit of robust OT (Balaji et al., 2020). Since our motivation lies in constructing a robust I2I translation architecture, we rather prioritize a balanced formulation of the same kind that results in an optimal plan, and eventually maps. As encouragement, we draw an immediate connection between the robust penalization and ROT (Le et al., 2021). Observe that, the $(4, 2)$-GW distance between Euclidean mm spaces can be fragmented as $d^2_{\mathrm{GW}}(\mu, \nu) = S_1 + S_2$, where

$$S_2(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \Gamma(\pi) \coloneqq \inf_{\pi\in\Pi(\mu,\nu)} \int -\|x\|^2\|y\|^2 d\pi(x,y) - 2 \sum_{\substack{1\le i\le d\\ 1\le j\le d'}} \Big(\int x_i y_j d\pi(x,y)\Big)^2,$$

and S_1 depends solely on the marginals μ, ν (Zhang et al., 2024). It enables us to define

$$\tilde{S}_2(\mu,\nu) = \inf_{\substack{\tilde\mu\in\mathcal{P}(\mathcal{X})\\ \tilde\nu\in\mathcal{P}(\mathcal{Y})}} \min_{\pi\in\Pi(\tilde\mu,\tilde\nu)} \Gamma(\pi) + \lambda_1 D_f(\tilde\mu|\mu) + \lambda_2 D_f(\tilde\nu|\nu), \tag{4.21}$$

where D_f is some f-divergence and $\lambda_1, \lambda_2 > 0$.

Lemma 4.2 (Duality). *Given $(\mu, \nu) \in \mathcal{P}_4(\mathbb{R}^d) \times \mathcal{P}_4(\mathbb{R}^{d'})$, if $D_f \equiv d_{KL}$, we have*

$$\tilde{S}_2(\mu,\nu) = \inf_{\substack{\tilde\mu\in\mathcal{P}(\mathcal{X})\\ \tilde\nu\in\mathcal{P}(\mathcal{Y})}} \inf_{\boldsymbol{A}\in\mathcal{D}_{M_{\tilde\mu,\tilde\nu}}} 8\|\boldsymbol{A}\|_F^2 + ROT_{c_{\boldsymbol{A}}}(\mu,\nu),$$

where $\mathcal{D}_{M_{\tilde\mu,\tilde\nu}} = [-M^\infty_{\tilde\mu,\tilde\nu}/2, M^\infty_{\tilde\mu,\tilde\nu}/2]^{d\times d'}$ (see Theorem 4.2) and $ROT_{c_{\boldsymbol{A}}}(\mu,\nu) = OT_{c_{\boldsymbol{A}}}(\tilde\mu,\tilde\nu) + \lambda_1 d_{KL}(\tilde\mu|\mu) + \lambda_2 d_{KL}(\tilde\nu|\nu)$ is based on the cost $c_{\boldsymbol{A}}$ mapping $(x, y) \mapsto -\|x\|^2\|y\|^2 - 8x^T\boldsymbol{A}y$.

As such, the optimization underlying robust alignment is essentially a moment-constrained robust OT. If we only regularize based on one marginal (e.g., μ only), the optimization boils down to solving a semi-constrained problem, namely RSOT. The GW estimate corresponding to such a $\tilde{S}_2(\mu, \nu)$ can be made robust by plugging in the optimal marginals so obtained, $\tilde\mu, \tilde\nu$ in the functional S_1. This marks the potential that the robust penalization (4.21) has. In search of learnable maps between the spaces, we may narrow down the feasible set of couplings following Hur et al. (2024). Instead of $\Pi(\tilde\mu, \tilde\nu)$, we may choose the path-restricted distributions that follow the *binding constraint*

$$\{\pi : \pi = (\mathrm{Id}, F)_\#\tilde\mu = (G, \mathrm{Id})_\#\tilde\nu\} \subset \Pi(\tilde\mu, \tilde\nu),$$

where F, G are measurable maps between $\mathcal{X}, \mathcal{Y}$ (see illustration 4.18) and $\tilde\mu, \tilde\nu$ are supposedly 'clean' marginals. One can also relax this by choosing a larger subclass of Π that imposes only

$F_{\#}\tilde{\mu} = \tilde{\nu}$ and $G_{\#}\tilde{\nu} = \tilde{\mu}$. In any case, the resultant robust distance— an upper bound to GW under a similar penalization (Hur et al. (2024), Proposition 5.4)— only inculcates maps that transport inlying marginals to those in the other space. In a Huber contamination model, this readily implies that the corresponding set of couplings is non-empty. However, it is in principle different from learning a map possessing a denoising ability in the sense $F_{\#}\mu = \tilde{\nu}$. Maps of the latter kind promote carrying out the optimization over couplings given as

$$\{\pi : \pi = (\tilde{\mathrm{Id}}, F)_{\#}\mu = (G, \tilde{\mathrm{Id}})_{\#}\nu\},$$

where $\tilde{\mathrm{Id}}$ denotes the denoising operation that pushes forward μ to $\tilde{\mu}$ (also ν in its ambient space). The set containing such couplings is also non-empty since partial mass transport guarantees the existence of such 'robustifiers' $\tilde{\mathrm{Id}}$, and hence a pair of amenable maps (F, G). Given $\epsilon \in [0, 1)$, let us define the partial couplings

$$\Pi_{\epsilon}(\mu, \nu) = \{\pi : \pi = (\tilde{\mathrm{Id}}^{\epsilon}, F)_{\#}\mu = (G, \tilde{\mathrm{Id}}^{\epsilon})_{\#}\nu\}, \tag{4.22}$$

where $\tilde{\mathrm{Id}}^{\epsilon}_{\#}\alpha \leq (1-\epsilon)\alpha$, for $\alpha \in \mathcal{P}(\mathcal{X})$. The inequality should be understood setwise. We can also generalize the notion based on distinct mass fractions to be clipped in the two spaces. It is (4.22) that we base our final proposition on constructing a robust I2I translation model. We call the term $\inf_{\pi \in \Pi_{\epsilon}(\mu,\nu)} ||d_X - d_Y||_{L^p(\pi \otimes \pi)}$, the *robust reversible Gromov-Monge* (RRGM) distance. The formulation essentially is a robust surrogate to the RGM distance due to Hur et al. (2024) based on a partial alignment. To favor comparative analysis, we only consider $p = 2$ in our experiments. During transform sampling, RGM uses additional penalization imposing the constraints of measure preservation. The preferred metric for the same is often chosen as Maximum Mean Discrepancy (MMD). To eliminate the critical question of the ideal kernel given an empirical problem, we employ instead W_1. Under the same, the RRGM loss[11] can be written in a Lagrangian form given as

$$\begin{aligned} \inf_{F,G} \Big[\int \Big(d_X(\tilde{\mathrm{Id}}^{\epsilon}(x), G(y)) - d_Y(\tilde{\mathrm{Id}}^{\epsilon}(y), F(x)) \Big)^2 d\mu \otimes \nu \Big]^{\frac{1}{2}} \\ + \lambda_1 W_1(\tilde{\mathrm{Id}}^{\epsilon}_{\#}\mu, G_{\#}\nu) + \lambda_2 W_1(F_{\#}\mu, \tilde{\mathrm{Id}}^{\epsilon}_{\#}\nu), \end{aligned} \tag{4.23}$$

where the infimum is over measurable maps and $\lambda_1, \lambda_2 > 0$. We may find a further lower bound to the loss owing to the fact that

$$W_1(\tilde{\mathrm{Id}}^{\epsilon}_{\#}\mu, G_{\#}\nu) \geq W_1^{\epsilon}(\mu, G_{\#}\nu), \tag{4.24}$$

where W_1^{ϵ} only carries out a partial transport of μ asymmetrically (see (4.13)). The same

[11] The loss (4.23) relaxes the binding constraint (as in (4.22)) and only imposes robust measure preservation. As such, it is essentially a lower bound to RRGM.

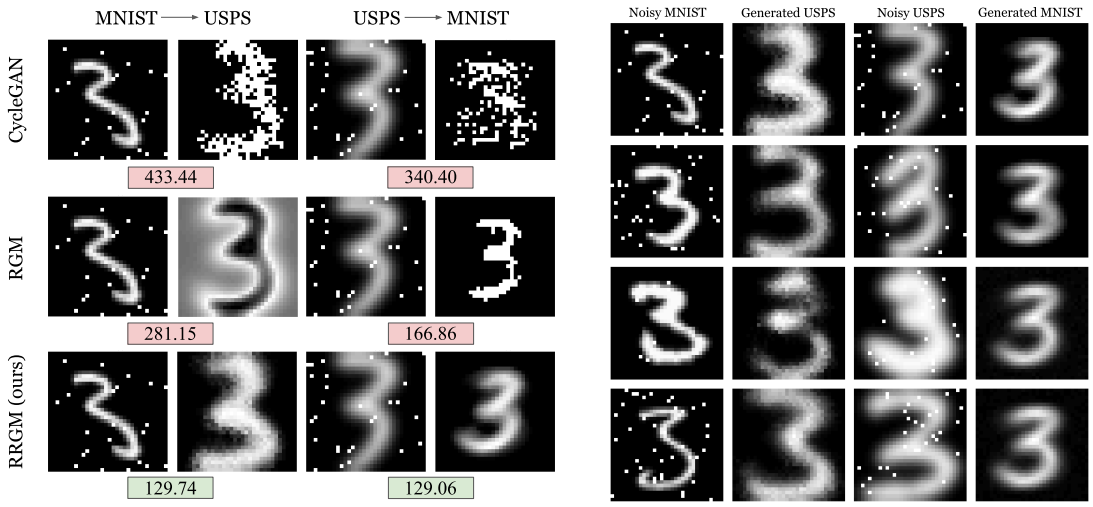
argument holds for the other term, given an asymmetric W_1^ϵ employment on the other space. During demanding I2I translations, it is often beneficial to have learnable discriminators over 1-Lipschitz dual maps. The usage of W_1 is also advantageous since it enables deploying a larger class of neural network-induced critics. As such, the two added losses can be optimized using WGAN-GP (Gulrajani et al., 2017) architectures. Despite promoting a different redistribution pathway of inlying mass, the sample complexity of transform sampling under W_1^ϵ should be of a similar order to LR constraints (Nietert et al. (2023), Theorem 4). We note that the convergence rate corresponding to the latter depends only on the inlying sample size (see Appendix) if outliers remain bounded in number.

Experiment: Image-to-image translation with noise

We test RRGM in a noisy MNIST↔USPS domain translation experiment. The contamination regime for the experiment remains the same as that in RUNIT. Maintaining $\alpha = 0.4$, we randomly select pixel locations following a Gaussian law and set their values to 1.0 (bright white) for all channels, adding visible outlier points to the image. Since handwritten digit images have all information regarding the numerical in the shape of white pixels, this contamination becomes quite challenging for an I2I model. For example, CycleGAN (Zhu et al., 2017) performs poorly despite employing a cycle-consistency component and generative losses in both directions. In contrast, RRGM (based on (4.24) with $\epsilon = 0.5$) under $\lambda_i = 0.2$; $i = 1, 2$ generates significantly sharper and denoised samples (Figure 4.7a). For a fair comparison, we also present both clean and noisy samples to the discriminators in CycleGAN. Even if the discriminators are shown noisy observations only, the generation quality of RRGM surpasses that of CycleGAN. As a reference, we maintain a similar parameter selection for RGM (Hur et al., 2024), which deploys an additional MMD loss to impose the binding constraint. However, in the absence of dedicated critic modules, it lags behind. Our model outperforms both techniques by a significant margin, in terms of both quantitative and qualitative measures.

4.5 Discussion

We explore three major possibilities for the robustification problem in a GW setup. Drawing from classical techniques in robust statistics, we propose novel GW surrogate distances TGW (and HGW) and LRGW that limit contamination due to outliers. We study their interrelation and their respective dependence on the truncation parameters. For TGW, we comment on its population-level robust guarantees and the resilience it offers to underlying distributions. Based on a data-dependent parameter selection scheme, we present a working algorithm to solve the HGW distance, which exhibits superior protection against outliers compared to existing methods in shape-matching tasks. On the other hand, solving LRGW-type measures boils down to calculating an OT loss between trimmed samples from the underlying distri-



(a) (b)

Figure 4.7: (a) FID scores corresponding to robust cross-domain generations between USPS and MNIST data under Gaussian contamination, using CycleGAN, RGM, and RRGM (ours). (b) Denoised generated samples using RRGM in both domains. The robust recovery empirically demonstrates the concentration around unperturbed losses (Proposition 4.6).

butions. It also hints at the effective level of trimming necessary (λ) given a sample problem. We generalize our notion to probabilistic mm spaces, which allows one to define LR alignment between mixtures of distributions of distinct dimensions. We extend this setup to introduce robust image translation networks that surpass existing benchmarks. We also propose in RRGM a cross-domain transform sampling framework that is robust to outliers. It promotes robust concentration and uniform deviation bounds, besides significantly improving image quality in I2I translations.

Our robust solutions to GW (also GM) and associated robust OT upper bounds can be immediately adopted to WAE and CycleGAN-type networks. In fact, we introduce superior translation architectures surpassing earlier benchmarks. As a potential future direction, we suggest integrating our solutions with GWGAN (Bunne et al., 2019) and GWAE (Nakagawa et al., 2023) models, which extend the learning of GAN and WAE-type architectures to diverse participating spaces.

4.6 Appendix: Proofs and Experimental Details

4.6.1 Proof of Proposition 4.1

(i) First, let us prove the triangle inequality for the 'norm' $\|\cdot\|_{\mathcal{T}_p}$ (not a norm following the formal definition). Assume $f, g \in L^p(\mu)$, where $\mu \in \mathcal{P}(\mathcal{X})$. Now,

$$\begin{aligned}
\|f+g\|^p_{\mathcal{T}_p(\mu)} &= \int_{\mathcal{X}} \mathcal{T}_p(f(x)+g(x))d\mu(x) \\
&= \int_{\mathcal{X}} \Big(\mathcal{T}_p(f+g)^{\frac{1}{p}}\Big)\Big(\mathcal{T}_p(f+g)^{\frac{p-1}{p}}\Big)d\mu \\
&\leq \int_{\mathcal{X}} \Big(\mathcal{T}_p^{\frac{1}{p}}(f)+\mathcal{T}_p^{\frac{1}{p}}(g)\Big)\Big(\mathcal{T}_p(f+g)^{\frac{p-1}{p}}\Big)d\mu \qquad (4.25)\\
&= \int_{\mathcal{X}} \mathcal{T}_p^{\frac{1}{p}}(f)\Big(\mathcal{T}_p(f+g)^{\frac{p-1}{p}}\Big)d\mu + \int_{\mathcal{X}} \mathcal{T}_p^{\frac{1}{p}}(g)\Big(\mathcal{T}_p(f+g)^{\frac{p-1}{p}}\Big)d\mu \\
&\leq \Big[\Big(\int_{\mathcal{X}} \mathcal{T}_p(f)d\mu\Big)^{\frac{1}{p}} + \Big(\int_{\mathcal{X}} \mathcal{T}_p(g)d\mu\Big)^{\frac{1}{p}}\Big]\Big(\int_{\mathcal{X}} \mathcal{T}_p(f+g)d\mu\Big)^{\frac{p-1}{p}} \qquad (4.26)\\
&= \Big(\|f\|_{\mathcal{T}_p(\mu)} + \|g\|_{\mathcal{T}_p(\mu)}\Big)\|f+g\|^{p-1}_{\mathcal{T}_p(\mu)},
\end{aligned}$$

where inequality (4.25) is due to the subadditivity of $\mathcal{T}_p^{\frac{1}{p}}$ (Musco et al. (2021), Lemma C.12) and Hölder inequality implies (4.26). In the process, we assume that the norm itself is not 0.

Given arbitrary $\varepsilon > 0$, one can obtain feasible optimal couplings $\pi_{XZ} \in \Pi(\mu_X, \mu_Z)$ and $\pi_{ZY} \in \Pi(\mu_Z, \mu_Y)$ that satisfy the following

$$\frac{1}{2}\|\Lambda_{X,Z}\|_{\mathcal{T}_p(\pi_{XZ}\otimes\pi_{XZ})} + \frac{1}{2}\|\Lambda_{Z,Y}\|_{\mathcal{T}_p(\pi_{ZY}\otimes\pi_{ZY})} = d_{\mathrm{TGW}}(Z,Y) + d_{\mathrm{TGW}}(X,Z) + 2\varepsilon. \qquad (4.27)$$

The Gluing lemma ensures the existence of $\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$ with marginals π_{XZ} on $\mathcal{X}\times\mathcal{Z}$ and π_{ZY} on $\mathcal{Z}\times\mathcal{Y}$. Also, let π_{XY} be the marginal of π on $\mathcal{X}\times\mathcal{Y}$. Observe that, given $x, x' \in \mathcal{X}$, $y, y' \in \mathcal{Y}$ and $z, z' \in \mathcal{Z}$, the triangle inequality of Λ_1 implies

$$\Lambda_1(d_X(x,x'), d_Y(y,y')) \leq \Lambda_1(d_X(x,x'), d_Z(z,z')) + \Lambda_1(d_Z(z,z'), d_Y(y,y'))$$

$(\pi \otimes \pi)$-a.e. Now,

$$\begin{aligned} d_{\mathrm{TGW}}(X,Y) &\leq \frac{1}{2}\|\Lambda_{X,Y}\|_{\mathcal{T}_p(\pi_{XY}\otimes\pi_{XY})} = \frac{1}{2}\|\Lambda_{X,Y}\|_{\mathcal{T}_p(\pi\otimes\pi)} \\ &\leq \frac{1}{2}\|\Lambda_{X,Z}+\Lambda_{Z,Y}\|_{\mathcal{T}_p(\pi\otimes\pi)} \\ &\leq \frac{1}{2}\|\Lambda_{X,Z}\|_{\mathcal{T}_p(\pi\otimes\pi)} + \frac{1}{2}\|\Lambda_{Z,Y}\|_{\mathcal{T}_p(\pi\otimes\pi)} \qquad (4.28) \\ &= \frac{1}{2}\|\Lambda_{X,Z}\|_{\mathcal{T}_p(\pi_{XZ}\otimes\pi_{XZ})} + \frac{1}{2}\|\Lambda_{Z,Y}\|_{\mathcal{T}_p(\pi_{ZY}\otimes\pi_{ZY})} \\ &= d_{\mathrm{TGW}}(Z,Y) + d_{\mathrm{TGW}}(X,Z) + 2\varepsilon, \end{aligned}$$

where (4.28) is due to the triangle inequality of the Tukey norm. Since the choice of $\varepsilon > 0$ is arbitrary, this completes the proof.

(ii) The proof of this part follows from the monotonicity of L^p norms. Observe that

$$\|\Lambda_{X,Y}\|_{\mathcal{T}_p(\gamma\otimes\gamma)} = \left\|\mathcal{T}_p^{\frac{1}{p}}(\Lambda_{X,Y})\right\|_{L^p(\gamma\otimes\gamma)} \leq \left\|\mathcal{T}_{p'}^{\frac{1}{p'}}(\Lambda_{X,Y})\right\|_{L^{p'}(\gamma\otimes\gamma)} = \|\Lambda_{X,Y}\|_{\mathcal{T}_{p'}(\gamma\otimes\gamma)}.$$

(iii) The definition of the Tukey loss implies that the corresponding distance is non-decreasing in τ.

4.6.2 Proof of Theorem 4.1

Triangle inequality of d_X implies that given $(x,y),(x',y') \sim \mu' \otimes \nu$

$$2\Lambda_1(d_X,d_X) = \left|d_X(x,x') - d_X(y,y')\right| \leq d_X(x,y) + d_X(x',y').$$

Thus, for a coupling $\pi \in \Pi(\mu',\nu)$, the triangle inequality of the norm $\|\cdot\|_{\mathcal{T}_p}$ implies

$$\|2\Lambda_1\|_{\mathcal{T}_p(\pi\otimes\pi)} \leq 2\|d_X\|_{\mathcal{T}_p(\pi)}. \qquad (4.29)$$

Now,

$$\begin{aligned} \inf_{\pi\in\Pi}\|d_X\|_{\mathcal{T}_p(\pi)} &= \inf_{\pi\in\Pi}\left(\int \mathcal{T}_p(d_X(x,y))d\pi\right)^{\frac{1}{p}} = W_{\mathcal{T}_p}(\mu',\nu) \\ &\leq W_{\mathcal{T}_p}((1-\epsilon)\mu + \epsilon\mu_c,\mu) + W_{\mathcal{T}_p}(\mu,\nu) \qquad (4.30) \\ &\leq W_{\mathcal{T}_p}(\epsilon\mu,\epsilon\mu_c) + W_{\mathcal{T}_p}(\mu,\nu) \qquad (4.31) \\ &\leq \epsilon^{\frac{1}{p}} W_{\mathcal{T}_p}(\mu,\mu_c) + W_{\mathcal{T}_p}(\mu,\nu) \\ &\leq \tau\epsilon^{\frac{1}{p}} + W_{\mathcal{T}_p}(\mu,\nu), \qquad (4.32) \end{aligned}$$

where (4.31) is due to the fact that for any $\alpha, \beta \in \mathcal{P}(\mathcal{X})$, we have $\mathrm{OT}_{d_X}(\alpha, \beta) \leq \mathrm{OT}_{d_X}(\alpha - \alpha \wedge \beta, \beta - \alpha \wedge \beta)$. The triangle inequality of $W_{\mathcal{T}_p}$ (4.30) follows a similar proof to that of the p-Wasserstein distance. Inequality (4.32) is due to the trivial upper bound of the Tukey function. This, along with the previous observation, completes the proof.

4.6.3 Proof of Corollary 4.1

For any $\tilde{\mu} \leq \frac{1}{1-\varepsilon}\mu$, by choosing an appropriate $\beta \in \mathcal{P}(\mathcal{X})$ we can write $\mu = (1-\varepsilon)\tilde{\mu} + \varepsilon\beta$. Now,

$$\begin{aligned} W_{\mathcal{T}_p}(\mu, \tilde{\mu}) &\leq \varepsilon^{\frac{1}{p}}(W_{\mathcal{T}_p}(\beta, \delta_{x_0}) + W_{\mathcal{T}_p}(\delta_{x_0}, \tilde{\mu})) \\ &= \varepsilon^{\frac{1}{p}}\Big[\|d_X(Y, x_0)\|_{\mathcal{T}_p(\beta)} + \|d_X(Y, x_0)\|_{\mathcal{T}_p(\tilde{\mu})}\Big] \\ &\leq 2\varepsilon^{\frac{1}{p}} \sup_{\alpha \in \mathcal{P}(\mathcal{X}), \alpha \leq \frac{1}{1-\varepsilon'}\mu} \|d_X(Y, x_0)\|_{\mathcal{T}_p(\alpha)}, \end{aligned} \tag{4.33}$$

where $\varepsilon' \coloneqq \varepsilon \vee (1-\varepsilon)$. Given that the expectation is finite, the definition of $\mathcal{T}_p$ implies $\mathbb{E}_\alpha[\mathcal{T}_p(d_X(Y, x_0))] \leq \tau^p$ for all such α as in (4.33). Moreover,

$$\begin{aligned} \mathbb{E}_\alpha[\mathcal{T}_p(d_X(Y, x_0))] &\leq |\mathbb{E}_\alpha[\mathcal{T}_p(d_X(Y, x_0))] - \mathbb{E}_\mu[\mathcal{T}_p(d_X(Z, x_0))]| + \mathbb{E}_\mu[\mathcal{T}_p(d_X(Z, x_0))] \\ &\leq \Big(1 \vee \frac{1-\varepsilon}{\varepsilon}\Big)\rho + (\sigma^p \wedge \tau^p), \end{aligned} \tag{4.34}$$

where the first inequality is due to triangle inequality. As such, combining inequality (4.34) with the bound (4.33) yields,

$$W_{\mathcal{T}_p}(\mu, \tilde{\mu}) \leq 2\Big((\rho^{\frac{1}{p}} + \varepsilon^{\frac{1}{p}}(\sigma \wedge \tau)) \wedge \varepsilon^{\frac{1}{p}}\tau\Big).$$

4.6.4 Existence of optimal couplings in LRGW

Given Polish mm spaces $(\mathcal{X}, d_X, \mu)$, $(\mathcal{Y}, d_Y, \nu)$ define the locally robust p-distortion realized by a coupling $\pi \in \Pi(\mu, \nu)$ as, $p \in [0, \infty]$

$$J_p^\lambda(\pi) \coloneqq \|l_\lambda(d_X) - l_\lambda(d_Y)\|_{L^p(\pi \otimes \pi)},$$

where $\lambda > 0$. Then,

Lemma 4.3. *There exists a coupling $\pi^\lambda \in \Pi(\mu, \nu)$ such that, (p, λ)-LRGW $= J_p^\lambda(\pi)$.*

The lemma can be proved by extending Corollary 10.1 of Mémoli (2011) for the mm spaces with modified metrics $l_\lambda(d_X)$ and $l_\lambda(d_Y)$. We give a version of the proof for completeness.

Proof. First, observe that the set of couplings $\Pi(\mu, \nu)$ is sequentially compact in $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ (Sturm (2023), lemma 1.2). Hence, for $1 \leq p < \infty$, it suffices to show that $J_p^\lambda(\pi)$ is continuous

on $\Pi(\mu,\nu)$. Let us choose the metric $d^\lambda((x,y),(x',y')) = l_\lambda(d_X(x,x'))+l_\lambda(d_Y(y,y'))$ mapping $\mathcal{X}\times\mathcal{Y}$ to $[0,2\lambda]$. Given $((x_i,y_i),(x_i',y_i'))\sim\pi\otimes\pi$ for $i=1,2$, observe that

$$
\begin{aligned}
&\big|\big|l_\lambda(d_X(x_1,x_1')) - l_\lambda(d_Y(y_1,y_1'))\big| - \big|l_\lambda(d_X(x_2,x_2')) - l_\lambda(d_Y(y_2,y_2'))\big|\big| \\
&\le \big|l_\lambda(d_X(x_1,x_1')) - l_\lambda(d_X(x_2,x_2')) + l_\lambda(d_Y(y_2,y_2')) - l_\lambda(d_Y(y_1,y_1'))\big| \\
&\le \big|l_\lambda(d_X(x_1,x_1')) - l_\lambda(d_X(x_2,x_2'))\big| + \big|l_\lambda(d_Y(y_2,y_2')) - l_\lambda(d_Y(y_1,y_1'))\big| \\
&\le [l_\lambda(d_X(x_1,x_2)) + l_\lambda(d_Y(y_1,y_2))] + [l_\lambda(d_X(x_1',x_2')) + l_\lambda(d_Y(y_1',y_2'))] \qquad (4.35)\\
&= d^\lambda((x_1,y_1),(x_2,y_2)) + d^\lambda((x_1',y_1'),(x_2',y_2')), \qquad (4.36)
\end{aligned}
$$

which implies the Lipschitz continuity of $f_1 \coloneqq 2\Lambda_1(l_\lambda(d_X), l_\lambda(d_Y))$ for the metric given by (4.36). The step (4.35) follows from the triangle inequalities of $l_\lambda(d_X)$ and $l_\lambda(d_Y)$. Now, consider a function $f_2:[0,2\lambda]\to\mathbb{R}_+$ mapping $t\mapsto t^p$, which in turn implies that $f_2\circ f_1$ is Lipschitz with constant $\le p(2\lambda)^{p-1}$. Thus, given a sequence $\{\pi_n\}_{n\in\mathbb{N}}\subset\mathcal{P}(\mathcal{X}\times\mathcal{Y})$ such that $\pi_n \xrightarrow{w} \pi$, we have

$$
\begin{aligned}
&\Big|\int \underbrace{\int f_2\circ f_1\,\pi_n(d(x,y))}_{\coloneqq f_{\pi_n}(x',y')}\,\pi_n(d(x',y')) - \int \underbrace{\int f_2\circ f_1\,\pi(d(x,y))}_{\coloneqq f_{\pi}(x',y')}\,\pi(d(x',y')\Big| \\
&\le \Big|\int (f_{\pi_n} - f_\pi)\,\pi_n(d(x',y'))\Big| + \Big|\int f_\pi\,\pi_n(d(x',y')) - f_\pi\,\pi(d(x',y'))\Big| \\
&\le \max_{(x',y')} |f_{\pi_n} - f_\pi| + \Big|\int f_\pi\,d\pi_n - f_\pi\,d\pi\Big|,
\end{aligned}
$$

where the first term on the right-hand side vanishes as $n\to\infty$ due to the uniformly convergent $f_{\pi_n}\to f_\pi$ (based on the point-wise convergence of f_{π_n} and the Lipschitz continuity of $f_2\circ f_1$). The weak convergence of π_n ensures that the second term also vanishes. As such, $J_p^\lambda(\pi_n)\to J_p^\lambda(\pi)$ as $n\to\infty$. Hence, the proof.

To show the result for $p=\infty$, first observe that given π, the sequence $\{J_l^\lambda(\pi)\}$ is non-decreasing in $1\le l\le\infty$ (using Jensen's inequality) and $\lim_{p\to\infty} J_p^\lambda(\pi)\to J_\infty^\lambda(\pi) = \sup\{J_p^\lambda : 1\le p<\infty\}$. As such, $J_\infty^\lambda(\pi)$ is lower semi-continuous. Now, invoking the compactness argument proves the infimum is achieved in $\Pi(\mu,\nu)$. □

4.6.5 Proof of Proposition 4.2

(i) Given $\lambda'>\lambda\ge 0$, observe that

$$
|l_\lambda(d_X) - l_\lambda(d_Y)| \le |l_{\lambda'}(d_X) - l_{\lambda'}(d_Y)|,
$$

a.s. $(\mu \otimes \nu)^{\otimes 2}$, which proves the first claim. Also, for all $a \in \mathbb{R}$ the following equality holds

$$l_{\lambda'}(a) - l_{\lambda}(a) = \min\{\lambda', \max\{a, \lambda\}\} - \lambda = l_{\lambda'}(\bar{l}_{\lambda}(a)) - \lambda.$$

We point out that, even a stronger equality holds in general almost surely

$$|l_{\lambda'}(d_X) - l_{\lambda'}(d_Y)| = |l_{\lambda}(d_X) - l_{\lambda}(d_Y)| + \big|l_{\lambda'}(\bar{l}_{\lambda}(d_X)) - l_{\lambda'}(\bar{l}_{\lambda}(d_Y))\big|. \tag{4.37}$$

Now, given any π that solves the (p, λ)-LRGW(μ, ν) problem

$$\begin{aligned}(p, \lambda')\text{-LRGW}(\mu, \nu) - (p, \lambda)\text{-LRGW}(\mu, \nu) &\leq \|l_{\lambda'}(d_X) - l_{\lambda'}(d_Y)\|_{L^p(\pi\otimes\pi)} - (p, \lambda)\text{-LRGW}(\mu, \nu) \\ &\leq \big\|l_{\lambda'}(\bar{l}_{\lambda}(d_X)) - l_{\lambda'}(\bar{l}_{\lambda}(d_Y))\big\|_{L^p(\pi\otimes\pi)},\end{aligned}$$

where the second inequality is due to (4.37) and the the triangle inequality of L^p norms. The proof follows since the choice of π is arbitrary in Π^{λ}. We also present an upper bound that depends on the spaces' regulated p-diameters. First, $\forall a \in \mathbb{R}$ let us write

$$|l_{\lambda+\lambda'}(a) - l_{\lambda'}(a)| = \bar{l}_0(l_{\lambda}(a - \lambda')).$$

Now, for any optimal coupling $\pi \in \Pi(\mu, \nu)$ for the (p, λ')-LRGW problem, we get

$$\begin{aligned}(p, \lambda + \lambda')\text{-LRGW}(\mu, \nu) &\leq \|l_{\lambda+\lambda'}(d_X) - l_{\lambda+\lambda'}(d_Y)\|_{L^p(\pi\otimes\pi)} \\ &\leq \|l_{\lambda+\lambda'}(d_X) - l_{\lambda'}(d_X)\|_{L^p(\mu\otimes\mu)} + \|l_{\lambda'}(d_X) - l_{\lambda'}(d_Y)\|_{L^p(\pi\otimes\pi)} \\ &\qquad + \|l_{\lambda'}(d_Y) - l_{\lambda+\lambda'}(d_Y)\|_{L^p(\nu\otimes\nu)} \\ &\leq \big\|\bar{l}_0(l_{\lambda}(d_X(x, x') - \lambda'))\big\|_{L^p(\mu\otimes\mu)} + (p, \lambda')\text{-LRGW}(\mu, \nu) \\ &\qquad + \big\|\bar{l}_0(l_{\lambda}(d_Y(y, y') - \lambda'))\big\|_{L^p(\nu\otimes\nu)} \\ &= d_{\mathcal{X},p}^{\lambda,\lambda'} + d_{\mathcal{Y},p}^{\lambda,\lambda'} + (p, \lambda')\text{-LRGW}(\mu, \nu),\end{aligned} \tag{4.38}$$

where

$$\begin{aligned}\|l_{\lambda}(d_X)\|_{L^p(\mu\otimes\mu)} &\geq d_{\mathcal{X},p}^{\lambda,\lambda'} \\ &= \big\|\bar{l}_0(l_{\lambda}(d_X(x, x') - \lambda'))\big\|_{L^p(\mu\otimes\mu)} \\ &= \big\|\bar{l}_{\lambda'}(l_{\lambda+\lambda'}(d_X(x, x'))) - \lambda'\big\|_{L^p(\mu\otimes\mu)} \\ &\geq \bar{l}_{\lambda'}\big(\big\|l_{\lambda+\lambda'}(d_X(x, x'))\big\|_{L^p(\mu\otimes\mu)}\big) - \lambda' \\ &= \bar{l}_0\big(\|l_{\lambda+\lambda'}(d_X)\|_{L^p(\mu\otimes\mu)} - \lambda'\big).\end{aligned}$$

As such, $(p, \lambda+\lambda')$-LRGW$(\mu,\nu) - (p,\lambda')$-LRGW$(\mu,\nu) \leq 2\max_{\mathcal{Z}\in(\mathcal{X},\mathcal{Y})} d_{\mathcal{Z},p}^{\lambda,\lambda'}$.
(ii) Choosing $\lambda' = \infty$ in the first argument of *(i)* proves the result.

(iii) For $a, b \geq 0$ and $p \in [1,\infty)$, the following inequality holds $l_{\lambda^p}(|a-b|^p) \geq |l_\lambda(a) - l_\lambda(b)|^p$. Based on the same, given arbitrary $x, x' \sim \mu$ and $y, y' \sim \nu$, we find

$$|l_\lambda[d_X(x,x')] - l_\lambda[d_Y(y,y')]|^p \leq (\lambda \wedge 2M)^{p-q} |l_\lambda[d_X(x,x')] - l_\lambda[d_Y(y,y')]|^q$$

holds a.e. Here, $\lambda \geq M = \mathrm{diam}(\mathcal{X}) \vee \mathrm{diam}(\mathcal{Y})$. Now, picking a feasible coupling $\pi \in \Pi(\mu,\nu)$ yields

$$(p,\lambda)\text{-LRGW}(\mu,\nu) \leq \|l_\lambda(d_X) - l_\lambda(d_Y)\|_{L^p(\pi\otimes\pi)} \leq (\lambda\wedge 2M)^{1-\frac{q}{p}} \Big(\|l_\lambda(d_X) - l_\lambda(d_Y)\|_{L^q(\pi\otimes\pi)} \Big)^{\frac{q}{p}}.$$

This proves the result since π is arbitrary.

4.6.6 Proof of Proposition 4.3

The Proposition can be proved by extending Theorem 3.1 of Chowdhury and Mémoli (2019) for the mm spaces with modified metrics $l_\lambda(d_X)$ and $l_\lambda(d_Y)$. We give a version of the proof for completeness. To show the inequality (III), observe that

$$\begin{aligned} d_{\mathrm{LRGW}}(X,Y) &= \inf_{\pi\in\Pi(\mu,\nu)} \|l_\lambda(d_X) - l_\lambda(d_Y)\|_{L^p(\pi\otimes\pi)} \\ &\geq \inf_{\pi,\pi'\in\Pi(\mu,\nu)} \|l_\lambda(d_X) - l_\lambda(d_Y)\|_{L^p(\pi\otimes\pi')} \\ &\geq \inf_{\pi\in\Pi(\mu,\nu)} \Big\| \inf_{\pi'\in\Pi(\mu,\nu)} \|l_\lambda[d_X(x,\cdot)] - l_\lambda[d_Y(y,\cdot)]\|_{L^p(\pi')} \Big\|_{L^p(\pi)} \\ &= \inf_{\pi\in\Pi(\mu,\nu)} \Big\| \mathcal{E}^\lambda_{p,\mathcal{X},\mathcal{Y}} \Big\|_{L^p(\pi)}, \end{aligned} \tag{4.39}$$

where the first inequality is achieved by taking infimum over a larger set. Now, given the fact that for measurable maps $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$ we have $W_p(f_\#\mu, g_\#\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \|f-g\|_{L^p(\pi)}$ (Chowdhury and Mémoli, 2019), the following holds

$$\mathcal{E}^\lambda_{p,\mathcal{X},\mathcal{Y}} = \inf_{\pi\in\Pi(\mu,\nu)} \|l_\lambda[d_X(x,\cdot)] - l_\lambda[d_Y(y,\cdot)]\|_{L^p(\pi)} = W_p\big(l_\lambda[d_X(x,\cdot)]_\#\mu, l_\lambda[d_Y(y,\cdot)]_\#\nu\big) = C(x,y).$$

Now, we show that (III) $\geq$ (I). Let us first denote $\mathcal{E}^\lambda_{p,\mathcal{X}}(x) \coloneqq \|l_\lambda[d_X(x,\cdot)]\|_{L^p(\mu)}$ mapping $\mathcal{X} \mapsto R_+$, such that $\mathrm{size}^\lambda_p(\mathcal{X}) = \Big\|\mathcal{E}^\lambda_{p,\mathcal{X}}\Big\|_{L^p(\mu)}$. Observe that for a coupling π, using Minkowski's inequality

$$\Big\|\mathcal{E}^\lambda_{p,\mathcal{X}} - \mathcal{E}^\lambda_{p,\mathcal{Y}}\Big\|_{L^p(\pi)} \geq \Big| \Big\|\mathcal{E}^\lambda_{p,\mathcal{X}}\Big\|_{L^p(\pi)} - \Big\|\mathcal{E}^\lambda_{p,\mathcal{Y}}\Big\|_{L^p(\pi)} \Big| = \Big|\mathrm{size}^\lambda_p(\mathcal{X}) - \mathrm{size}^\lambda_p(\mathcal{Y})\Big|,$$

where

$$\|l_\lambda[d_X(x,\cdot)] - l_\lambda[d_Y(y,\cdot)]\|_{L^p(\pi)} \geq \Big| \|l_\lambda[d_X(x,\cdot)]\|_{L^p(\pi)} - \|l_\lambda[d_Y(y,\cdot)]\|_{L^p(\pi)} \Big| = \Big| \mathcal{E}^\lambda_{p,\mathcal{X}}(x) - \mathcal{E}^\lambda_{p,\mathcal{Y}}(y) \Big|,$$

again due to Minkowski's inequality. Hence, using (4.39) we conclude the proof.

Finally, to show (II), observe that an optimal $\pi \in \Pi(\mu,\nu)$ satisfies $\pi \otimes \pi \in \Pi(\mu^{\otimes 2}, \nu^{\otimes 2})$. As such,

$$d_{\mathrm{LRGW}}(X,Y) \geq \inf_{\omega \in \Pi(\mu^{\otimes 2}, \nu^{\otimes 2})} \|l_\lambda(d_X) - l_\lambda(d_Y)\|_{L^p(\omega)},$$

where assuming $f = l_\lambda(d_X)$ and $g = l_\lambda(d_Y)$ proves the bound.

4.6.7 Proof of Theorem 4.2

The proof follows the decomposition of the GW cost due to Zhang et al. (2024). Recall the decomposition of the squared LRIGW cost: $d^2_{\mathrm{LRIGW}}(\mu,\nu) = F_1 + F_2$, where $F_2(\mu,\nu;\lambda) = \inf_{\pi\in\Pi(\mu,\nu)} -2\int l_\lambda(\langle x,x'\rangle)\, l_\lambda(\langle y,y'\rangle) d\pi\otimes\pi(x,y,x',y')$.

Now, for all $x, x' \sim \mu \in \mathcal{P}_4(\mathbb{R}^d_{\geq 0})$

$$l_\lambda(\langle x,x'\rangle) = l_\lambda\Big(\sum_{i=1}^d x_i x_i'\Big) \geq \sum_{i=1}^d l_{\frac{\lambda}{d}}(x_i x_i') \geq \sum_{i=1}^d l_{\sqrt{\frac{\lambda}{d}}}(x_i) l_{\sqrt{\frac{\lambda}{d}}}(x_i') = \Big\langle l_{\sqrt{\frac{\lambda}{d}}}(x), l_{\sqrt{\frac{\lambda}{d}}}(x') \Big\rangle.$$

In the last step, the function l_λ applies componentwise. Similarly, the inequality $l_\lambda(\langle y,y'\rangle) \geq \sum_{j=1}^{d'} l_{\sqrt{\lambda/d'}}(y_j) l_{\sqrt{\lambda/d'}}(y_j')$ holds for all $y, y' \sim \nu \in \mathcal{P}_4(\mathbb{R}^{d'}_{\geq 0})$. Let us generalize by defining $M^{(\lambda,\lambda')}_{\mu,\nu} \coloneqq \sqrt{M_2(\mu;\lambda) M_2(\nu;\lambda')}$, where $M_2(\rho;\lambda) = \int \|l_\lambda(x)\|^2 d\rho(x)$ for any ρ. Also, let $\mathcal{D}_{M^{(\lambda,\lambda')}_{\mu,\nu}} \coloneqq [0, M^{(\lambda,\lambda')}_{\mu,\nu}/2]^{d\times d'}$. Hence,

$$F_2 \leq \inf_{\pi\in\Pi(\mu,\nu)} -2 \sum_{\substack{1\leq i\leq d\\ 1\leq j\leq d'}} \Big(\int l_{\sqrt{\frac{\lambda}{d}}}(x_i) l_{\sqrt{\frac{\lambda}{d'}}}(y_j) d\pi(x,y) \Big)^2 \tag{4.40}$$

$$= \inf_{\pi\in\Pi(\mu,\nu)} \sum_{\substack{1\leq i\leq d\\ 1\leq j\leq d'}} \inf_{0\leq a_{ij}\leq \frac{M^{\bar\lambda}_{\mu,\nu}}{2}} 8\Big(a_{ij}^2 - \int a_{ij} l_{\sqrt{\frac{\lambda}{d}}}(x_i) l_{\sqrt{\frac{\lambda}{d'}}}(y_j) d\pi(x,y) \Big) \tag{4.41}$$

$$= \inf_{\mathbf{A}\in\mathcal{D}_{M^{\bar\lambda}_{\mu,\nu}}} \inf_{\pi\in\Pi(\mu,\nu)} \sum_{\substack{1\leq i\leq d\\ 1\leq j\leq d'}} 8\Big(a_{ij}^2 - \int a_{ij} l_{\sqrt{\frac{\lambda}{d}}}(x_i) l_{\sqrt{\frac{\lambda}{d'}}}(y_j) d\pi(x,y) \Big)$$

$$= \inf_{\mathbf{A}\in\mathcal{D}_{M^{\bar\lambda}_{\mu,\nu}}} 8\|\mathbf{A}\|_F^2 + \inf_{\pi\in\Pi(\mu,\nu)} \int c^{\bar\lambda}_{\mathbf{A}}(x,y) d\pi(x,y),$$

where $\bar\lambda = (\sqrt{\lambda/d}, \sqrt{\lambda/d'})$ and $c^{\bar\lambda}_{\mathbf{A}} : (x,y) \in \mathbb{R}^d_{\geq 0} \times \mathbb{R}^{d'}_{\geq 0} \mapsto -8 l_{\sqrt{\lambda/d}}(x)^T \mathbf{A} l_{\sqrt{\lambda/d'}}(y)$. The optimization in $\mathbf{A}$ can be made unconstrained, however, the optimal a_{ij} in (4.41) is achieved

at $\frac{1}{2}\int l_{\sqrt{\lambda/d}}(x_i) l_{\sqrt{\lambda/d'}}(y_j) d\pi(x,y) \in [0, M^{\bar{\lambda}}_{\mu,\nu}/2]$ (due to Cauchy-Schwarz inequality), which enables us to restrict the optimization to $\mathcal{D}_{M^{\bar{\lambda}}_{\mu,\nu}}$.

A simple parametrization can result in uniform λ-thresholding over the two spaces. Specifically, for the threshold $(d \vee d')\lambda^2$, we achieve the desired upper bound to F_2, as in (4.40), satisfying

$$\bar{F}_2(\mu,\nu;(d\vee d')\lambda^2) = \inf_{\mathbf{A}\in\mathcal{D}_{M^{\lambda}_{\mu,\nu}}} 8\|\mathbf{A}\|_F^2 + \mathrm{OT}_{c^{\lambda}_{\mathbf{A}}}(\mu,\nu). \tag{4.42}$$

Given an optimal coupling $\pi^*_{\mathbf{A}}$ for $\mathrm{OT}_{c^{\lambda}_{\mathbf{A}}}$, a solution $\mathbf{A}^*$ achieving the infimum in (4.42) can be expressed as $\mathbf{A}^* = \frac{1}{2}\int l_\lambda(x) l_\lambda(y)^T d\pi^*_{\mathbf{A}^*}(x,y)$. The associated optimal value of the upper bound becomes

$$\bar{F}_2 = -2\int \Big\langle l_\lambda(x), l_\lambda(x')\Big\rangle \Big\langle l_\lambda(y), l_\lambda(y')\Big\rangle d\pi^*_{\mathbf{A}^*}\otimes\pi^*_{\mathbf{A}^*}(x,y,x',y').$$

4.6.8 Relation between W_p^ϵ and truncated OT

Given $\alpha,\beta \in \mathcal{P}(X)$ such that X is compact, the dual formulation of W_p^ϵ for $p\in[1,\infty)$ becomes (Nietert et al. (2022), Theorem 2)

$$(1-\epsilon)W_p^\epsilon(\alpha,\beta)^p = \sup_{\phi\in C_b(X)} \int \phi\, d\alpha - \int \phi^c\, d\beta - \epsilon\,\mathrm{Range}(\phi), \tag{4.43}$$

where $C_b(X) \coloneqq \{f: X\to\mathbb{R} : f \text{ is continuous, } \|f\|_\infty < \infty\}$ and ϕ^c denotes the c-transform of ϕ w.r.t. the cost $d_X(\cdot,\cdot)^p$. On the other hand, the Kantorovich potential ϕ that solves $\mathrm{OT}_{l_\lambda(d_X)}(\alpha,\beta) \coloneqq \inf_{\pi\in\Pi(\alpha,\beta)}\int_{X\times X}\min\{d_X(x,y),\lambda\}^p d\pi(x,y)$ is a solution to the dual

$$\sup_{\substack{\phi\in C_b(X):\\ \mathrm{Range}(\phi)\le\lambda}} \int \phi\, d\alpha - \int \phi^c\, d\beta, \tag{4.44}$$

such that $\phi^c = \phi$ (Ma et al. (2023), Theorem 2.1), $\lambda > 0$. The latter is a constrained formulation of the regularized dual (4.43). Given any tolerable margin $\lambda < \infty$ on the range of potentials, the solution to (4.44) satisfies (4.43).

4.6.9 Proof of Proposition 4.4

(i) The result is a direct consequence of the fact $W_p^0 = W_p$.
(ii) For $0 \le \epsilon \le \epsilon' \le 1$,

$$d_{\mathrm{LRGW}}(\mu,\nu;\epsilon) = \inf_{\pi\in\Pi(\mu,\nu)} \big\|W_p^\epsilon(x,x') - W_p(y,y')\big\|_{L^p(\pi\otimes\pi)}. \tag{4.45}$$

Now, for $(x, y), (x', y') \sim \mu \otimes \nu$

$$\left|W_p^{\epsilon}(x, x') - W_p(y, y')\right| \leq \left|W_p^{\epsilon}(x, x') - W_p^{\epsilon'}(x, x')\right| + \left|W_p^{\epsilon'}(x, x') - W_p(y, y')\right|. \tag{4.46}$$

Due to the dual form (see, Appendix 4.6.8), given any $\alpha, \beta \in \mathcal{P}(X)$ we may write

$$\begin{aligned}(1-\epsilon)W_p^{\epsilon}(\alpha, \beta)^p &= \sup_{\phi \in C_b(X)} \int \phi \, d\alpha - \int \phi^c \, d\beta - \epsilon \, \text{Range}(\phi) \\ &\leq \sup_{\phi \in C_b(X)} \int \phi \, d\alpha - \int \phi^c \, d\beta - \epsilon' \, \text{Range}(\phi) + 2(\epsilon' - \epsilon)\|\phi\|_{\infty}.\end{aligned}$$

Since $\phi \in C_b(X)$, $\exists K > 0$ such that $W_p^{\epsilon}(\alpha, \beta) - \left(\frac{1-\epsilon'}{1-\epsilon}\right)^{\frac{1}{p}} W_p^{\epsilon'}(\alpha, \beta) \leq K\left(\frac{\epsilon'-\epsilon}{1-\epsilon}\right)^{\frac{1}{p}}$. As such,

$$\begin{aligned}0 &\leq W_p^{\epsilon}(x, x') - W_p^{\epsilon'}(x, x') \\ &\leq \left[\left(\frac{1-\epsilon'}{1-\epsilon}\right)^{\frac{1}{p}} - 1\right] W_p^{\epsilon'}(x, x') + K\left(\frac{\epsilon'-\epsilon}{1-\epsilon}\right)^{\frac{1}{p}} \\ &\leq \left[\left(\frac{1-\epsilon'}{1-\epsilon}\right)^{\frac{1}{p}} - 1\right] \underline{W}_{p,\mu}^{\epsilon'} + K\left(\frac{\epsilon'-\epsilon}{1-\epsilon}\right)^{\frac{1}{p}},\end{aligned} \tag{4.47}$$

where $\underline{W}_{p,\mu}^{\epsilon'} = \inf_{x,x' \sim \mu} W_p^{\epsilon'}(x, x')$, which may only be non-zero given a sample problem. The last inequality, along with (4.46) and the triangle inequality of L^p norms implies that

$$d_{\text{LRGW}}(\mu, \nu; \epsilon) - d_{\text{LRGW}}(\mu, \nu; \epsilon') \leq \left[\left(\frac{1-\epsilon'}{1-\epsilon}\right)^{\frac{1}{p}} - 1\right] \underline{W}_{p,\mu}^{\epsilon'} + K\left(\frac{\epsilon'-\epsilon}{1-\epsilon}\right)^{\frac{1}{p}}.$$

Since (4.46) holds both ways (in ϵ, ϵ'), invoking the trivial upper bound to (4.47) we obtain

$$\left|d_{\text{LRGW}}(\mu, \nu; \epsilon) - d_{\text{LRGW}}(\mu, \nu; \epsilon')\right| \lesssim \left(\frac{\epsilon'-\epsilon}{1-\epsilon}\right)^{\frac{1}{p}}.$$

4.6.10 Proof of Proposition 4.5

The proof follows as a modification to Huber (1981), Corollary 4.3. Given any $\pi \in \Pi(\alpha, \beta)$, observe that

$$\begin{aligned}\mathbb{E}_{\pi}[l_{\lambda}(d_X(x, y))] &\leq \varepsilon \, \mathbb{P}(l_{\lambda}(d_X(x, y)) \leq \varepsilon) + \lambda \, \mathbb{P}(l_{\lambda}(d_X(x, y)) > \varepsilon) \\ &= \varepsilon + (\lambda - \varepsilon) \, \mathbb{P}(l_{\lambda}(d_X(x, y)) > \varepsilon).\end{aligned}$$

Now, consider $\varepsilon = \hat{\rho}_{\lambda}(\alpha, \beta)$. As such there exists π such that $\pi(\{(x, y) : l_{\lambda}(d_X(x, y)) > \varepsilon\}) \leq \varepsilon$. Thus

$$\mathbb{E}_{\pi}[l_{\lambda}(d_X(x, y))] \leq \varepsilon + (\lambda - \varepsilon)\varepsilon \leq (1 + \lambda)\varepsilon.$$

Taking infimum over all couplings result in $\frac{1}{1+\lambda}W_{1,\lambda} \leq \hat{\rho}_\lambda$. To show the upper bound, by using Markov's inequality, we write

$$\mathbb{P}(l_\lambda(d_X(x,y)) > \varepsilon) \leq \frac{\mathbb{E}_\pi[l_\lambda(d_X(x,y))]}{\varepsilon},$$

where $\pi \in \Pi(\alpha,\beta)$ is a feasible optimal coupling for $W_{1,\lambda}$. Observe that we can always choose $\varepsilon^2 = W_{1,\lambda}$. As such, $\hat{\rho}_\lambda \leq \sqrt{W_{1,\lambda}}$.

4.6.11 Proof of Theorem 4.3

Before proving the first inequality, recall that given two mm spaces $(\mathcal{X}, d_X, \mu)$ and $(\mathcal{Y}, d_Y, \nu)$, a metric $\tilde{d}$ on $\mathcal{X} \sqcup \mathcal{Y}$ (disjoint union) is said to be a coupling of d_X and d_Y if and only if $\tilde{d}(x,x') = d_X(x,x')$ and $\tilde{d}(y,y') = d_Y(y,y')$ hold for all $x, x' \in \mathcal{X}$, $y, y' \in \mathcal{Y}$. Let us denote by $\mathscr{D}(d_X, d_Y)$ the collection of all such couplings.

Now, given $(x,y),(x',y') \sim \mu \otimes \nu$ and $\lambda > 0$, we have $|l_\lambda(d_X(x,x')) - l_\lambda(d_Y(y,y'))| \leq l_\lambda(\tilde{d}(x,y)) + l_\lambda(\tilde{d}(x',y'))$. Hence, for $p \in [1,\infty)$

$$J_p^\lambda(\pi) \coloneqq \|l_\lambda(d_X) - l_\lambda(d_Y)\|_{L^p(\pi\otimes\pi)} \leq 2||l_\lambda(\tilde{d})||_{L^p(\pi)}$$

hold for all $\pi \in \Pi(\mu,\nu)$. Hence, the inequality. To show the upper bound, we require some additional definitions.

Definition 4.3 (Modulus of trimmed mass distribution)**.** *For $\delta \geq 0$, the modulus of λ-trimmed mass distribution of μ, having full support is defined as*

$$v_\delta^\lambda(\mu) \coloneqq \inf\{\varepsilon > 0 : \mu(\{x \in \mathcal{X} : \mu(B_X^\lambda(x,\varepsilon)) \leq \delta\}) \leq \varepsilon\},$$

where $B_X^\lambda(x,\varepsilon) = \{y \in \mathcal{X} : l_\lambda(d_X(x,y)) < \varepsilon\}$ is the open ball of λ-trimmed radius $\varepsilon > 0$ around $x \in \mathcal{X}$.

Here, we uniquely consider μ to be a probability measure. Observe that only when $\varepsilon < \lambda$, we get $B_X^\lambda(x,\varepsilon) = B_X(x,\varepsilon)$: the usual open ball around x. Otherwise, the ball becomes the entire $\mathcal{X}$. As such, we only account for the 'thin points' residing in the trimmed support. $v_\delta^\lambda(\mu)$ is essentially equal to $\min\{\lambda, v_\delta(\mu)\}$, where v_δ is the modulus under the metric d_X. This preserves the continuity in the sense that $v_\delta^\lambda(\mu) \xrightarrow{\delta\to 0} 0$ (Greven et al. (2009), lemma 6.5). The relation also implies that an effective trimming requires $\lambda \leq 1$ for probability measures. The proof for the upper bound requires showing a similar statement to lemma 10.3 under the altered metrics $l_\lambda(d_X)$ and $l_\lambda(d_Y)$.

Step 1 *(Construction of ε-nets):* Let $\delta \in (0, \frac{1}{2})$, and $\pi \in \Pi(\mu,\nu)$ such that $J_p^\lambda(\pi) < \delta^5$. Since the altered metric space $(\mathcal{X}, l_\lambda(d_X))$ also contains a maximal 2ε-separated net for any $\varepsilon \geq 0$, we have the following statement.

Lemma 4.4 (Greven et al. (2009), lemma 6.9)**.** *Given $\delta > 0$ and $v_\delta^\lambda(\mu) < \varepsilon$, there exist points $\{x_i\}_{i=1}^N \in \mathcal{X}$ with $N \leq \lfloor \frac{1}{\delta} \rfloor$ such that $\mu(B_X^\lambda(x_i, \varepsilon)) > \delta$, and $\mu(\bigcup_{i=1}^N B_X^\lambda(x, 2\varepsilon)) > 1-\varepsilon$, $\forall\, i = 1, \cdots, N$. Also, for all $i \neq j = 1, \cdots, N$, $l_\lambda(d_X(x_i, x_j)) > \varepsilon$.*

Observe that, since the effective range of permissible ε remains $(0, \lambda)$, a maximal net of $(\mathcal{X}, d_X)$ may be a feasible candidate satisfying the lemma. Beyond the range, the argument becomes trivial.

Now, set $\varepsilon = 4v_\delta^\lambda(\mu)$. As such, we can find a set of points $\{x_i\}_{i=1}^N \in \mathcal{X}$ which ensures $\mu(\bigcup_{i=1}^N B_X^\lambda(x, \varepsilon)) > 1 - \varepsilon$ with $l_\lambda(d_X(x_i, x_j)) > \varepsilon/2$ for all $i \neq j = 1, \cdots, N$. The set may contain arbitrarily far lying observations, yet the argument holds until $\lambda > \varepsilon/2$. Thus, following Mémoli (2011), Claim 10.2 we argue that $\forall\, i = 1, \cdots, N\ \exists\, y_i \in \mathcal{Y}$ such that

$$\pi\Big(B_X^\lambda(x_i, \varepsilon) \times B_Y^\lambda(y_i, 2(\varepsilon + \delta))\Big) \geq (1 - \delta^2)\mu(B_X^\lambda(x_i, \varepsilon)) > \delta(1 - \delta^2). \tag{4.48}$$

Observe that if $\lambda < 2(\varepsilon+\delta)$, the first inequality holds trivially. We denote by $S = \{(x_i, y_i), i = 1, \cdots, N\} \subset \mathcal{X} \times \mathcal{Y}$ the set of points constructing the nets.

Step 2 *(Bounding locally robust distortions):* Consider $\{x_i\}_{i=1}^N$ and $\{y_i\}_{i=1}^N$ that satisfy (4.48) with $\mu(B_X^\lambda(x_i, \varepsilon)) > \delta$. Then, for all $i, j = 1, \cdots, N$

$$|l_\lambda(d_X(x_i, x_j)) - l_\lambda(d_Y(y_i, y_j))| \leq 6(\varepsilon + \delta).$$

To prove the claim, let us first assume that there exists a pair (i, j) for which it does not hold. As such, for all $x' \in B_X^\lambda(x_i, \varepsilon)$, $x'' \in B_X^\lambda(x_j, \varepsilon)$, $y' \in B_Y^\lambda(y_i, 2(\varepsilon+\delta))$, and $y'' \in B_Y^\lambda(y_j, 2(\varepsilon+\delta))$ we have

$$\begin{aligned}
&\big|l_\lambda(d_X(x', x'')) - l_\lambda(d_Y(y', y''))\big| \\
&\geq |l_\lambda(d_X(x_i, x_j)) - l_\lambda(d_Y(y_i, y_j))| - \big|l_\lambda(d_X(x_i, x_j)) - l_\lambda(d_X(x', x''))\big| \\
&\qquad\qquad - \big|l_\lambda(d_Y(y', y'')) - l_\lambda(d_Y(y_i, y_j))\big| \\
&\geq 6(\varepsilon + \delta) - 3\varepsilon - 4(\varepsilon + \delta) = 2\delta.
\end{aligned}$$

Then,

$$\begin{aligned}
J_1^\lambda(\pi) &\geq 2\delta\, \pi\Big(B_X^\lambda(x_i, \varepsilon) \times B_Y^\lambda(y_i, 2(\varepsilon + \delta))\Big)\pi\Big(B_X^\lambda(x_j, \varepsilon) \times B_Y^\lambda(y_j, 2(\varepsilon + \delta))\Big) \\
&\geq 2\delta^3(1 - \delta^2)^2 > 2\delta^5,
\end{aligned}$$

since $\delta \leq \frac{1}{2}$. This contradicts our initial assumption.

Step 3 *(Constructing a suitable metric S):* Define $\tilde{d}_S^\lambda$ on $\mathcal{X} \sqcup \mathcal{Y}$ as

$$(x, y) \mapsto \inf_{(x', y') \in S} \Big[l_\lambda(d_X(x, x')) + \|l_\lambda(d_X) - l_\lambda(d_Y)\|_{L^\infty(S \times S)} + l_\lambda(d_Y(y, y'))\Big],$$

also assuming $\tilde{d}_S^\lambda = l_\lambda(d_X)$ on $\mathcal{X}\times\mathcal{X}$ and $\tilde{d}_S^\lambda = l_\lambda(d_Y)$ on $\mathcal{Y}\times\mathcal{Y}$. Using Step 2, we get

$$\tilde{d}_S^\lambda(x,y) \leq 2\lambda + 6(\varepsilon+\delta). \tag{4.49}$$

However, for $i = 1,\cdots,N$, given $(x,y) \in B_X^\lambda(x_i,\varepsilon)\times B_Y^\lambda(y_i, 2(\varepsilon+\delta))$ we have

$$\begin{aligned}\tilde{d}_S^\lambda(x,y) &\leq \varepsilon + 2(\varepsilon+\delta) + \tilde{d}_S^\lambda(x_i,y_i)\\ &\leq \varepsilon + 8(\varepsilon+\delta),\end{aligned} \tag{4.50}$$

where the last inequality is due to Mémoli (2011), lemma 10.1. Now, in pursuit of fragmenting $\mathcal{X}\times\mathcal{Y}$ based on balls around the points that constitute the maximal net, define $L = \bigcup_{i=1}^N B_X^\lambda(x_i,\varepsilon)\times B_Y^\lambda(y_i,2(\varepsilon+\delta))$. Hence,

$$\begin{aligned}\int_{\mathcal{X}\times\mathcal{Y}} [\tilde{d}_S^\lambda(x,y)]^p d\pi(x,y) &= \int_L [\tilde{d}_S^\lambda(x,y)]^p d\pi(x,y) + \int_{\mathcal{X}\times\mathcal{Y}\setminus L} [\tilde{d}_S^\lambda(x,y)]^p d\pi(x,y)\\ &\leq (9(\varepsilon+\delta))^p + (\varepsilon+\delta)[2\lambda + 6(\varepsilon+\delta)]^p,\end{aligned}$$

where the last inequality is due to (4.49), (4.50) and the fact that $\pi(\mathcal{X}\times\mathcal{Y}\setminus L) \leq \varepsilon+\delta$ (Mémoli (2011), Claim 10.5), which is obvious if $\lambda < 2(\varepsilon+\delta)$. As such, for $p\in[1,\infty)$

$$d_{l\text{RSGW}}(\mu,\nu) \leq (4\lambda+\delta)^{\frac{1}{p}}(62\lambda + \frac{15}{2}),$$

since $\varepsilon \leq 4\lambda$ and $\delta \leq \frac{1}{2}$.

4.6.12 Sample Complexity of Transform Sampling Using Tukey and LR-guided RGM Under Contamination

While the formulation (4.23) proposes altering the set of amenable couplings, based on our discussion in the first two sections, it is quite intuitive to think of a formulation that rather penalizes the norms. For example, we may construct a robust transform sampler drawing from both LR and Tukey's robustification techniques as follows:

$$\inf_{F,G} \int \mathcal{T}_2\Big(d_X(x,G(y)) - d_Y(y,F(x))\Big) d\mu\otimes\nu(x,y) + \lambda_1 W_1^\lambda(\mu, G_\#\nu) + \lambda_2 W_1^\lambda(F_\#\mu,\nu), \tag{4.51}$$

where the parameter underlying $\mathcal{T}_2$ is $\tau \geq 0$ and $\lambda \geq 0$. Typically, in an empirical problem, both τ,λ need to be tuned, and the infimum is taken over arbitrary measurable maps F, G between spaces $\mathcal{X} \rightleftarrows \mathcal{Y}$. They need not follow the binding constraint, as the Lagrangian conditions ensure their measure preservation only. We assume that they are continuous and component-wise uniformly bounded in the following sense.

Assumption 4.1. *Given that $\mathcal{X}\subseteq\mathbb{R}^d$ and $\mathcal{Y}\subseteq\mathbb{R}^{d'}$, let $F_k(\cdot):\mathcal{X}\to\mathbb{R}$ (similarly, $G_l(\cdot)$:*

$\mathcal{Y} \to \mathbb{R}$) denote the kth (and lth, $l = 1, \cdots, d$) coordinate of F, which is continuous (similarly, G), $k = 1, \cdots, d'$. There exists $b > 0$ such that for all $k = 1, \cdots, d'$, and $l = 1, \cdots, d$

$$\sup_F \|F_k\|_\infty, \sup_G \|G_l\|_\infty \leq b.$$

We denote such classes of functions as $\mathcal{F}_b^{\mathcal{X}\to\mathcal{Y}}$ and $\mathcal{F}_b^{\mathcal{Y}\to\mathcal{X}}$ respectively.

4.6.12.1 Robust Concentration Inequalities

Given a pair of feasible maps $(F, G) \in \mathcal{F}_b^{\mathcal{X}\to\mathcal{Y}} \times \mathcal{F}_b^{\mathcal{Y}\to\mathcal{X}}$ we observe the non-asymptotic deviation of realized values of (4.51) from a robust population benchmark. We assume, without loss of generality, that $\lambda_1 = \lambda_2 = 1$. Also, let $x_1, x_2, \cdots, x_m$ are sampled following the $\mathcal{O} \cup \mathcal{I}$ framework with μ being the inlier distribution. Similarly, we have $y_1, y_2, \cdots, y_n$ from the other space with inliers drawn i.i.d from ν. The inlying (outlying) set of samples are indexed using $\mathcal{I}^X$ and $\mathcal{I}^Y$ ($\mathcal{O}^X$ and $\mathcal{O}^Y$) respectively. Let us denote

$$T(\mu, \nu, F, G) \coloneqq \int \mathcal{T}_2\Big(d_X(x, G(y)) - d_Y(y, F(x))\Big) d\mu \otimes \nu(x, y),$$
$$L(\mu, \nu, F, G) \coloneqq W_1^\lambda(\mu, G_\# \nu) + W_1^\lambda(F_\# \mu, \nu).$$

As such, the population loss function in (4.51) can be written as $C(\mu, \nu, F, G) = T(F, G) + L(F, G)$. The empirical loss under contaminated measures $\hat{\mu}_m, \hat{\nu}_n$ is given as $T(\hat{\mu}_m, \hat{\nu}_n, F, G)$. However, to emphasize the robust translations we write the empirical version of $L(\mu, \nu, F, G)$ as $L(\hat{\mu}_m, \hat{\nu}_n, F, G) \coloneqq W_1^\lambda(\mu, G_\# \hat{\nu}_n) + W_1^\lambda(F_\# \hat{\mu}_m, \nu)$. The following two results combined, present the concentration of $C(\hat{\mu}_m, \hat{\nu}_n, F, G)$ around

$$\frac{|\mathcal{I}^X||\mathcal{I}^Y|}{mn} \mathbb{E}_{\mu\otimes\nu}[T(\hat{\mu}_m^{\mathcal{I}}, \hat{\nu}_n^{\mathcal{I}}, F, G)] + \frac{|\mathcal{I}^Y|}{n} \mathbb{E}_\nu[W_1^\lambda(\mu, G_\# \hat{\nu}_n^{\mathcal{I}})] + \frac{|\mathcal{I}^X|}{m} \mathbb{E}_\mu[W_1^\lambda(F_\# \hat{\mu}_m^{\mathcal{I}}, \nu)].$$

Proposition 4.6. *There exists a constant $K > 0$ depending on τ^2 such that for $\delta > 0$*

$$\left| T(\hat{\mu}_m, \hat{\nu}_n, F, G) - \frac{|\mathcal{I}^X||\mathcal{I}^Y|}{mn} \mathbb{E}_{\mu\otimes\nu}[T(\hat{\mu}_m^{\mathcal{I}}, \hat{\nu}_n^{\mathcal{I}}, F, G)] \right| \leq K \frac{|\mathcal{I}^X||\mathcal{I}^Y|}{mn} \sqrt{\frac{\ln\left(\frac{4(|\mathcal{I}^X| \vee |\mathcal{I}^Y|)}{\delta}\right)}{|\mathcal{I}^X| \wedge |\mathcal{I}^Y|}} + \tau^2\Big(\frac{|\mathcal{O}^X|}{m} + \frac{|\mathcal{O}^Y|}{n}\Big)$$

holds with probability at least $1 - \delta$.

Proof of Proposition 4.6

Let us denote $t_{F,G}(x,y) \coloneqq \mathcal{T}_2\Big(d_X(x,G(y)) - d_Y(y,F(x))\Big)$. Now, following the $\mathcal{O}\cup\mathcal{I}$ setup

$$\begin{aligned}
T(\hat{\mu}_m,\hat{\nu}_n,F,G) &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathcal{T}_2\Big(d_X(x_i,G(y_j)) - d_Y(y_j,F(x_i))\Big)\\
&\leq \frac{1}{mn}\sum_{i\in\mathcal{I}^X}\sum_{j\in\mathcal{I}^Y}\mathcal{T}_2\Big(d_X(x_i,G(y_j)) - d_Y(y_j,F(x_i))\Big)\\
&\quad + \frac{\tau^2}{mn}\Big(|\mathcal{O}^X||\mathcal{O}^Y| + |\mathcal{O}^X||\mathcal{I}^Y| + |\mathcal{I}^X||\mathcal{O}^Y|\Big)\\
&= \frac{|\mathcal{I}^X||\mathcal{I}^Y|}{mn}T(\hat{\mu}^{\mathcal{I}}_m,\hat{\nu}^{\mathcal{I}}_n,F,G) + \tau^2\Big(\frac{|\mathcal{O}^X|}{m} + \frac{|\mathcal{O}^Y|}{n} - \frac{|\mathcal{O}^X||\mathcal{O}^Y|}{mn}\Big).
\end{aligned}$$

Also,

$$\frac{|\mathcal{I}^X||\mathcal{I}^Y|}{mn}T(\hat{\mu}^{\mathcal{I}}_m,\hat{\nu}^{\mathcal{I}}_n,F,G) - \tau^2\Big(\frac{|\mathcal{O}^X|}{m} + \frac{|\mathcal{O}^Y|}{n} - \frac{|\mathcal{O}^X||\mathcal{O}^Y|}{mn}\Big) \leq T(\hat{\mu}_m,\hat{\nu}_n,F,G).$$

As such, combining the two inequalities we get

$$\begin{aligned}
&\Big|T(\hat{\mu}_m,\hat{\nu}_n,F,G) - \frac{|\mathcal{I}^X||\mathcal{I}^Y|}{mn}\mathbb{E}_{\mu\otimes\nu}[T(\hat{\mu}^{\mathcal{I}}_m,\hat{\nu}^{\mathcal{I}}_n,F,G)]\Big|\\
&\leq \frac{|\mathcal{I}^X||\mathcal{I}^Y|}{mn}\Big|T(\hat{\mu}^{\mathcal{I}}_m,\hat{\nu}^{\mathcal{I}}_n,F,G) - \mathbb{E}_{\mu\otimes\nu}[T(\hat{\mu}^{\mathcal{I}}_m,\hat{\nu}^{\mathcal{I}}_n,F,G)]\Big| + \tau^2\Big(\frac{|\mathcal{O}^X|}{m} + \frac{|\mathcal{O}^Y|}{n} - \frac{|\mathcal{O}^X||\mathcal{O}^Y|}{mn}\Big).
\end{aligned} \tag{4.52}$$

Now, observe that

$$\begin{aligned}
&\Big|T(\hat{\mu}^{\mathcal{I}}_m,\hat{\nu}^{\mathcal{I}}_n,F,G) - \mathbb{E}_{\mu\otimes\nu}[T(\hat{\mu}^{\mathcal{I}}_m,\hat{\nu}^{\mathcal{I}}_n,F,G)]\Big|\\
&= \left|\frac{1}{|\mathcal{I}^X||\mathcal{I}^Y|}\sum_{i\in\mathcal{I}^X}\sum_{j\in\mathcal{I}^Y}t_{F,G}(x_i,y_j) - \mathbb{E}_{\mu\otimes\nu}[t_{F,G}(x,y)]\right|\\
&\leq \left|\frac{1}{|\mathcal{I}^X|}\sum_{i\in\mathcal{I}^X}\left(\frac{1}{|\mathcal{I}^Y|}\sum_{j\in\mathcal{I}^Y}t_{F,G}(x_i,y_j) - \mathbb{E}_{\nu}[t_{F,G}(x_i,y)]\right)\right| + \left|\frac{1}{|\mathcal{I}^X|}\sum_{i\in\mathcal{I}^X}\mathbb{E}_{\nu}[t_{F,G}(x_i,y)] - \mathbb{E}_{\mu\otimes\nu}[t_{F,G}(x,y)]\right|.
\end{aligned} \tag{4.53}$$

Recall that $M = \operatorname{diam}(\mathcal{X})\vee\operatorname{diam}(\mathcal{Y})$. The function $|\mathcal{I}^X|^{-1}\sum_{i\in\mathcal{I}^X}t_{F,G}(x_i,y)$ satisfies the bounded difference inequality with parameter $|\mathcal{I}^X|^{-1}(4M^2\wedge\tau^2)$. Hence, due to McDiarmid's

inequality

$$\left|\frac{1}{|\mathcal{I}^X|}\sum_{i\in\mathcal{I}^X}\mathbb{E}_\nu[t_{F,G}(x_i,y)]-\mathbb{E}_{\mu\otimes\nu}[t_{F,G}(x,y)]\right|$$
$$\leq \mathbb{E}_\nu\left|\frac{1}{|\mathcal{I}^X|}\sum_{i\in\mathcal{I}^X}t_{F,G}(x_i,y)-\mathbb{E}_\mu[t_{F,G}(x,y)]\right| \leq \sqrt{\frac{(4M^2\wedge\tau^2)^2\ln(2/\delta)}{2|\mathcal{I}^X|}}$$

holds with probability at least $1-\delta$, where the first inequality follows from Jensen's inequality. For the first term in (4.53), using the union bound over a similar argument, we get

$$\left|\frac{1}{|\mathcal{I}^X|}\sum_{i\in\mathcal{I}^X}\left(\frac{1}{|\mathcal{I}^Y|}\sum_{j\in\mathcal{I}^Y}t_{F,G}(x_i,y_j)-\mathbb{E}_\nu[t_{F,G}(x_i,y)]\right)\right| \leq \sqrt{\frac{(4M^2\wedge\tau^2)^2\ln(2|\mathcal{I}^X|/\delta)}{2|\mathcal{I}^Y|}}$$

with probability at least $1-\delta$. As such,

$$\left|T(\hat{\mu}_m^{\mathcal{I}},\hat{\nu}_n^{\mathcal{I}},F,G)-\mathbb{E}_{\mu\otimes\nu}[T(\hat{\mu}_m^{\mathcal{I}},\hat{\nu}_n^{\mathcal{I}},F,G)]\right| \lesssim \sqrt{\frac{\ln\left(\frac{2(|\mathcal{I}^X|\vee|\mathcal{I}^Y|)}{\delta}\right)}{|\mathcal{I}^X|\wedge|\mathcal{I}^Y|}}$$

holds with probability $\geq 1-2\delta$. Hence, putting this back to (4.52) proves the result.

The proof also shows that it is always possible to replace the term $\frac{|\mathcal{I}^X||\mathcal{I}^Y|}{mn}\mathbb{E}_{\mu\otimes\nu}[T(\hat{\mu}_m^{\mathcal{I}},\hat{\nu}_n^{\mathcal{I}},F,G)]$ by $T(\mu,\nu,F,G)$, only by incurring an additional term on the upper bound of $\mathcal{O}\Big(\frac{|\mathcal{O}^X|}{m}+\frac{|\mathcal{O}^Y|}{n}\Big)$.

Proposition 4.7. *There exists a constant $\tilde{K}>0$ depending on λ such that for $\delta>0$*

$$\left|W_1^\lambda(\mu,G_\#\hat{\nu}_n)+W_1^\lambda(F_\#\hat{\mu}_m,\nu)-\frac{|\mathcal{I}^Y|}{n}\mathbb{E}[W_1^\lambda(\mu,G_\#\hat{\nu}_n^{\mathcal{I}})]-\frac{|\mathcal{I}^X|}{m}\mathbb{E}[W_1^\lambda(F_\#\hat{\mu}_m^{\mathcal{I}},\nu)]\right|$$
$$\leq \tilde{K}\Big(\frac{\sqrt{|\mathcal{I}^X|}}{m}+\frac{\sqrt{|\mathcal{I}^Y|}}{n}\Big)\sqrt{\ln(4/\delta)}+\lambda\Big(\frac{|\mathcal{O}^X|}{m}+\frac{|\mathcal{O}^Y|}{n}\Big)$$

holds with probability at least $1-\delta$.

Proof of Proposition 4.7

First, let us note that $F_\#\hat{\mu}_m=\frac{1}{m}\sum_{i=1}^m\delta_{F(x_i)}$ based on the transformed observations $\{F(x_i)\}_{i=1}^m$. Technically, this is rather the empirical distribution $\widehat{(F_\#\mu)}_m$. Our consideration remains valid if F is taken as an information preserving transform (IPT), which is in abundance (e.g., Lipschitz maps) Chakrabarty et al. (2025a). Hence, similar to the proof of Proposition 4.6, we observe

$$\left|W_1^\lambda(F_\#\hat{\mu}_m,\nu)-\frac{|\mathcal{I}^X|}{m}W_1^\lambda(F_\#\hat{\mu}_m^{\mathcal{I}},\nu)\right| \leq \frac{\lambda|\mathcal{O}^X|}{m}.$$

This implies that

$$\left|W_1^\lambda(F_\#\hat{\mu}_m,\nu)-\frac{|\mathcal{I}^X|}{m}\mathbb{E}[W_1^\lambda(F_\#\hat{\mu}_m^{\mathcal{I}},\nu)]\right|\leq\frac{|\mathcal{I}^X|}{m}\left|W_1^\lambda(F_\#\hat{\mu}_m^{\mathcal{I}},\nu)-\mathbb{E}[W_1^\lambda(F_\#\hat{\mu}_m^{\mathcal{I}},\nu)]\right|+\frac{\lambda|\mathcal{O}^X|}{m}. \tag{4.54}$$

Now, using the duality of W_1^λ (see, Appendix 4.6.8) we may write

$$W_1^\lambda(F_\#\hat{\mu}_m^{\mathcal{I}},\nu)=\frac{1}{|\mathcal{I}^X|}\sup_{\substack{\phi\in C_b(\mathcal{Y}):\\ \mathrm{Range}(\phi)\leq\lambda}}\sum_{i\in\mathcal{I}^X}\phi(F(x_i))-\mathbb{E}_\nu\phi^c.$$

Due to Assumption 4.1, the composition $\phi\circ F\in C_{b'}(\mathcal{X})$ for some $b'>0$. As such, $W_1^\lambda(F_\#\hat{\mu}_m,\nu)$ satisfies the bounded differences property with upper bound $\mathcal{O}(\frac{1}{|\mathcal{I}^X|})$. Hence,

$$\left|W_1^\lambda(F_\#\hat{\mu}_m^{\mathcal{I}},\nu)-\mathbb{E}[W_1^\lambda(F_\#\hat{\mu}_m^{\mathcal{I}},\nu)]\right|\lesssim\sqrt{\frac{\ln(2/\delta)}{|\mathcal{I}^X|}}$$

holds with probability at least $1-\delta$. Putting the bound back in (4.54) yields,

$$\left|W_1^\lambda(F_\#\hat{\mu}_m,\nu)-\frac{|\mathcal{I}^X|}{m}\mathbb{E}[W_1^\lambda(F_\#\hat{\mu}_m^{\mathcal{I}},\nu)]\right|\leq K'\sqrt{\frac{|\mathcal{I}^X|}{m}}\sqrt{\frac{\ln(2/\delta)}{m}}+\frac{\lambda|\mathcal{O}^X|}{m},$$

that hold with probability at least $1-\delta$, where $K'>0$ depends on λ. Similarly, we can show that there exists $K''>0$ such that

$$\left|W_1^\lambda(\mu,G_\#\hat{\nu}_n)-\frac{|\mathcal{I}^Y|}{n}\mathbb{E}[W_1^\lambda(\mu,G_\#\hat{\nu}_n^{\mathcal{I}})]\right|\leq K''\sqrt{\frac{|\mathcal{I}^Y|}{n}}\sqrt{\frac{\ln(2/\delta)}{n}}+\frac{\lambda|\mathcal{O}^Y|}{n}$$

also holds with probability $\geq 1-\delta$. Combining the last two bounds proves the result.

Remark 4.6 (Uniform Deviations)**.** *The concentration inequalities make it easier to comment on the uniform deviation* $\sup_{(F,G)\in\mathcal{F}_b^{\mathcal{X}\to\mathcal{Y}}\times\mathcal{F}_b^{\mathcal{Y}\to\mathcal{X}}}|T(\hat{\mu}_m,\hat{\nu}_n,F,G)-T(\mu,\nu,F,G)|$*. Let us assume both d_X and d_Y to be the Euclidean metrics in their respective spaces. Due to (4.52), the problem boils down to finding*

$$\sup_{(F,G)\in\mathcal{F}_b^{\mathcal{X}\to\mathcal{Y}}\times\mathcal{F}_b^{\mathcal{Y}\to\mathcal{X}}}\left|T(\hat{\mu}_m^{\mathcal{I}},\hat{\nu}_n^{\mathcal{I}},F,G)-\mathbb{E}_{\mu\otimes\nu}[T(\hat{\mu}_m^{\mathcal{I}},\hat{\nu}_n^{\mathcal{I}},F,G)]\right|=:T_{m,n}^{\mathcal{I}}.$$

Since the underlying loss depends solely on inlying observations, using Proposition 4.6 of Hur et al. (2024), we obtain for any $\varepsilon>0$

$$T_{m,n}^{\mathcal{I}}\lesssim\sqrt{\frac{\ln\left(\frac{2(|\mathcal{I}^X|\vee|\mathcal{I}^Y|)}{\delta}\right)}{|\mathcal{I}^X|\wedge|\mathcal{I}^Y|}}+\varepsilon+\sqrt{\frac{\sum_{k=1}^{d'}\log N_\infty(\varepsilon,\mathcal{F}_k,|\mathcal{I}^X|)+\sum_{l=1}^{d}\log N_\infty(\varepsilon,\mathcal{G}_l,|\mathcal{I}^Y|)}{|\mathcal{I}^X|\wedge|\mathcal{I}^Y|}},$$

holds with probability $\geq 1-\delta$*, where* $\mathcal{F}_k$ *and* $\mathcal{G}_l$ *are respectively the collections of amenable functions* F_k *and* G_l *satisfying Assumption 4.1. Classes* $\mathcal{F}_k$ *and* $\mathcal{G}_l$ *whose metric entropies scale according to* $\mathcal{O}(1/\varepsilon)^a$*, for some* $a > 0$ *are abundant. For example, Sobolev or Lipschitz-smooth functions defined on the unit interval* $[0,1]^d$*. One can similarly derive uniform deviation bounds corresponding to* $L(\hat{\mu}_m, \hat{\nu}_n, F, G)$ *and hence, eventually* $C(\hat{\mu}_m, \hat{\nu}_n, F, G)$*.*

4.6.13 Existence of latent chaining

While it is difficult to characterize a suitable $\mathcal{Z}$ that follows the chaining argument given arbitrary μ and ν, we give examples that conform to our experiments. Assume that $\mu, \nu \in \mathcal{P}_2^{\text{ac}}(\mathbb{R}^d)$ are fully supported. Moreover, $\mathcal{Z}$ is convex, endowed with an absolutely continuous ω (e.g., Lebesgue). Then, due to Brenier's polar factorization (Brenier, 1991), any transport map $T : \mathcal{Z} \to \mathcal{X}$ can be decomposed as $T = (\nabla\varphi) \circ s$ a.e., where $\varphi : \mathcal{Z} \to \mathbb{R}$ is convex and $s : \mathcal{Z} \to \mathcal{Z}$ is measure-preserving, both uniquely defined a.e. In fact, $(\nabla\varphi)$ is the unique optimal transport map between ω and μ under the Euclidean cost. Hence, we can construct $\phi_X \coloneqq [(\nabla\varphi) \circ s']^{-1}$, where s' is also bijective (see diagram (4.18)). Similarly, define $\phi'_Y \coloneqq (\nabla\varrho) \circ s''$, where $(\nabla\varrho)$ is the OT map between ω and ν, and $s'' : \mathcal{Z} \to \mathcal{Z}$ preserves measure. As such, $F = (\nabla\varrho) \circ s'' \circ [(\nabla\varphi) \circ s']^{-1}$. One can similarly define G.

The same can be extended to $\mathcal{Z}$ (Also, $\mathcal{X}$ and $\mathcal{Y}$) being a connected, compact, C^3-smooth Riemannian Manifold without boundary (McCann (2001), Theorem 11). Then, any volume-preserving transport T is represented as $\exp(-\nabla\varphi) \circ s$ a.e. Based on this, the rest of the construction follows exactly.

4.6.14 Implementation details

We refer to the repository `https://github.com/SankhaSubhra/LRGW/tree/master` for codes corresponding to LR and `https://anonymous.4open.science/r/RCDA/` for the rest along with execution instructions. All experiments were carried out on an RTX 3090 GPU.

4.6.14.1 Parameter selection in TGW and HGW

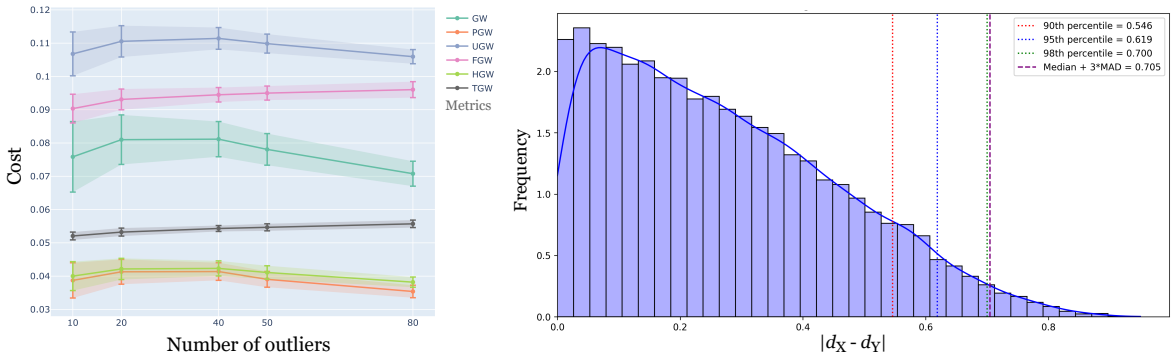
As mentioned before, we select $\tau = \tilde{m} + 3\tilde{\sigma}$, where $\tilde{m}$ and $\tilde{\sigma}$ are respectively the median and the mean deviation about median of the deviation values $J_{X,Y} = |d_X - d_Y|$. Observe that, for an univariate standard folded Normal random variable Z

$$\mathbb{P}(Z \leq \tilde{m}) = 2\Phi(\tilde{m}) - 1 = \frac{1}{2},$$

where $\Phi(\cdot)$ is the distribution function of $N(0,1)$. As such, $\tilde{m} \approx 0.69$, the third quartile of $N(0,1)$. Now, given $a > 0$,

$$\begin{aligned}\tilde{\sigma} = \mathbb{E}\left[\,|Z-a|\right] &= \int_0^\infty |z-a| f(z)dz = \sqrt{\frac{2}{\pi}} \int_0^\infty |z-a| e^{-\frac{z^2}{2}} dz \\ &= \sqrt{\frac{2}{\pi}} \Big(\int_0^a (a-z) e^{-\frac{z^2}{2}} dz + \int_a^\infty (z-a) e^{-\frac{z^2}{2}} dz \Big) \\ &= \sqrt{\frac{2}{\pi}} (2e^{-\frac{a^2}{2}} - 1) + 4a\Phi(a) - 3a.\end{aligned}$$

As such at $a = 0.69$, we have $\tilde{\sigma} \approx 0.46$. The corresponding estimate for τ turns out to be ≈ 2.07, which coincides approximately with the 96-percentile of Z. In our experiments, the deviation between pairwise distances under contamination does not follow such a law. Thus, calculating only a certain percentile becomes insufficient. Given a set of observations, we calculate the statistics independently and only use the percentiles as a reference.

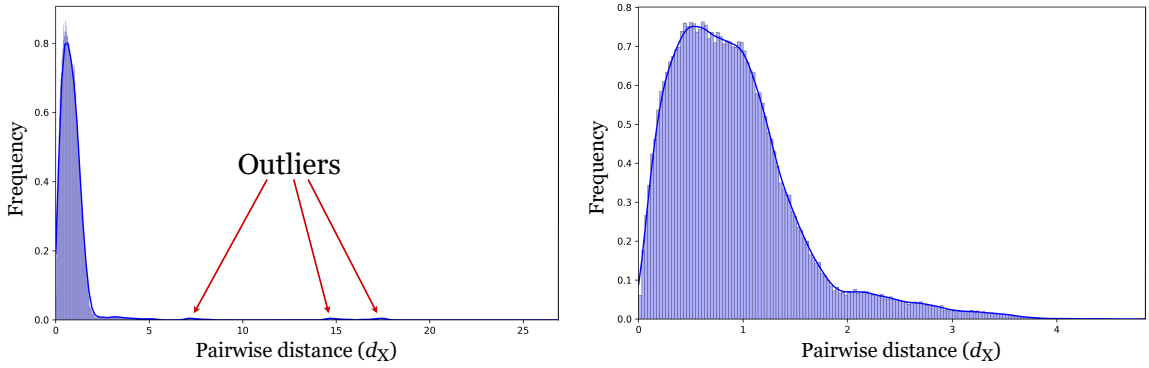


(a) (b)

Figure 4.8: (a) Average losses under increasing proportion of bi-variate standard Gaussian outliers $(0.02, 0.04, 0.08, 0.1, 0.16)$ in source. (b) Empirical distribution of $J_{X,Y}$ under 80 Gaussian outliers. Realized 95-percentile and $\tilde{m} + 3\tilde{\sigma}$ are 0.619 and 0.705 respectively. TGW follows the 95-percentile selection scheme while HGW is calculated based on $\tilde{m} + 3\tilde{\sigma}$.

Performances of the competing methods alter under a contaminating distribution with a thinner tail. While Cauchy implants vastly outlying observations with higher frequency, Gaussian outliers flock to the immediate neighborhood (see Figure 4.2(b)). As a result, we observe a much more pronounced distortion of the shape instead of extremely large distortion values ($J_{X,Y}$). Moreover, as the number of outliers increases, only 'moderately' high-valued distances (d_X) in the source increase (see Figure 4.9). This noising phenomenon is more difficult to eliminate using our thresholding scheme as 'outlying' d_X values are erroneously considered legitimate. As a result, the vanilla GW value, due to averaging, decreases even at elevated contamination levels. This is misleading since it does not reflect the distortion in local geometry. PGW and HGW perform well, and remarkably, TGW not only stays stable

but also exhibits minute increments, indicating intensifying noise. The OT component's increase in FGW also demonstrates the same effect.



(a) Cauchy noise (b) Gaussian noise

Figure 4.9: (a) Empirical density of pairwise distances $d_X(x, x')$ in the source shape (cat) with 40 outliers.

This motivates a rather finer thresholding technique we call local robustification. Instead of trimming extreme distortion values, we shift our focus to individual pairwise distances d_X and d_Y.

4.6.14.2 LR translation using GcGAN and UNIT

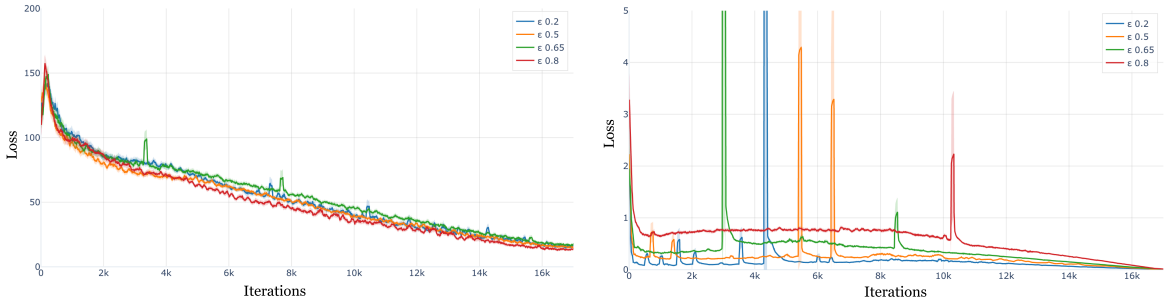
For images, contamination regimes become more complex than point data. In case there are clear outliers ($n\epsilon$ out of n) from other image sources (e.g., MNIST samples in a pool of facial images (Nietert et al., 2023)), the discriminators can still distinguish between them. In contrast, if all images have noise injected in them in a predefined proportion (α), the resultant generations get much more affected. This is mainly due to the discriminators misclassifying them. Observe that if $\alpha = 1$, the noisy image from the second regime becomes an outlier from the first case. We maintain a flexible framework, striking a balance between the two.

For the experiment under GcGAN, first, we create a copy of the original image tensor to ensure its data remains unchanged. Next, we generate standard Gaussian noise with the same shape as the tensor, scaling it by $\alpha = 0.2$ to control the noise intensity. Finally, we add this scaled noise to the copied tensor to produce a noisy version of the original.

In the case of UNIT, we inject random bright pixels following a Gaussian law into the random images based on the image ratio. The wrapper supports datasets with or without labels by handling tuple or single image outputs and seamlessly integrates into existing data loaders.

Ablation study: Parameter selection

Our experimental framework begins with analyzing the generator and discriminator loss propagation (Figure 4.10) under varying values of ϵ. While smaller values imply weaker protection against noise, larger values tend to degrade discrimination performance (Figure 4.10b). Coupled with the quantitative scrutiny, we introduce an additional experiment based on the qualitative outcomes as in Figure 4.11. Here, we observe increased meddling in background color and oversaturation as ϵ increases. On the other hand, a lower parameter value increases the likelihood of noise being manifested in the resulting images. Based on a trade-off between both examinations, we infer that $\epsilon = 0.5$ consistently demonstrates balanced performance during training, prompting us to an optimal value.



(a) Convergence of loss

(b) Discriminator performance

Figure 4.10: (a) Realized robust GcGAN loss for varying ϵ under Gaussian noise ($\alpha = 0.2$). There is no perceptible difference between ϵ values in this regard. (b) The discriminators also eventually perform similarly.

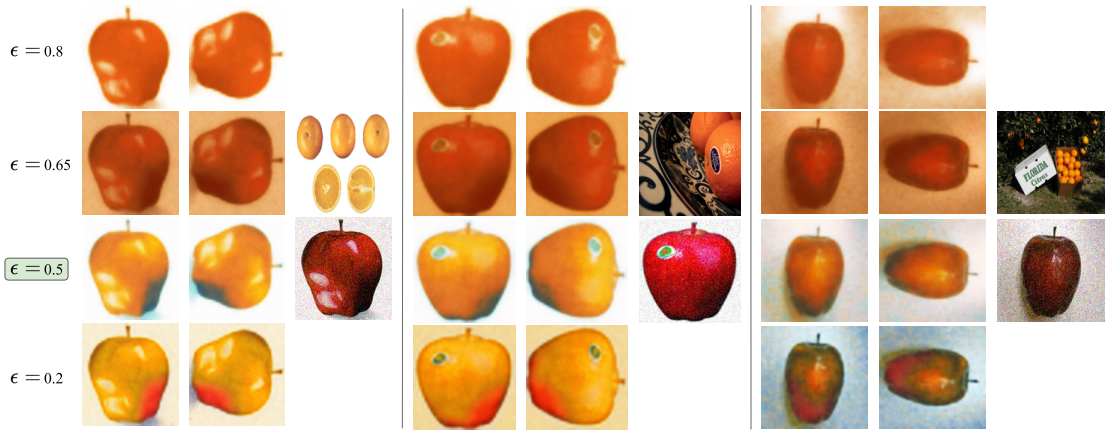


Figure 4.11: Style transfer performance of robust GcGAN for varying ϵ. While small values ($\epsilon = 0.2$) produce inadequate denoising, high values ($\epsilon = 0.8$) distort the style and oversaturate images.

4.6.14.3 LR alignment under Gaussian Contamination

In this section, we place the experimental results corresponding to Gaussian contamination in both domains in LR. Compared to Cauchy outliers, it is much more difficult to find a threshold λ that clearly distinguishes between inlying and outlying samples. This is primarily because standard bi-variate Gaussian noise clouds the shapes locally. As such, the distribution of pairwise distances distorts throughout rather than increasing only extremely large deviations. As a result, often the $\tilde{m}+3\tilde{\sigma}$ criteria produces thresholds close to fixed percentiles. However, the margin of error being much lower under Gaussian contamination, it still produces superior outcomes.

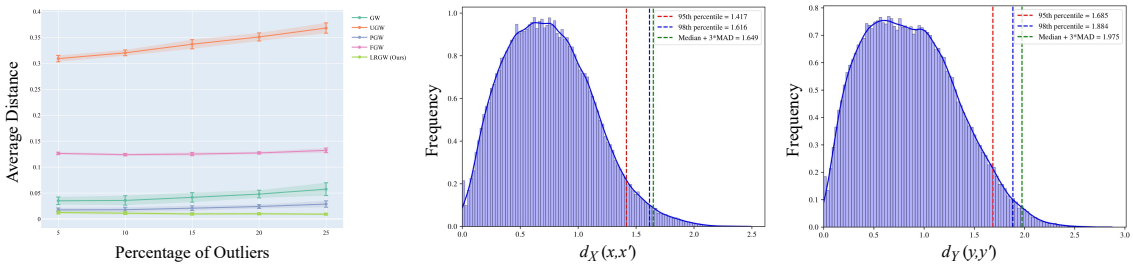


Figure 4.12: (*left*) Average loss under increasing percentage $(5, 10, 15, 20, 25)$ of outlier points in the source domain (cat) drawn independently from bi-variate Gaussian. The target (heart) shape contains 20% Gaussian outliers. (*center*) Empirical distribution of pairwise distances under 20% Gaussian outliers in the target (heart). Realized 98-percentile and $\tilde{m}+3\tilde{\sigma}$ are 1.616 and 1.649 respectively. (*right*) Empirical distribution of pairwise distances under 20% Gaussian outliers in the source (cat). Realized 98-percentile and $\tilde{m}+3\tilde{\sigma}$ are 1.884 and 1.975 respectively.

For example, as seen in Figure 4.13, it is only LRGW that accurately recovers the original (noise-free) GW value ≈ 0.03 during shape matching. We emphasize that the thresholds, being data-dependent and tunable, enable recovering robust surrogates throughout the spectrum of increasing noise in both domains.

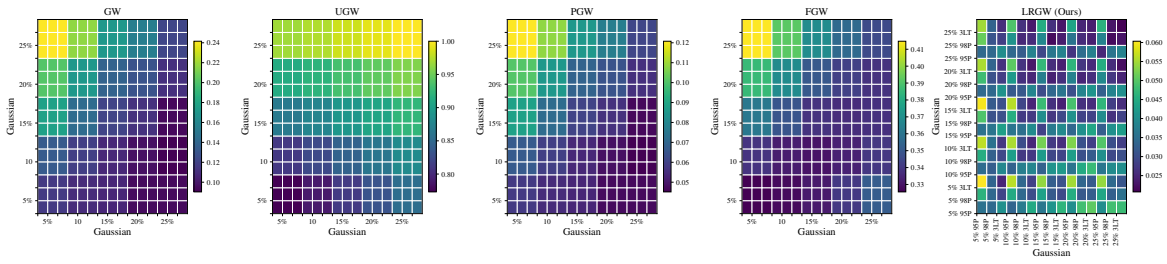


Figure 4.13: Average distances under varying levels of Gaussian contamination in both domains.

The heightened distortion in shapes due to Gaussian outliers also becomes predominant in barycenters. However, as pointed out in Figure 4.14, LRGW improves over GW by preserving underlying shapes better.

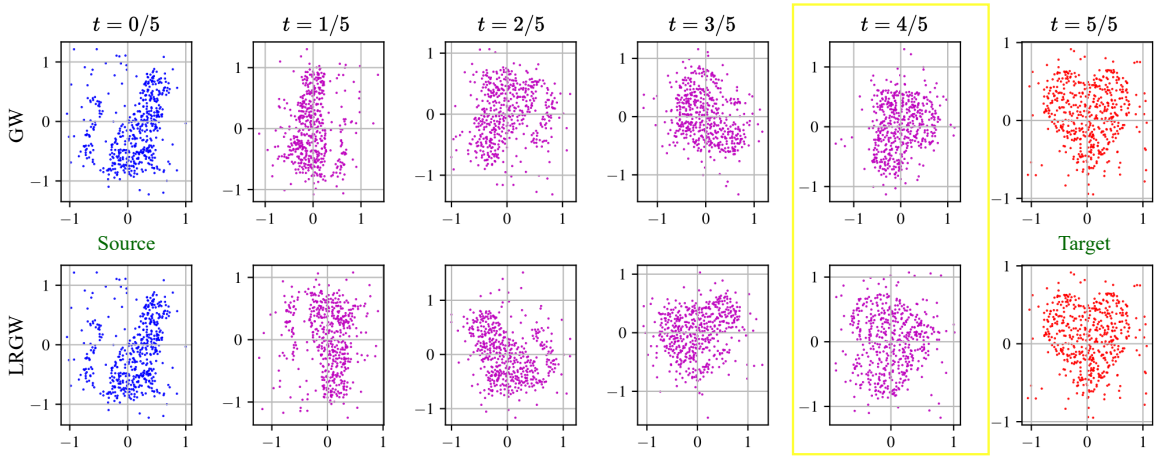


Figure 4.14: GW and LRGW barycenters between source (cat) and target (heart) datasets at levels $t = 0/5, 1/5, \cdots, 5/5$. Both the source and target shapes are contaminated with 15% Gaussian outliers. The λ for both domains are chosen following the $\tilde{m} + 3\tilde{\sigma}$ threshold, i.e., 1.87 and 1.59 respectively. LRGW barycenters tend to recover robust structures better than GW (notice, as marked in yellow).

Chapter 5

Conclusion

Summary

In this chapter, we summarize the contributions of the previous chapters and point towards potential future extensions. Moreover, we briefly outline the impact our work has had in the literature. We also pose some interesting open questions that we have come across during our investigations.

5.1 Contributions and Impact

This thesis has centered around two fundamental questions: First, do deep generative models involving distinct spaces accurately approximate the true data distributions under suitable characterization? Second, what are the principal ways to robustify the models without altering their architecture, given their extent of inherent denoising capacity? These questions have been explored specifically in the contexts of Wasserstein autoencoders and unsupervised unpaired cross-domain translators, maintaining cycle-consistency.

The data approximation aspect, in our approach, is reformulated as a density estimation task, which, based on our introductory discussion, becomes equivalent to learning a sampler from the viewpoint of information theory. In Wasserstein autoencoders, under smoothness assumptions on the data distribution, achieving successful generation is sufficiently dependent upon learning an information preservation (IP) encoding transform. We establish supporting deterministic non-parametric upper bounds on latent-space errors, which adapt to the intrinsic geometry of the input distribution and remain invariant under group actions. Moreover, optimal encoders can be provably obtained by solving a minimum-distance problem in latent space. Additionally, reconstructions via WAE-MMD naturally follow from latent consistency under regular and invariant kernels. Our approach has since inspired similar studies in VAEs in a PAC-Bayes setup (Mbacke et al., 2023) and has been adapted to find error bounds depending on the Pseudo-dimension of input densities (Chakraborty and Bartlett, 2024).

Building on this framework, the thesis extends IP-based methods to unpaired cycle-

consistent cross-domain generative models. It analyzes the statistical error introduced by ill-posed translation tasks and shows that, under smooth input distributions, the use of L^1 and Wasserstein losses in cyclic objectives yields similar outcomes. Crucially, it demonstrates that ensuring translation consistency suffices to achieve cycle-consistency in total variation, provided smoothness is preserved. Our framework has since been adopted to draw similar inferences under a misspecified model (Chakraborty and Bartlett, 2025) and to further enforce identifiability (Shrestha and Fu, 2024).

In the final part, the thesis addresses the robustness of such DGMs by proposing new variants of the Gromov-Wasserstein (GW) distance tailored for contaminated data. These include Tukey-GW and Huber-GW distances, which offer outlier resilience while preserving metric properties and remain computationally tractable. A more conservative lower bound, Locally Robust GW, is also introduced, providing stronger robustness at the expense of a removable degeneracy. These methods extend to probabilistic metric-measure spaces and enable the design of robust, cycle-consistent cross-domain models. Empirical evaluations confirm that the proposed techniques outperform existing approaches in tasks such as image synthesis, shape matching, and interpolation.

5.2 Open Questions

While this thesis advances several core aspects of the subject, it also brings to light a number of compelling and unresolved questions that merit further exploration. Below, we outline some of the most intriguing directions for future work.

Consistent truncation in TGW and LRGW

Our prescriptions of suitable truncation parameters for Tukey's GW (TGW) and Locally Robust GW (LRGW) in Chapter 4 rely on $\tau \coloneqq \tilde{m}+3\tilde{\sigma}$, where $\tilde{m}$ and $\tilde{\sigma}$ are the median and mean absolute deviation (MAD) about median of distortions and pairwise distances, respectively. While such a data-dependent thresholding is empirically effective in both norm penalization (Chakrabarty et al., 2024) and local robustification (Chakrabarty et al., 2025b), exploring a statistically consistent algorithmic solution for choosing τ, such that it estimates the effective diameter of inlying observations, seems urgent. Our method gives an optimization-free, low-complexity solution, which may suffer from rigidity and hence suboptimal performance.

Optimal latent dimension in WAEs

As already hinted in Chapter 2, despite recent empirical progress, a fundamental lack of theoretical guidance remains regarding the optimal latent dimension of a WAE. For instance, ARD-VAE introduces a hierarchical prior over latent dimensions to dynamically identify

and prune irrelevant axes, effectively discovering an appropriate latent size for each dataset, with demonstrated improvements in FID and disentanglement on benchmarks (Saha et al., 2025). In contrast, Leeb et al. (2022) introduces the concept of 'latent responses', which leverages interventions in the latent space— i.e., selectively altering one latent dimension while keeping others fixed— to probe the causal and structural dependencies within the learned representation. Their framework quantifies how one dimension influences others through a latent response matrix, a tool for detecting inactive dimensions (posterior collapse) and revealing cross-couplings in the latent manifold. Together, these approaches represent a shift from heuristic hyperparameter tuning toward adaptive, data-driven determination of latent size, although a unified theoretical framework for optimality is still forthcoming.

Guarantee of Denoising Diffusion Probabilistic Models (DDPMs)

As also hinted in the introduction, the empirical success of Score-based Generative Models (SGMs) (Song et al., 2021b) has recently attracted scores of theoretical scrutiny into its mechanism. SGMs, instead of modeling the data density p_μ itself, estimate the gradient of the log-density (score function) in a discretized (Euler-Maruyama (EM) scheme) process, where the forward-pass is typically governed by the Ornstein-Uhlenbeck (OU) process. As generation guarantees seem to establish, most works have focused on showing deterministic upper bounds on the discrepancy between p_μ and the law of $\{\overrightarrow{X_t}\}$, where the backward process $\{\overrightarrow{X_t}\}$ is terminated after time t^*, based on predefined step sizes. An immediate investigation may look into whether the generation process of SGMs maintains IP, and the ensuing error bounds can be obtained as an extension to our non-parametric approach. We point out that from the existing works as a whole, several patterns emerge in terms of their limitations and technical considerations. **(1)** For example, most, if not all, rely on the assumption that p_μ is log-concave, only recently being relaxed to 'weak' log-concavity (Silveri and Ocello, 2025). Given that log-concave densities are inherently unimodal, such characterization does not accommodate most practical cases, where datasets consist of images with multiple classes and text having definite groups. Their exponentially decaying tail behavior also makes the studies incapable of discussing robustness under outlying observations. **(2)** Moreover, all convergence bounds to date consider the divergence $\Omega(\cdot,\cdot)$ measuring the discrepancy $\Omega(p_\mu, \mathcal{L}\{\overrightarrow{X_{t^*}}\})$ to be either KL (implied to TV due to Pinsker's inequality) or W_p (Wasserstein, given $p = 1, 2$, also implied from KL). While the choice remains solely technical, as the underlying processes do not dictate a favorite, this poses an interesting question: *Can sharper error bounds be established based on MMD, under specific choices of kernels?* This is particularly intriguing since, under stochastic localization, matching scores essentially boil down to comparing moments. Also, as highlighted in our work, kernels *exist* that metrize the Wasserstein space, and showing convergence of probability measures under W_p becomes equivalent to showing the same under corresponding MMD (e.g., Energy kernels (Modeste and Dombry, 2024)).

The potential benefit of an MMD error bound might be its parametric dependence on data dimension, which under W_2 turns out optimally $\sqrt{d}$ (Bruno et al., 2025). It is yet to be seen how convergence bounds adapt to the intrinsic dimensionality of data supports. **(3)** Works so far also assume the expected *score approximation error* to be arbitrarily small or bounded by a deterministic quantity (see Bruno et al. (2025)). Given that most SGMs employ attention-based architectures, it would be interesting to explore the parametric optimality conditions that enable such approximations.

Faster calculation of TGW

The algorithmic computation of TGW (similarly, HGW and LRGW) follows GW under entropic regularization. While this is the most common and frugal way of solving a matching problem in higher dimensions, often, computing proxies are beneficial. In this context, *slicing* has been identified as an effective technique to solve OT, boiling it down to mass allocation in 1D. In GW, employing a similar technique (Vayer et al., 2019) enables one to obtain an optimal plan by matching samples through permutations. However, choosing slicing directions uniformly at random over $\mathcal{U}(\mathbb{S}^{d-1})$ distorts the relation between pairwise distances post-projection. Furthermore, the complexity ($\mathcal{O}(n \log n)$) does not hold uniformly over all possible samples. Future work may investigate effective and efficient slicing methods that incur low cost and offer robust plans for TGW simultaneously.

List of Publications

A. Chakrabarty and S. Das. Statistical regeneration guarantees of the wasserstein autoencoder with latent space consistency. *Advances in neural information processing systems*, 34: 17098–17110, 2021. 5, 9, 11, 12, 17, 23, 28, 33, 35

A. Chakrabarty and S. Das. On translation and reconstruction guarantees of the cycle-consistent generative adversarial networks. *Advances in Neural Information Processing Systems*, 35:23607–23620, 2022. 5, 111

A. Chakrabarty, A. Basu, and S. Das. On robust cross domain alignment. *arXiv preprint arXiv:2412.15861 (Submitted)*, 2024. 5, 142

A. Chakrabarty, A. Basu, and S. Das. Information preservation with wasserstein autoencoders: generation consistency and adversarial robustness. *Statistics and Computing*, 35 (5):1–27, 2025a. 5, 9, 133

A. Chakrabarty, S. Subhra Mullick, and S. Das. Locally robust alignment between distinct spaces. *Stat*, 14(3):e70093, 2025b. 5, 142

References

J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh. Sorting with adversarial comparators and application to density estimation. In *2014 IEEE International Symposium on Information Theory*, pages 1682–1686. IEEE, 2014. 24

P. C. Alvarez-Esteban, E. Del Barrio, J. A. Cuesta-Albertos, and C. Matran. Trimmed comparison of distributions. *Journal of the American Statistical Association*, 103(482): 697–704, 2008. 104

C. Anil, J. Lucas, and R. Grosse. Sorting out Lipschitz function approximation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 291–301. PMLR, 2019. 14, 18, 78

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 2017. 3, 66, 69

S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In *Proceedings of the 34th International Conference on Machine Learning*, pages 224–232. PMLR, 2017. 66

S. Arya, A. Auddy, R. A. Clark, S. Lim, F. Memoli, and D. Packer. The gromov–wasserstein distance between spheres. *Foundations of Computational Mathematics*, pages 1–56, 2024. 91

H. Asatryan, H. Gottschalk, M. Lippert, and M. Rottmann. A convenient infinite dimensional framework for generative adversarial learning. *Electronic Journal of Statistics*, 17(1):391–428, 2023. 14, 23, 31

H. Ashtiani and A. Mehrabian. Some techniques in density estimation, 2018.

H. Ashtiani, S. Ben-David, and A. Mehrabian. Sample-efficient learning of mixtures. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 17

Y. Bai, R. D. Martin, A. Kothapalli, H. Du, X. Liu, and S. Kolouri. Partial gromov-wasserstein metric. *arXiv preprint arXiv:2402.03664*, 2024. 89, 101, 107, 114

Y. Balaji, R. Chellappa, and S. Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems*, 33:12934–12944, 2020. 92, 95, 115

L. Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of multivariate analysis*, 88(1):190–206, 2004. 44

A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993. 21, 57

Y. Bartal, B. Recht, and L. J. Schulman. Dimensionality reduction: beyond the johnson-lindenstrauss bound. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 868–887. SIAM, 2011. 22

D. Bashkirova, B. Usman, and K. Saenko. Adversarial self-defense for cycle-consistent gans. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019. 76

M. Bauer, F. Mémoli, T. Needham, and M. Nishino. The z-gromov-wasserstein distance. *arXiv preprint arXiv:2408.08233*, 2024. 91, 106, 108

D. Belomestny, E. Moulines, A. Naumov, N. Puchkin, and S. Samsonov. Rates of convergence for density estimation with gans. *arXiv preprint arXiv:2102.00199*, 2021.

A. Ben-Israel. The change-of-variables formula using matrix volume. *SIAM Journal on Matrix Analysis and Applications*, 21(1):300–312, 1999. 23

S. Benaim and L. Wolf. One-sided unsupervised domain mapping. *Advances in neural information processing systems*, 30, 2017. 111

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. 8, 43

R. Bennett. The intrinsic dimensionality of signal collections. *IEEE Transactions on Information Theory*, 15(5):517–525, 1969. 42

J. Benton, V. D. Bortoli, A. Doucet, and G. Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024. 4

K. Bertin, S. E. Kolei, and N. Klutchnikoff. Adaptive density estimation on bounded domains. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 55(4):1916 – 1947, 2019. 16

G. Biau, B. Cadre, M. Sangnier, and U. Tanielian. Some theoretical properties of GANS. *The Annals of Statistics*, 48:1539 – 1566, 2020. 66

G. Biau, M. Sangnier, and U. Tanielian. Some theoretical insights into wasserstein gans. *Journal of Machine Learning Research*, 22(119):1–45, 2021. 66

J. Birrell, M. A. Katsoulakis, L. Rey-Bellet, and W. Zhu. Structure-preserving gans. In *International Conference on Machine Learning*, 2022. 26, 59

J. Blanchet, A. Jambulapati, C. Kent, and A. Sidford. Towards optimal running timesfor optimal transport. *Operations Research Letters*, 52:107054, 2024. 91

A. J. Blumberg, I. Gal, M. A. Mandell, and M. Pancia. Robust statistics, hypothesis testing, and confidence intervals for persistent homology on metric measure spaces. *Foundations of Computational Mathematics*, 14:745–789, 2014. 93, 111

V. D. Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. 4

J. Bourgain. On lipschitz embedding of finite metric spaces in hilbert space. *Israel Journal of Mathematics*, 52:46–52, 1985. 22

G. E. Box and M. E. Muller. A note on the generation of random normal deviates. *The Annals of Mathematical Statistics*, 29(2):610–611, 1958. 2

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, USA, 2004. ISBN 0521833787. 70

Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991. 23, 72, 135

F.-X. Briol, A. Barp, A. B. Duncan, and M. Girolami. Statistical inference for generative models with maximum mean discrepancy. *arXiv preprint arXiv:1906.05944*, 2019. 54

S. Bruno, Y. Zhang, D. Lim, O. D. Akyildiz, and S. Sabanis. On diffusion-based generative models and their error bounds: The log-concave case with full convergence estimates. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. 4, 144

C. Bunne, D. Alvarez-Melis, A. Krause, and S. Jegelka. Learning generative models across incomparable spaces. In *International conference on machine learning*, pages 851–861. PMLR, 2019. 4, 118

L. A. Caffarelli. The regularity of mappings with a convex potential. *Journal of the American Mathematical Society*, 5(1):99–104, 1992. 72

L. A. Caffarelli. Monotonicity properties of optimal transportation and the fkg and related inequalities. *Communications in Mathematical Physics*, 214:547–563, 2000. 23

Y. Cai and L.-H. Lim. Distances between probability distributions of different dimensions, 2021.

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011. 8

A. Caragea, P. Petersen, and F. Voigtlaender. Neural network approximation and estimation of classifiers with classification boundary in a barron class. *arXiv preprint arXiv:2011.09363*, 2020. 20

M. Chae and S. G. Walker. Wasserstein upper bounds of the total variation for smooth densities. *Statistics and Probability Letters*, 163:108771, 2020. 74

A. Chakrabarty and S. Das. Statistical regeneration guarantees of the wasserstein autoencoder with latent space consistency. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:17098–17110, 2021. 5, 9, 11, 12, 17, 23, 28, 33, 35

A. Chakrabarty and S. Das. On translation and reconstruction guarantees of the cycle-consistent generative adversarial networks. *Advances in Neural Information Processing Systems*, 35:23607–23620, 2022. 5, 111

A. Chakrabarty, A. Basu, and S. Das. On robust cross domain alignment. *arXiv preprint arXiv:2412.15861 (Submitted)*, 2024. 5, 142

A. Chakrabarty, A. Basu, and S. Das. Information preservation with wasserstein autoencoders: generation consistency and adversarial robustness. *Statistics and Computing*, 35 (5):1–27, 2025a. 5, 9, 133

A. Chakrabarty, S. Subhra Mullick, and S. Das. Locally robust alignment between distinct spaces. *Stat*, 14(3):e70093, 2025b. 5, 142

S. Chakraborty and P. Bartlett. A statistical analysis of wasserstein autoencoders for intrinsically low-dimensional data. In *The 12th International Conference on Learning Representations*, 2024. 9, 29, 141

S. Chakraborty and P. Bartlett. Statistical guarantees for unpaired image-to-image cross-domain analysis using gans. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. 75, 142

L. Chapel, M. Z. Alaya, and G. Gasso. Partial optimal tranport with applications on positive-unlabeled learning. *Advances in Neural Information Processing Systems*, 33:2903–2913, 2020. 88, 89, 99

M. Chen, C. Gao, and Z. Ren. A general decision theory for Huber's ϵ-contamination model. *Electronic Journal of Statistics*, 10:3752 – 3774, 2016. 38

M. Chen, H. Jiang, W. Liao, and T. Zhao. Efficient approximation of deep relu networks for functions on low dimensional manifolds. *Advances in neural information processing systems*, 32, 2019. 18

M. Chen, W. Liao, H. Zha, and T. Zhao. Statistical guarantees of generative adversarial networks for distribution estimation, 2020. 66, 69

M. Chen, K. Huang, T. Zhao, and M. Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pages 4672–4712. PMLR, 2023a. 4

S. Chen, S. Chewi, J. Li, Y. Li, A. Salim, and A. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023b. 4

Z. Chen, M. Katsoulakis, L. Rey-Bellet, and W. Zhu. Sample complexity of probability divergences under group symmetry. In *International Conference on Machine Learning*, pages 4713–4734. PMLR, 2023c. 27, 60

A. Chernodub and D. Nowicki. Norm-preserving orthogonal permutation linear unit activation functions (oplu). *arXiv preprint arXiv:1604.02313*, 2016. 15

J. Chhoa, M. Ivanitskiy, F. Jiang, S. Li, D. McBride, T. Needham, and K. O'Hare. Metric properties of partial and robust gromov-wasserstein distances. *arXiv preprint arXiv:2411.02198*, 2024.

P.-A. Chiappori, R. J. McCann, and B. Pass. Multi-to one-dimensional optimal transport. *Communications on Pure and Applied Mathematics*, 70(12):2405–2444, 2017. 23

Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 111

S. Chowdhury and F. Mémoli. The gromov–wasserstein distance between networks and stable network invariants. *Information and Inference: A Journal of the IMA*, 8(4):757–787, 2019. 103, 124

S. Chowdhury and T. Needham. Generalized spectral clustering via gromov-wasserstein learning. In *International Conference on Artificial Intelligence and Statistics*, pages 712–720. PMLR, 2021. 87

C. Chu, A. Zhmoginov, and M. Sandler. Cyclegan, a master of steganography, 2017.

R. A. Clark, T. Needham, and T. Weighill. Generalized dimension reduction using semi-relaxed gromov-wasserstein distance. *arXiv preprint arXiv:2405.15959*, 2024. 87

K. Clarkson, R. Wang, and D. Woodruff. Dimensionality reduction for tukey regression. In *International Conference on Machine Learning*, pages 1262–1271. PMLR, 2019. 94

K. L. Clarkson and D. P. Woodruff. Sketching for m-estimators: A unified approach to robust regression. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 921–939. SIAM, 2014. 98

G. Cleanthous, A. G. Georgiadis, and E. Porcu. Minimax density estimation on sobolev spaces with dominating mixed smoothness, 2019. 85

G. Cleanthous, A. Georgiadis, and O. Lepski. Adaptive estimation of the $\mathbb{L}_2$-norm of a probability density and related topics ii. upper bounds via the oracle approach. *The Annals of Statistics*, 53(3):1275–1297, 2025. 2

J. D. Clemens, S. Gao, and A. S. Kechris. Polish metric spaces: their classification and isometry groups. *Bulletin of Symbolic Logic*, 7(3):361–375, 2001.

M. Colombo and M. Fathi. Bounds on optimal transport maps onto log-concave measures. *Journal of Differential Equations*, 271:1007–1022, 2021. ISSN 0022-0396. 23, 72

J. Corander, U. Remes, and T. Koski. On the Jensen-Shannon divergence and the variation distance for categorical probability distributions. *Kybernetika*, 57(6):879–907, 2021.

N. Courty, R. Flamary, and M. Ducoffe. Learning wasserstein embeddings. In *International Conference on Learning Representations*, 2018. 22

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 98

C. Czado and A. Munk. Assessing the similarity of distributions-finite sample performance of the empirical mallows distance. *Journal of Statistical Computation and Simulation*, 60 (4):319–346, 1998. 104

B. Dai and D. Wipf. Diagnosing and enhancing VAE models. In *International Conference on Learning Representations*, 2019. 9

B. Dai, Y. Wang, J. Aston, G. Hua, and D. Wipf. Hidden talents of the variational autoencoder. *arXiv preprint arXiv:1706.05148*, 2017.

B. Dai, Y. Wang, J. Aston, G. Hua, and D. Wipf. Connections with robust pca and the role of emergent sparsity in variational autoencoder models. *The Journal of Machine Learning Research*, 19(1):1573–1614, 2018. 8

I. Daubechies, R. DeVore, S. Foucart, B. Hanin, and G. Petrova. Nonlinear approximation and (deep) relu networks. *Constructive Approximation*, 55(1):127–172, 2022. 18

G. David and S. Semmes. Regular mappings between dimensions. *Publicacions Matematiques*, pages 369–417, 2000. 34

A. G. de G. Matthews, J. Hron, M. Rowland, R. E. Turner, and Z. Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018. 80

N. De Ponti and A. Mondino. Entropy-transport distances between unbalanced metric measure spaces. *Probability Theory and Related Fields*, 184(1):159–208, 2022. 89

T. De Ryck, S. Lanthaler, and S. Mishra. On the approximation of functions by tanh neural networks. *Neural Networks*, 143:732–750, 2021. 72

N. Deb, P. Ghosal, and B. Sen. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections, 2021.

P. Demetci, R. Santorella, B. Sandstede, W. S. Noble, and R. Singh. Scot: Single-cell multiomics alignment with optimal transport. *Journal of Computational Biology*, 29(1):3–18, 2022. 87

L. Devroye and L. Gyorfi. No empirical probability measure can converge in the total variation sense for all distributions. *The Annals of Statistics*, pages 1496–1499, 1990. 17

L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2001. 17, 24, 73, 82

P. J. Diggle and R. J. Gratton. Monte carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 46(2): 193–212, 1984. 1

L. Ding, L. Zou, W. Wang, S. Shahrampour, and R. Tuo. High-dimensional non-parametric density estimation in mixed smooth sobolev spaces, 2021.

K. Do and T. Tran. Theory and evaluation metrics for learning disentangled representations. In *International Conference on Learning Representations*, 2020. 43

D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995. 2

D. Dowson and B. Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982. 108

T. Dumont, T. Lacombe, and F.-X. Vialard. On the existence of monge maps for the gromov–wasserstein problem. *Foundations of Computational Mathematics*, pages 1–48, 2024. 93

D. Dutta, A. Chakrabarty, and S. Das. Lost in translation: GANs' inability to generate simple probability distributions. In *The Second Tiny Papers Track at ICLR 2024*, 2024. 44

P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 2018. 25

G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, page 258–267, 2015. 69

S. Y. Efroimovich. Nonparametric estimation of a density of unknown smoothness. *Theory of Probability & Its Applications*, 30(3):557–568, 1986. 2

D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003. 24

E. Eneva, K. Kumaraswami, and M. Matteucci. Wekkem: A study in fractal dimension and dimensionality reduction. In *Workshop on Fractals and Self-similarity in Data Mining: Issues and Approaches*, 2002. 42

E. Facco, M. d'Errico, A. Rodriguez, and A. Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):1–8, 2017. 42

G. Fasano and A. Franceschini. A multidimensional version of the kolmogorov–smirnov test. *Monthly Notices of the Royal Astronomical Society*, 225(1):155–170, 1987. 44

C. Fefferman, S. Mitter, and H. Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016. 42

R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html. 107

C. Franz. cramer: multivariate nonparametric cramer-test for the two-sample-problem. *R package version 0.8-1*, 2006. 44

B. E. Fristedt and L. F. Gray. *A modern approach to probability theory*. Springer Science & Business Media, 2013. 104, 106

H. Fu, M. Gong, C. Wang, K. Batmanghelich, K. Zhang, and D. Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2427–2436, 2019. 111

I. Gat, T. Remez, N. Shaul, F. Kreuk, R. T. Chen, G. Synnaeve, Y. Adi, and Y. Lipman. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37:133345–133385, 2024. 3

B. Gaujac, I. Feige, and D. Barber. Learning disentangled representations with the wasserstein autoencoder. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 69–84. Springer, 2021. 43

A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002. 75

E. Giné and A. Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 38, pages 907–921. Elsevier, 2002. 25

E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2021. 25, 62, 63, 83, 85

A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3):1608 – 1632, 2011. 2, 16

F. Gong, Y. Nie, and H. Xu. Gromov-wasserstein multi-modal alignment and clustering. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 603–613, 2022. 87

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014. 3, 31, 64

L.-A. Gottlieb, A. Kontorovich, and R. Krauthgamer. Efficient regression in metric spaces via approximate lipschitz extension. *IEEE Transactions on Information Theory*, 63(8): 4838–4849, 2017. 16

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. 10

A. Greven, P. Pfaffelhuber, and A. Winter. Convergence in distribution of random metric measure spaces (λ-coalescent measure trees). *Probability Theory and Related Fields*, 145 (1):285–322, 2009. 128, 129

R. Gribonval, G. Kutyniok, M. Nielsen, and F. Voigtlaender. Approximation spaces of deep neural networks. *Constructive approximation*, 55(1):259–367, 2022. 18

M. Groppe and S. Hundrieser. Lower complexity adaptation for empirical entropic optimal transport. *arXiv preprint arXiv:2306.13580*, 2023. 105

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017. 117

X. Guo and L. Zhao. A systematic survey on deep generative models for graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5370–5390, 2022. 3

M. Haas and S. Richter. Statistical analysis of wasserstein gans with applications to time series forecasting, 2020. 66

M. Hammami, D. Friboulet, and R. Kechichian. Cycle gan-based data augmentation for multi-organ detection in ct images via yolo. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 390–393, 2020.

J. Han, M. R. Min, L. Han, L. E. Li, and X. Zhang. Disentangled recurrent wasserstein autoencoder. In *International Conference on Learning Representations*, 2021. 43

F. He and D. Tao. Recent advances in deep learning theory, 2021.

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. 3

I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018. 43

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

Z. Hu, Z. Yang, R. Salakhutdinov, and E. P. Xing. On unifying deep generative models. In *International Conference on Learning Representations*, 2018.

J. Huang, Y. Jiao, Z. Li, S. Liu, Y. Wang, and Y. Yang. An error analysis of generative adversarial networks for learning distributions, 2021. 66, 81

X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018. 111

P. Huber. *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley, 1981. 127

P. J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964. 94

Y. Hur, W. Guo, and T. Liang. Reversible gromov–monge sampler for simulation-based inference. *SIAM Journal on Mathematics of Data Science*, 6(2):283–310, 2024. 93, 115, 116, 117, 134

H. Husain, R. Nock, and R. C. Williamson. A primal-dual link between gans and autoencoders. *Advances in Neural Information Processing Systems*, 32, 2019. 8, 9, 11

T. Huster, C.-Y. J. Chiang, and R. Chadha. Limitations of the lipschitz constant as a defense against adversarial examples. In *ECML PKDD 2018 Workshops*. Springer International Publishing, 2019. 14

A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999. 15

A. Jain and A. Orlitsky. A general method for robust learning from batches. *Advances in Neural Information Processing Systems*, 33:21775–21785, 2020. 17

Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1965–1972, 2017. 22

W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Conference in modern analysis and probability*, 26:189–206, 1984. 22

G. Kerkyacharian, D. Picard, and K. Tribouley. Lp adaptive density estimation. *Bernoulli*, 2(3):229–247, 1996. 2

I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020. 15

H. Kim and A. Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018. 43

J. Kim, J. Shin, A. Rinaldo, and L. Wasserman. Uniform convergence rate of the kernel density estimator adaptive to intrinsic volume dimension. In *International Conference on Machine Learning*, pages 3398–3407. PMLR, 2019. 1, 18

J. Kim, M. Kim, H. Kang, and K. H. Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *International Conference on Learning Representations*, 2020. 65, 69

T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1857–1865, 2017. 3, 65, 68

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations*, 2014a. 7

D. P. Kingma and M. Welling. Auto-encoding variational bayes, 2014b. 3, 65

J. M. Klusowski and A. R. Barron. Approximation by combinations of relu and squared relu ridge functions with l^1 and l^0 controls. *IEEE Transactions on Information Theory*, 64(12): 7649–7656, 2018. 21

F. Koehler, V. Mehta, C. Zhou, and A. Risteski. Variational autoencoders in the presence of low-dimensional data: landscape and implicit bias. In *International Conference on Learning Representations*, 2022.

D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009. 1

L. Kong, J. Li, J. Tang, and A. M.-C. So. Outlier-robust gromov-wasserstein for graph data. *Advances in Neural Information Processing Systems*, 36, 2024. 90, 115

J. Lafferty, H. Liu, and L. Wasserman. Concentration of measure. *On-line]. Available: http://www. stat. cmu. edu/ larry/= sml/Concentration. pdf*, 2008. 52

S. Langer. Approximating smooth functions by deep neural networks with sigmoid activation function. *Journal of Multivariate Analysis*, 182:104696, 2021. 72

K. G. Larsen and J. Nelson. Optimality of the johnson-lindenstrauss lemma. In *58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 633–638. IEEE, 2017. 22

K. Le, H. Nguyen, Q. M. Nguyen, T. Pham, H. Bui, and N. Ho. On robust optimal transport: Computational complexity and barycenter computation. *Advances in Neural Information Processing Systems*, 34:21947–21959, 2021. 90, 92, 115

G. Lecué and M. Lerasle. Robust machine learning by median-of-means: Theory and practice. *The Annals of Statistics*, 48(2):906 – 931, 2020. 97

Y. LeCun, C. Cortes, and C. Burges J.c. MNIST handwritten digit database. URL `https://yann.lecun.com/exdb/mnist/`. 29

H. Lee, R. Ge, T. Ma, A. Risteski, and S. Arora. On the ability of neural nets to express distributions. In *Conference on Learning Theory*, pages 1271–1296. PMLR, 2017. 21, 57

F. Leeb, S. Bauer, M. Besserve, and B. Schölkopf. Exploring the latent space of autoencoders with interventional assays. *Advances in neural information processing systems*, 35:21562–21574, 2022. 143

J. Lei. Convergence and concentration of empirical measures under wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1), 2020. ISSN 1350-7265.

N. Lei, Y. Guo, D. An, X. Qi, Z. Luo, S.-T. Yau, and X. Gu. Mode collapse and regularity of optimal transportation maps, 2019. 73

N. Lei, D. An, Y. Guo, K. Su, S. Liu, Z. Luo, S.-T. Yau, and X. Gu. A geometric understanding of deep learning. *Engineering*, 6(3):361–374, 2020.

E. Levina and P. Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17, 2004. 42

C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos. MMD GAN: Towards deeper understanding of moment matching network. *Advances in neural information processing systems*, 30, 2017. 3

C. T. Li and F. Farnia. Mode-seeking divergences: Theory and applications to gans. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 8321–8350. PMLR, 2023. 41

G. Li, Y. Wei, Y. Chen, and Y. Chi. Towards non-asymptotic convergence for diffusion-based generative models. In *The Twelfth International Conference on Learning Representations*, 2024. 4

M. Li, H. Huang, L. Ma, W. Liu, T. Zhang, and Y.-G. Jiang. Unsupervised image-to-image translation with stacked cycle-consistent adversarial networks. In *ECCV*, 2018. 65

M. Li, J. Yu, H. Xu, and C. Meng. Efficient approximation of gromov-wasserstein distance using importance sparsification. *Journal of Computational and Graphical Statistics*, 32(4): 1512–1523, 2023. 99

T. Liang. How well can generative adversarial networks learn densities: A nonparametric view, 2018. 79

T. Liang. How well generative adversarial networks learn distributions. *Journal of Machine Learning Research*, 22(228):1–41, 2021. 66, 69

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t. 3

H. Liu and C. Gao. Density estimation with contamination: minimax rates and theory of adaptation. *Electronic Journal of Statistics*, 13:3613 – 3653, 2019. 38

M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017a. 4, 65, 111

Q. Liu, J. Xu, R. Jiang, and W. H. Wong. Density estimation using deep generative neural networks. *Proceedings of the National Academy of Sciences*, 118(15), 2021a. ISSN 0027-8424. 69

S. Liu, O. Bousquet, and K. Chaudhuri. Approximation and convergence properties of generative adversarial learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 5551–5559, 2017b. 66

S. Liu, Y. Yang, J. Huang, Y. Jiao, and Y. Wang. Non-asymptotic error bounds for bidirectional GANs. In *Advances in Neural Information Processing Systems*, 2021b. 66

T. Liu, P. Kumar, R. Zhou, and X. Liu. Learning from few samples: Transformation-invariant svms with composition and locality at multiple scales. *Advances in Neural Information Processing Systems*, 35:9151–9163, 2022. 39

X. Liu, C. Gong, and Q. Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. 3

Z. Liu and P.-L. Loh. Robust W-GAN-based estimation under Wasserstein contamination. *Information and Inference: A Journal of the IMA*, 12(1):312–362, 2022. 38

F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019. 43

P.-L. Loh. Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics*, 45(2):866 – 896, 2017. 98

G. Lu, Z. Zhou, Y. Song, K. Ren, and Y. Yu. Guiding the one-to-one mapping in cyclegan via optimal transport. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*. AAAI Press, 2019. 72

Y. Lu and J. Lu. A universal approximation theorem of deep neural networks for expressing probability distributions, 2020.

Y. Ma, H. Liu, D. La Vecchia, and M. Lerasle. Inference via robust optimal transportation: theory and methods. *arXiv preprint arXiv:2301.06297*, 2023. 105, 126

D. J. MacKay. Bayesian neural networks and density networks. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 354(1):73–80, 1995.

S. Mahabadi, K. Makarychev, Y. Makarychev, and I. Razenshteyn. Nonlinear dimension reduction via outer bi-lipschitz extensions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1088–1101, 2018. 33

A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

E. Mathieu, T. Rainforth, N. Siddharth, and Y. W. Teh. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, pages 4402–4412. PMLR, 2019. 43

S. D. Mbacke, F. Clerc, and P. Germain. Statistical guarantees for variational autoencoders using pac-bayesian theory. *Advances in Neural Information Processing Systems*, 36:56903–56915, 2023. 141

R. J. McCann. Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, 80(2):309 – 323, 1995. 23

R. J. McCann. Polar factorization of maps on riemannian manifolds. *Geometric & Functional Analysis GAFA*, 11(3):589–608, 2001. 135

R. J. McCann and B. Pass. Optimal transportation between unequal dimensions. *Archive for Rational Mechanics and Analysis*, 238(3):1475–1520, 2020. 23

F. Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11:417–487, 2011. 87, 91, 94, 95, 104, 121, 129, 130

F. Memoli, Z. Smith, and Z. Wan. The wasserstein transform. In *International Conference on Machine Learning*, pages 4496–4504. PMLR, 2019. 108

G. Mena and J. Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Advances in neural information processing systems*, 32, 2019. 91

L. Mescheder, S. Nowozin, and A. Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 2391–2400, 2017.

T. Modeste and C. Dombry. Characterization of translation invariant mmd on rd and connections with wasserstein distances. *Journal of Machine Learning Research*, 25(237):1–39, 2024. 19, 36, 143

H. Montanelli and Q. Du. New error bounds for deep relu networks using sparse grids. *SIAM Journal on Mathematics of Data Science*, 1(1):78–92, 2019. 18

H. Montanelli, H. Yang, and Q. Du. Deep relu networks overcome the curse of dimensionality for bandlimited functions. *arXiv preprint arXiv:1903.00735*, 2019.

N. Moriakov, J. Adler, and J. Teuwen. Kernel of cyclegan as a principal homogeneous space. In *International Conference on Learning Representations*, 2020. 11, 65, 66, 69, 70, 73

Y. Mroueh and M. Rigotti. Unbalanced sobolev descent. *Advances in Neural Information Processing Systems*, 33:17034–17043, 2020. 99

Y. Mroueh, C.-L. Li, T. Sercu, A. Raj, and Y. Cheng. Sobolev GAN. In *International Conference on Learning Representations*, 2018. 70

D. Mukherjee, A. Guha, J. M. Solomon, Y. Sun, and M. Yurochkin. Outlier-robust optimal transport. In *International Conference on Machine Learning*, pages 7850–7860. PMLR, 2021. 6, 96

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997. 10

C. Musco, C. Musco, D. P. Woodruff, and T. Yasuda. Active linear regression for l_p norms and beyond. *arXiv preprint arXiv:2111.04888*, 2021. 119

N. Nakagawa, R. Togo, T. Ogawa, and M. Haseyama. Gromov-wasserstein autoencoders. In *The Eleventh International Conference on Learning Representations*, 2023. 4, 77, 118

Q. M. Nguyen, H. H. Nguyen, Y. Zhou, and L. M. Nguyen. On unbalanced optimal transport: Gradient methods, sparsity and approximation error. *The Journal of Machine Learning Research*, 24(1):18390–18430, 2023. 89

R. Nickl and B. M. Pötscher. Bracketing metric entropy rates and empirical central limit theorems for function classes of besov-and sobolev-type. *Journal of Theoretical Probability*, 20:177–199, 2007. 16

S. Nietert, Z. Goldfeld, and R. Cummings. Outlier-robust optimal transport: Duality, structure, and statistical analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 11691–11719. PMLR, 2022. 92, 109, 126

S. Nietert, R. Cummings, and Z. Goldfeld. Robust estimation under the wasserstein distance. *arXiv preprint arXiv:2302.01237*, 2023. 92, 96, 117, 137

S. Nowozin, B. Cseke, and R. Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 66

Y. Pang, J. Lin, T. Qin, and Z. Chen. Image-to-image translation: Methods and applications, 2021.

G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021. 3

J. A. Peacock. Two-dimensional goodness-of-fit testing in astronomy. *Monthly Notices of the Royal Astronomical Society*, 202(3), 1983. 43

O. Pele and M. Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th international conference on computer vision*, pages 460–467. IEEE, 2009. 96

P. Petersen and F. Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330, 2018. 14, 71

G. Peyré, M. Cuturi, and J. Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, pages 2664–2672. PMLR, 2016. 98, 99, 107

P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021. 29, 42

C. Puritz, E. Ness-Cohn, and R. Braun. fasano. franceschini. test: An implementation of a multidimensional ks test in r. *arXiv preprint arXiv:2106.10539*, 2021. 44

S. Raghvendra, P. Shirzadian, and K. Zhang. A new robust partial p-Wasserstein-based metric for comparing distributions. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 41867–41885. PMLR, 21–27 Jul 2024. 92, 110

D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. 3

J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31:2351–2364, June 2023. ISSN 2329-9290. 3

G. Rioux, Z. Goldfeld, and K. Kato. Entropic gromov-wasserstein distances: Stability and algorithms. *arXiv preprint arXiv:2306.00182*, 2023. 92

M. Rolinek, D. Zietlow, and G. Martius. Variational autoencoders pursue pca directions (by accident). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12406–12415, 2019. 12, 15

P. K. Rubenstein, B. Schoelkopf, and I. Tolstikhin. Learning disentangled representations with wasserstein auto-encoders, 2018. 43

S. Saha, S. Joshi, and R. Whitaker. ARD-VAE: A Statistical Formulation to Find the Relevant Latent Dimensions of Variational Autoencoders . In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 889–898, 2025. 143

A. Salmona, A. Desolneux, and J. Delon. Gromov-wasserstein-like distances in the gaussian mixture models space. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. 108, 109

V. Sandfort, K. Yan, P. Pickhardt, and R. Summers. Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. *Scientific Reports*, 9, 2019. 65

M. Scetbon, G. Peyré, and M. Cuturi. Linear-time gromov wasserstein distances using low rank couplings and costs. In *International Conference on Machine Learning*, pages 19347–19365. PMLR, 2022. 92

N. Schreuder. Bounding the expectation of the supremum of empirical processes indexed by hölder classes, 2020. 81

N. Schreuder, V.-E. Brunel, and A. Dalalyan. Statistical guarantees for generative models without domination. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, pages 1051–1071, 2021. 66, 81

B. Schweizer, A. Sklar, et al. Statistical metric spaces. *Pacific J. Math*, 10(1):313–334, 1960. 106

T. Séjourné, F.-X. Vialard, and G. Peyré. The unbalanced gromov-wasserstein distance: Conic formulation and relaxation. *Advances in Neural Information Processing Systems*, 34:8766–8779, 2021. 88, 89, 100, 114

Z. Shen, H. Yang, and S. Zhang. Deep network approximation characterized by number of neurons. *arXiv preprint arXiv:1906.05497*, 2019. 20, 79

Z. Shen, H. Yang, and S. Zhang. Optimal approximation rate of relu networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, 157:101–135, 2022.

S. Shrestha and X. Fu. Towards identifiable unsupervised domain translation: A diversified distribution matching approach. In *The Twelfth International Conference on Learning Representations*, 2024. 142

M. G. Silveri and A. Ocello. Beyond log-concavity and score regularity: Improved convergence bounds for score-based generative models in w2-distance. In *Forty-second International Conference on Machine Learning*, 2025. 4, 143

B. Sim, G. Oh, J. Kim, C. Jung, and J. C. Ye. Optimal transport driven cyclegan for unsupervised learning in inverse problems. *SIAM Journal on Imaging Sciences*, 13(4): 2281–2306, 2020. 65, 72

S. Singh, A. Uppal, B. Li, C.-L. Li, M. Zaheer, and B. Póczos. Nonparametric density estimation with adversarial losses. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 10246–10257, 2018.

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, 2015. 3

J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. 3

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. 3, 143

Y. Song, P. Dhariwal, M. Chen, and I. Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 32211–32252, 2023. 3

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. Lanckriet. On integral probability metrics, ϕ-divergences and binary classification. *preprint arXiv:0901.2698*, 2009. 15, 54

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010. 19

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6: 1550 – 1599, 2012. 82, 103

G. Staerman, P. Laforgue, P. Mozharovskyi, and F. d'Alché Buc. When ot meets mom: Robust estimation of wasserstein distance. In *International Conference on Artificial Intelligence and Statistics*, pages 136–144. PMLR, 2021.

K.-T. Sturm. On the geometry of metric measure spaces. *Acta Mathematica*, 196(1):65 – 131, 2006. 89, 110, 111

K.-T. Sturm. *The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces*, volume 290. American Mathematical Society, 2023. 107, 121

T. Suzuki. Adaptivity of deep reLU network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019. 18

Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. In *5th International Conference on Learning Representations, ICLR 2017*, 2017. 65

R. Tang and Y. Yang. On empirical bayes variational autoencoder: An excess risk bound. In *Conference on Learning Theory*, pages 4068–4125. PMLR, 2021.

X. Tang, H. Dai, E. Knight, F. Wu, Y. Li, T. Li, and M. Gerstein. A survey of generative ai for de novo drug design: new frontiers in molecule and protein generation. *Briefings in Bioinformatics*, 25(4):bbae338, 2024. 3

U. Tanielian and G. Biau. Approximating lipschitz continuous functions with groupsort neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 442–450. PMLR, 2021. 20, 72

L. C. Tiao, E. V. Bonilla, and F. Ramos. Cycle-consistent adversarial learning as approximate bayesian inference, 2018. 66

I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018. 3, 4, 7, 8, 11, 13, 17

F. Topsoe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on information theory*, 46(4):1602–1609, 2000. 24

Q. H. Tran, H. Janati, N. Courty, R. Flamary, I. Redko, P. Demetci, and R. Singh. Unbalanced co-optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10006–10016, 2023. 89, 114

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 2008. ISBN 0387790519. 1

A. Uppal, S. Singh, and B. Poczos. Nonparametric density estimation and convergence rates for gans under besov ipm losses. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996. 16

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000. 26, 52

R. Van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014. 17

T. Vayer, R. Flamary, R. Tavenard, L. Chapel, and N. Courty. Sliced gromov-wasserstein. *arXiv preprint arXiv:1905.10124*, 2019. 144

T. Vayer, L. Chapel, R. Flamary, R. Tavenard, and N. Courty. Fused gromov-wasserstein distance for structured objects. *Algorithms*, 13(9):212, 2020. 99

C. Villani. *Optimal transport : old and new*. Grundlehren der mathematischen Wissenschaften. Springer, 2009. ISBN 978-3-540-71049-3. 11, 67, 80, 91

C. Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021. 110

A. Virmaux and K. Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31, 2018. 18

M. Vladimirova, S. Girard, H. Nguyen, and J. Arbel. Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9(1), 2020. 17

L. Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006. 1

J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620 – 2648, 2019. 29, 62, 79

J. Weed and Q. Berthet. Estimation of smooth densities in wasserstein distance. In *conference on Learning Theory*, pages 3118–3119. PMLR, 2019. 62

R. Wei, C. Garcia, A. El-Sayed, V. Peterson, and A. Mahmood. Variations in variational autoencoders - a comparative evaluation. *IEEE Access*, 8, 2020. 7

S. Wojtowytsch et al. Representation formulas and pointwise properties for barron functions. *Calculus of Variations and Partial Differential Equations*, 61(2):1–37, 2022. 20, 21, 57

Y. Yang, Z. Li, and Y. Wang. On the capacity of deep generative networks for approximating distributions. *Neural networks*, 145, 2022. 34, 36, 77

D. Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017. 14, 71

D. Yarotsky. Optimal approximation of continuous functions by very deep relu networks. In *Conference on learning theory*, pages 639–649. PMLR, 2018. 14, 71

Y. G. Yatracos. Rates of convergence of minimum distance estimators and kolmogorov's entropy. *The Annals of Statistics*, 13(2):768–774, 1985.

Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. *IEEE International Conference on Computer Vision (ICCV)*, pages 2868–2876, 2017. 3, 65, 68

Y. Yu, K. H. R. Chan, C. You, C. Song, and Y. Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. *Advances in Neural Information Processing Systems*, 33:9422–9434, 2020. 43

J. Yukich. Laws of large numbers for classes of functions. *Journal of Multivariate Analysis*, 17(3):245–260, 1985. ISSN 0047-259X. 52

P. Zhang, Q. Liu, D. Zhou, T. Xu, and X. He. On the discrimination-generalization tradeoff in GANs. In *International Conference on Learning Representations*, 2018. 66

Z. Zhang, Y. Mroueh, Z. Goldfeld, and B. Sriperumbudur. Cycle consistent probability divergences across different spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 7257–7285. PMLR, 2022. 66, 76, 88

Z. Zhang, Z. Goldfeld, Y. Mroueh, and B. K. Sriperumbudur. Gromov–wasserstein distances: Entropic regularization, duality and sample complexity. *The Annals of Statistics*, 52(4): 1616–1645, 2024. 97, 105, 115, 125

R. Zhou, C. Jiang, and Q. Xu. A survey on generative adversarial network-based text-to-image synthesis. *Neurocomputing*, 451:316–336, 2021. 3

B. Zhu, J. Jiao, and J. Steinhardt. Generalized resilience and robust statistics. *The Annals of Statistics*, 50(4):2256–2283, 2022. 38, 96

J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017. 3, 5, 64, 65, 66, 68, 69, 76, 113, 117