Journal of Machine Learning Research 23 (2022) 1-73 Submitted 1/21; Revised 5/22; Published 9/22

VINA: Variational Invertible Neural Architectures

Shubhanshu Shekhar* SHUBHAN@UMICH.EDU
EECS Department
University of Michigan, Ann Arbor

Mohammad Javad Khojasteh* MJKEME@RIT.EDU
EME Department
Rochester Institute of Technology

Ananya Acharya AA2334@RIT.EDU
EME Department
Rochester Institute of Technology

Tony Tohme TOHME@MIT.EDU
Department of Mechanical Engineering
Massachusetts Institute of Technology

Kamal Youcef-Toumi YOUCEF@MIT.EDU
Department of Mechanical Engineering
Massachusetts Institute of Technology

Editor:

Abstract

The distinctive architectural features of normalizing flows (NFs), notably bijectivity and tractable Jacobians, make them well-suited for generative modeling. Invertible neural networks (INNs) build on these principles to address supervised inverse problems, enabling direct modeling of both forward and inverse mappings. In this paper, we revisit these architectures from both theoretical and practical perspectives and address a key gap in the literature: the lack of theoretical guarantees on approximation quality under realistic assumptions, whether for posterior inference in INNs or for generative modeling with NFs.

We introduce a unified framework for INNs and NFs based on variational unsupervised loss functions, inspired by analogous formulations in related areas such as generative adversarial networks (GANs) and the Precision-Recall divergence for training normalizing flows. Within this framework, we derive theoretical performance guarantees, quantifying posterior accuracy for INNs and distributional accuracy for NFs, under assumptions that are weaker and more practically realistic than those used in prior work.

Building on these theoretical results, we conduct extensive case studies to distill general design principles and practical guidelines. We conclude by demonstrating the effectiveness of our approach on a realistic ocean-acoustic inversion problem.

Keywords: Invertible Neural Networks, Normalizing Flows, Variational Representation, Kernel Approximation, f-divergences.

∗. Equal contribution.

1 Introduction

Recently, machine learning approaches, particularly those based on deep neural networks, have emerged as effective alternatives to conventional inverse problem solvers. Among these, invertible neural networks (INNs) (Ardizzone et al., 2018) stand out due to their ability to model complex, non-linear relationships while ensuring invertibility between input and output spaces. INNs offer advantages such as bijectivity and computationally tractable Jacobians, which make them particularly suitable for solving inverse problems, where both forward and inverse mappings need to be computed accurately and efficiently. INNs have been applied for solving inverse problems in various fields, including epidemiology (Radev et al., 2021), astrophysics (Ardizzone et al., 2018), optics (Luce et al., 2023), geophysics (Zhang and Curtis, 2021; Wu et al., 2023), and reservoir engineering (Padmanabha and Zabaras, 2021).

The name INN in the literature has been used to refer to supervised models trained to solve inverse problems, where the training loss involves a supervised loss (to model observed outputs) as well as an unsupervised component (Ardizzone et al., 2019; Guan et al., 2024; Ardizzone et al., 2018; Hagemann and Neumayer, 2021). There exists a related class of models based on similar architectural principles, referred to as normalizing flows (NFs), which are primarily used for unsupervised density estimation and generative modeling (Papamakarios et al., 2021, 2017; Gomez et al., 2017; Kingma and Dhariwal, 2018; Dinh et al., 2014). For ease of exposition, we also use this convention and distinguish between INNs and NFs based on the existence of the supervised loss in the training process.

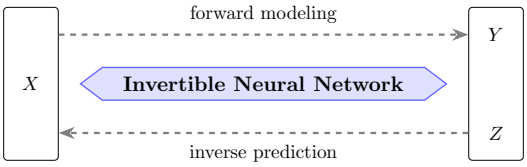


Figure 1: INNs can be represented by an invertible map T that approximates the relation from the input $X \in \mathbb{R}^{d_{\mathbf{x}}}$ to the output $Y \in \mathbb{R}^{d_{\mathbf{y}}}$ and a latent variable $Z \in \mathbb{R}^{d_{\mathbf{x}}-d_{\mathbf{y}}}$. The invertibility of T means that for any $\mathbf{y} \in \mathbb{R}^{d_{\mathbf{y}}}$, we also have an approximate posterior sampling distribution via $T^{-1}(\mathbf{y}, Z)$.

A typical inverse problem can be described as follows: We are given a dataset $\mathcal{D} = \{(X_i, Y_i) \in \mathbb{R}^{d_{\mathbf{x}}+d_{\mathbf{y}}}\}_{i=1}^{n}$ consisting of n input-output pairs, where $X_i \in \mathbb{R}^{d_{\mathbf{x}}}$ is the input feature vector and $Y_i \in \mathbb{R}^{d_{\mathbf{y}}}$ is the target vector, from a joint distribution $(X, Y) \sim P_{X,Y}$. Usually we have $d_{\mathbf{x}} \geq d_{\mathbf{y}}$ representing some inherent information loss in a measurement process (forward map). This joint distribution can be decomposed into a prior P_X and a possibly randomized transformation $P_{Y|X}$ representing the forward process. The general goal of the Bayesian approach to inverse problems (Stuart, 2010; Dashti et al., 2012) is to learn or approximate the posterior distribution $P_{X|Y}$ given the data $\mathcal{D}$. As depicted in Figure 1, the approach taken by INNs to achieve this goal is to incorporate an additional latent output variable, denoted by $Z \in \mathbb{R}^{d_{\mathbf{z}}}$ where $d_{\mathbf{z}} = d_{\mathbf{x}} - d_{\mathbf{y}}$. This latent variable is designed to represent information related to X that is not contained in Y, and its distribution P_Z is selected by the data analyst (Hagemann and Neumayer, 2021). An INN consists of a map $T \equiv (T_{\mathbf{y}}, T_{\mathbf{z}}) : \mathbb{R}^{d_{\mathbf{x}}} \to \mathbb{R}^{d_{\mathbf{y}} \times d_{\mathbf{z}}}$ that is continuous and invertible with a continuous inverse, and usually they are also assumed to be continuously differentiable to enable gradient based training procedures. Let $\mathcal{T}$ denote an appropriate parametrized collection

of such functions (we discuss some architectural details in Appendix A.1). The model is trained by minimizing an objective of the form

$$T^*_{\text{INN}} \in \arg\min_{T\in\mathcal{T}} \left(\mathbb{E}_{XY}[\|T_{\mathbf{y}}(X) - Y\|_2^2] \;+\; \lambda\, D(P_{Y,Z}, P_{Y,T_{\mathbf{z}}(X)})\right), \tag{1}$$

where the two terms correspond to the supervised and the unsupervised (population) losses, respectively, and $\lambda > 0$ is a regularizing constant. In this display, we use D to denote some divergence or distance measure (such as f-divergence, kernel-MMD, Wasserstein metric, etc.) between the joint distributions of $P_{Y,T_{\mathbf{z}}(X)}$ and $P_{Y,Z} = P_Y P_Z$. The supervised term encourages the component $T_{\mathbf{y}}$ to approximate the forward map $P_{Y|X}$, while the unsupervised term matches the joint distribution of $P_{Y,T_{\mathbf{z}}(X)}$ with the output-latent variable pair $P_{Y,Z}$. Crucially, due to the invertibility of T, it is possible to show that under certain conditions, making these two terms small induces an approximate posterior sampling distribution: for a given $Y = \mathbf{y}$, the distribution of $T^{-1}(\mathbf{y}, Z)$ for $Z \sim P_Z$ is close (in the same divergence D) to the true posterior $P_{X|Y=\mathbf{y}}$. Note that unlike Hagemann and Neumayer (2021), we use $P_{Y,T_{\mathbf{z}}(X)}$ instead of $P_{T_{\mathbf{y}}(X),T_{\mathbf{z}}(X)}$ as the second argument of $D(\cdot,\cdot)$ to simplify some of the technical arguments. We expect that our results can be extended to the unsupervised loss with $P_{T_{\mathbf{y}}(X),T_{\mathbf{z}}(X)}$, and we consider this choice in some of our experiments.

In the purely unsupervised setting, the same invertible models reduce to NFs. In particular, we have access to a dataset $\mathcal{D} = \{X_i \in \mathbb{R}^d\}_{i=1}^n$, and our goal is to estimate the distribution P_X of these observations. As in the case of INNs, the idea is to introduce a *latent variable* $Z \sim P_Z$, also in $\mathbb{R}^d$, and learn an invertible map $T : \mathbb{R}^d \to \mathbb{R}^d$ that minimizes some notion of divergence between P_Z and the forward map $T(X)$; that is,

$$T^*_{\text{NF}} \in \arg\min_{T\in\mathcal{T}} \; D(P_Z, P_{T(X)}).$$

By making the divergence between the distributions of Z and $T(X)$ small, the invertibility of T can be used to establish closeness (in the same divergence) between the distributions of the inverse map X, and $T^{-1}(Z)$, as desired. Within this formulation, in this paper, we view NFs as a purely unsupervised variant of the more general class of INNs.

The choice of divergence D in the training objective critically shapes the behavior of INNs (and NFs) as different choices encourage different behavior in the trained model. For instance, forward relative entropy (or KL divergence) induces mode coverage behavior, while reverse relative entropy encourages mode seeking. Jensen-Shannon (JS) divergence provides a natural compromise between these two extremes (Polyanskiy and Wu, 2025). Other popular choices include the Wasserstein metrics that are known to provide informative gradients, and kernel maximum mean discrepancy (MMD) which is an important instance of the family of integral probability metrics (IPMs) (Sriperumbudur et al., 2012). In practice, these choices influence not only the final model's eventual behavior but also the optimization dynamics, sample quality, and robustness to constraints such as limited data and architectural constraints.

1.1 Overview of our contributions

We now present an overview of our main contributions, which can be divided into the proposal of unifying variational framework, theoretical characterization of the quality of

approximation achieved by the empirical risk minimization (ERM)-based model, experimental case studies that verify and augment the theory, and a real-world ocean acoustic inversion application.

A unifying framework. The starting point of our work is the observation that a large class of divergence measures used to define the unsupervised loss term in the INN and NF training objective admit a variational (or dual) representation. This allows an interpretation of the training process as a saddle-point optimization problem. While we present the formal details in Section 3.1, we illustrate the idea using an f-divergence between two distributions on $\mathbb{R}^d$, denoted by $D_f(P \parallel Q)$, which admits the following classical Donsker-Varadhan (DV) type variational representation (Polyanskiy and Wu, 2025, Theorem 7.26):

$$D_f(P \parallel Q) = \sup_{g:\mathbb{E}_Q[f^*(g(X))]<\infty} \left\{\mathbb{E}_P[g(X)] - \mathbb{E}_Q[f^*(g(X))]\right\}, \quad \text{where} \quad f^*(b) = \sup_{a\in\mathbb{R}} ab - f(a)$$

denotes the convex conjugate of f (also known as Fenchel conjugate (Rockafellar, 1970)). With this formulation, the training of an INN given the dataset (augmented with samples from the latent distribution P_Z) $\mathcal{D} = \{(X_i, Y_i, Z_i) \sim P_{X,Y,Z}\}_{i=1}^n$, a model class $\mathcal{T}$, and a "critic class" $\mathcal{G} \subset \{g : \mathbb{E}_Q[f^*(g(X))] < \infty\}$, can be represented as

$$\hat{T}_n \in \arg\min_{T\in\mathcal{T}} \left\{ \frac{1}{n}\sum_{i=1}^n \|T_{\mathbf{y}}(X_i) - Y_i\|_2^2 + \lambda \sup_{g\in\mathcal{G}} \left(\frac{1}{n}\sum_{i=1}^n g(Y_i, Z_i) - \frac{1}{n}\sum_{j=1}^n f^*\big(g(Y, T_{\mathbf{z}}(X))\big) \right) \right\}.$$

Similar expressions can be obtained for the general family of IPMs, Wasserstein-1 (W_1) metric using Kantorovich-Rubinstein duality (Villani, 2008). We refer to this class of models as *Variational Invertible Neural Architectures* (VINA).

Theoretical analysis. In Section 3.2 (INNs) and Appendix B (NFs), we analyze the quality of the ERM model $\hat{T}_n$ under the unified variational framework. In both cases, we present our results at three "levels". At the first level (Theorem 1 and Theorem 14), we state how certain assumptions, such as realizability, uniform learnability, moment bounds, and representation ability of the critic class can allow us to convert the empirical loss into a high probability guarantee on the approximation quality of $\hat{T}_n$ in terms of the W_1 metric (see Remark 4 for justification of this choice). At the next level (Section 3.2.1 and Appendix B.1), we identify sufficient and more naturally verifiable conditions under which the requirements of the previous level are satisfied. In particular, we present conditions on the Rademacher complexity, Lipschitz regularity of the model class, and certain tightness properties associated with the latent distribution and model class. At the third and final level (Section 3.2.2 and Appendix B.2), we present a concrete instantiation of a practically relevant model class (iResNet models), and critic class (norm constrained reproducing kernel Hilbert spaces (RKHSs)), and verify that they satisfy all the conditions obtained at the previous level.

Unlike existing theoretical guarantees, our results incorporate some practical aspects of training INNs and NFs. Ardizzone et al. (2018) showed that if a model T achieves exactly zero population loss (supervised squared error + unsupervised kernel-MMD loss), then sampling via $T^{-1}(\mathbf{y}, Z)$ exactly recovers the posterior $P_{X|\mathbf{y}}$. In practice, however, zero population loss is seldom achieved. Hagemann and Neumayer (2021) addressed this issue by

bounding the W_1 distance between the true posterior and the estimated posterior in terms of the population loss (combination of supervised and unsupervised loss). However, their analysis places a rather strong bounded support assumption on the input and output of the network, which excludes common choices such as Gaussian latent variables. In this work, we relax this assumption by only requiring a finite-moment assumption which significantly broadens the scope of our result to more practical scenarios. In fact, we obtain the first formal quantification of the quality of the posterior distribution represented by INNs trained via empirical risk minimization (ERM) in terms of the Wasserstein (W_1) metric, and also discuss how it can be extended to other metrics such as kernel-MMD.

Empirical Results. We then perform a thorough empirical study to deliver practically useful insights and general design principles that complement the theoretical results mentioned above. More specifically, we study the impact of the different design choices involved in our framework, such as the effect of the choice of divergence / metric, the choice of the latent distribution, and the prior loss. The key insights obtained from these empirical evaluations are summarized below.

- In Section 4.1.1, we study the significance of integrating the knowledge of prior distribution on the input space ($\mathcal{X} = \mathbb{R}^{d_{\mathbf{x}}}$) into the training of INNs. Our empirical findings show that *employing a well-specified prior can significantly improve the performance of INNs*, but a misspecified can lead to a degradation in the quality of the generated samples.

- In Section 4.1.2, we compare coupling-based and iResNet architectures using f-divergence-based unsupervised losses in both forward and backward directions. We observe that training performance is influenced by multiple factors, including the network architecture, the training procedure, and the capacity of the critic class, with *backward f-divergence losses tending to yield lower inference error*. In line with prior work in the literature (e.g., Behrmann et al. (2019)), coupling-based architectures demonstrate improved efficiency relative to iResNet-based models.

- In Section 4.1.3, we observe *an approximately local U-shaped dependence between latent dimension and inference performance of INNs.* More specifically, as we change the latent dimension size $d_{\mathbf{z}}$ (and appropriately modify $d_{\mathbf{x}}$ by padding), we observe that increasing dimension initially improves results, and then starts degrading, and the overall trend is multimodal.

- In Section 4.1.4, we investigate the practical aspects of solving inverse problems using Wasserstein Distance (Villani, 2008). Calculating the Wasserstein distance is computationally challenging and susceptible to the curse of dimensionality (Fournier and Guillin, 2015). Here, we deploy the entropic estimate of the Wasserstein distance (W_2), both Sinkhorn approximation(Cuturi, 2013) and Sinkhorn divergence (Peyré et al., 2019), as a metric within our training process. We observe that *lower entropic regularization improves sample quality but increases training time for both Sinkhorn Divergence and Sinkhorn approximation of Wasserstein Distance, while Sinkhorn divergence demonstrates greater stability than the Sinkhorn approximation.*

- In Section 4.1.5, we study the effect of the support of the latent distribution when training with the KL divergence and the Sinkhorn divergence. It is well known that f-divergences perform poorly when the two distributions are mutually singular(Zhang et al., 2019, 2020), while Wasserstein distances can provide informative gradients even when supports do not overlap because it reflects the underlying geometric cost. Our empirical observations in this section validate this hypothesis, and the *models trained with the Sinkhorn divergence were more robust to mismatch between the supports of the latent and prior distributions.*

- In Section 4.1.6, we study the relation between the number of finite moments of X, and the performance of the NF model. As *predicted by our theoretical results*, we find that as the number of finite moment of X increases, the W_1 distance between the true and estimated distributions decreases.

Overall, our exploratory experiments in Section 4.1 provide empirical validation of some of the theoretical predictions of Section 3, and also provide some general practical insights in training invertible architectures.

Ocean-Acoustic Application. To demonstrate the practical relevance of our approach, we employ the insights gained from the exploratory experiments in Section 4.1 to the ocean–acoustic inversion setting of the SWellEx-96 experiment conducted off the coast of San Diego near Point Loma (Yardim et al., 2010; Meyer and Gemba, 2021). Geoacoustic inversion (GI), as demonstrated in the SWellEx-96 study, is a challenging and computationally intensive problem (Dosso and Dettmer, 2011; Huang et al., 2006; Chapman and Shang, 2021). This paper utilizes synthetic data to simplify the task for the initial application of an INN-based framework. These preliminary studies revealed that the latent-space dimensionality plays a critical role in estimating the posterior, and that physically meaningful priors, such as uniform priors for uncertain quantities like sound speed, improve stability. We demonstrate that invertible architectures can offer a favorable computational trade-off relative to likelihood-based MCMC sampling for GI, in which likelihood evaluations rely on repeated calls to the forward model KRAKEN (Porter, 1992). Although no method is universally best, moving this computational burden offline into training enables a pre-trained INN to support rapid, near-real-time posterior inference at test time.

To summarize, our work advances both the theory and practice of invertible neural architectures. We provide a first unified ERM analysis of invertible models that provides explicit bounds (in W_1 metric) on the quality of approximation achieved by the trained model under realistic moment and capacity assumptions. Through concrete instantiations we show that these conditions are satisfied by practically useful models, and finally we implement a series of empirical case studies and apply our ideas to a real-world ocean-acoustic inversion problem.

1.2 Organization of the paper

The remainder of the paper is organized as follows. We present a thorough review of the related work in Section 2, and then recall some background on the architecture and training details of existing invertible models. We present our main results in Section 3, and in particular, propose our variational training strategy in Section 3.1 and derive theoretical

results for INNs in Section 3.2. Also, we extend our theoretical analysis to variational NFs in Appendix B. We then move on to the empirical part of the paper in Section 4. In particular, in Section 4.1 we present a series of observations about the effect of various design choices on the performance of invertible models. Some of these observations verify the theoretical predictions from the previous section (such as the effect of the complexity of critic class, and the effect of the number of finite moments of the latent distributions), while others concern important aspects such as the latent dimension, choice of divergence measure etc. Finally, in Section 4.2, we apply our ideas to a real-world task of ocean-acoustic inversion.

2 Related Works

Invertible architectures (NFs and INNs): There are numerous ways to achieve invertibility in neural architectures (Tabak and Turner, 2013; Kobyzev et al., 2020; Keller et al., 2021), such as by using residual connections (Gomez et al., 2017; Jacobsen et al., 2018; Behrmann et al., 2019), triangularization (Bogachev et al., 2005; Marzouk et al., 2016; Parno et al., 2016; JM and Welling, 2017), and coupling-based normalizing flows (e.g., RealNVP Dinh et al. (2016) and Glow Kingma and Dhariwal (2018)). In practice, coupling-based architecture strikes the right balance between computational efficiency and expressivity. In fact, under appropriate assumptions, they have been shown to be universal diffeomorphism approximators (Teshima et al., 2020; Jin et al., 2024).

While NFs were developed largely for tractable density modeling, their supervised variants INNs have become increasingly popular for solving inverse problems. We explore this particular direction in Section 4.2. However, INNs (as defined in our paper) are not the only way to employ invertible neural networks to solve inverse problems. There exist other approaches, such as using conditional INNs (Ardizzone et al., 2019; Winkler et al., 2019). Another approach is to use conditional NFs to learn the likelihoods (Papamakarios et al., 2019), which can be used to estimate posterior by combining it with a prior distribution using Bayes rule. NFs have also been integrated with sampling-based Bayesian inference, to combine the advantages of both approaches (Song et al., 2017; Winter et al., 2023; Kruse et al., 2025), as will be discussed further below.

Our research focuses on INNs, which provide a probabilistic framework for addressing inverse problems. However, it's important to recognize the growing popularity of neural networks in solving inverse problem. For instance, In non-probabilistic contexts, works Ying (2022); Fan and Ying (2019); Khoo and Ying (2019) have designed specialized neural architectures that embed physical formulation to recover unknown parameters while reducing the reliance on large amounts of data.

INNs vs. sampling-based Bayesian methods: Conventional Markov Chain Monte Carlo (MCMC) methods (MacKay, 2003; Brooks et al., 2011; Andrieu et al., 2003; Doucet and Wang, 2005; Korattikara et al., 2014; Kungurtsev et al., 2023; Atchadé and Rosenthal, 2005) are widely used for sampling from complex probability distributions, including posterior distributions arising in inverse problems (Geweke, 1989). From a pushforward viewpoint, MCMC defines a Markov transition kernel with the target posterior as its stationary distribution; under standard ergodicity conditions, the chain converges asymptotically to the exact posterior. However, standard likelihood-evaluation-based MCMC methods often

require substantial computational time to achieve adequate sampling (Roy, 2020; Jones and Qin, 2022), which becomes a bottleneck, particularly in inverse problems where the forward process is computationally expensive to evaluate. Likelihood-free MCMC variants exist but can be substantially more computationally demanding in practice (Beaumont, 2019).

On the other hand, invertible generative models parameterize an invertible map that pushes a simple base distribution to an induced (learned) distribution that approximates the target distribution. When the Jacobian determinant is tractable, this learned distribution can be evaluated exactly via the change-of-variables formula. In context of inverse problem, once an INN is trained, approximating the posterior involves running the network in reverse, which is computationally inexpensive. We demonstrate that invertible architectures can offer a favorable computational trade-off relative to repeated likelihood-evaluation-based MCMC sampling for a practical inverse problem. This speedup comes from shifting computation offline into training. In fact, under appropriate assumptions, INNs can even perform in amortized settings, where a model trained extensively can generalize effectively to new instances without retraining (Radev et al., 2023, 2021).

Since practically relevant performance guarantees on invertible architecture are scarce in literature (we discuss some existing results later in this section), Gabrié et al. (2021, 2022); Brofos et al. (2022); Schönle and Gabrié (2023) have incorporated NFs to characterize the proposal distributions of MCMC methods, thereby accelerating convergence in practical tasks, while retaining theoretical guarantees for MCMC convergence. In this work, we establish error bounds for our proposed class of invertible models under appropriate assumptions, providing theoretical characterizations of posterior accuracy for INNs and generative accuracy for NFs. Taken together, our results clarify when offline-trained invertible maps with controlled approximation error can replace repeated online likelihood-based Monte Carlo sampling.

Posterior sampling in high dimensions is often computationally challenging (Montanari and Wu, 2023), and there is no universally best configuration for these problems. As demonstrated in our empirical studies, INNs exhibit sensitivity to several design and optimization choices such as the learning rate and architectural configuration. To address this, in our cases studies, we employed tree-structured Parzen estimator (Bergstra et al., 2011) to search for promising hyperparameters including: the size of the subnetworks, the latent dimensionality, and the number of coupling layers.

Unsupervised loss: The choice of training objective critically shapes the behavior of INNs as these models are trained through optimization. INNs were initially trained using MMD (Ardizzone et al., 2018), which only requires samples and avoids the need for explicit density modeling. However, due to challenges such as poor scaling in high dimensions and sensitivity to kernel choice (Ramdas et al., 2015), negative log-likelihood (NLL) (Dinh et al., 2016) has become more prevalent in recent INN architectures (Ren et al., 2020; Kruse et al., 2021). The use of NLL can be justified as an approximation of training by minimizing the KL-divergence (Kingma et al., 2019; Papamakarios et al., 2021). Training an INN using NLL in the literature is based on making assumptions about the distribution of the output samples, for instance Gaussian distribution has been used in prior work (Ren et al., 2020; Kruse et al., 2021). In this paper, we eliminate the need for such distributional assumptions by relying solely on samples and leveraging the variational formulation of f-divergences, as

introduced by Nguyen et al. (2010) and follow-up works (Nowozin et al., 2016; Ruderman et al., 2012; Ke et al., 2021).

Our approach is inspired by analogous formulations in related domains, such as generative adversarial networks (GANs) (Goodfellow et al., 2014; Nowozin et al., 2016), FlowGAN (Grover et al., 2018) and the Precision-Recall divergence for training NFs (Verine et al., 2023). To train invertible architectures, we take a unified look at the usage of a class of variational distance metrics over the space of probability measures, such as IPMs (Sriperumbudur et al., 2012), which include the MMD (Li et al., 2017) and Wasserstein metrics (Arjovsky et al., 2017; Coeurdoux et al., 2022), and the f-divergence family (Csiszár, 1967; Arjovsky and Bottou, 2017), and we study its implication in supervised inverse problems (variational INN) and in unsupervised generative modeling tasks (Variational NF). Training based on the variational representation of f-divergence enables stronger theoretical guarantees. By extending and unifying the results of Nguyen et al. (2010) and Hagemann and Neumayer (2021), we derive new convergence bounds for INN that are both tighter and hold under weaker assumptions than prior work. In particular, Hagemann and Neumayer (2021, Theorem 2) establish an approximation guarantee for the posterior distribution of an INN represented by a homeomorphism T in terms of the error bounds on the supervised and unsupervised losses. Our results extend their result in two main ways: first we relax a strong requirement of bounded observations imposed by Hagemann and Neumayer (2021) that omits usual choices like Gaussian latent distributions, and second, we analyze the performance of a data-driven model $\hat{T}_n$ trained using an ERM strategy, unlike Hagemann and Neumayer (2021) who work with a fixed T satisfying certain approximation guarantees.

Latent distribution: It is well established that the structure and topology of the latent space play a crucial role in the accuracy of generative models (Gurumurthy et al., 2017; Bevins et al., 2023; Stimper et al., 2022; Laszkiewicz et al., 2021; Hickling and Prangle, 2024; Fadel et al., 2021). While the most commonly used latent distributions for INNs is the standard Gaussian distribution (Ardizzone et al., 2018), the works in Behrmann et al. (2021); Hagemann and Neumayer (2021) shows that a simple Gaussian prior is often insufficient for multimodal problems and choosing a suitable latent distribution can improve robustness. In this paper, we study the effect of size of latent variable and support set of this variable in performance of INN.

A selective discussion of the existing architectural choices and design of INNs is provided in Appendix A.1.

3 Main Results

We begin this section by presenting a general unsupervised training cost for variational invertible neural architecture (VINA) in Section 3.1. As mentioned earlier, this formulation unifies several commonly used cost functions in the literature. For simplicity we focus our presentation in Section 3.1 for the case of INNs, but we then illustrate how it naturally applies to NFs in Appendix B.

Table 1: Definitions of important acronyms.

NF	normalizing flow	INN	invertible neural network
IK	inverse kinematics	BNN	Bayesian neural network
NLL	negative log-likelihood	MMD	maximum mean discrepancy
ERM	empirical risk minimization	KL	Kullback-Leibler
JS	Jensen-Shannon	IPM	integral probability metric
SL	supervised loss	USL	unsupervised loss
GAN	generative adversarial network	GI	Geoacoustic inversion

3.1 Variational unsupervised loss (USL)

Our main methodological contribution in this paper begins with the observation, also used in priors works such as Nowozin et al. (2016); Zhang et al. (2019); Grover et al. (2018), that several commonly used statistical divergence measures used in generative modeling admit variational representations. More specifically, with $T : \mathbb{R}^{d_{\mathbf{x}}} \to \mathbb{R}^{d_{\mathbf{y}}} \times \mathbb{R}^{d_{\mathbf{z}}}$ denoting an INN, we wish to minimize some divergence of the form

$$D_\phi(P_{Y,Z}, P_{Y,T_{\mathbf{z}}(X)}) = \sup_{g \in \mathcal{G}_\phi} \mathbb{E}\left[\phi(g, X, Y, Z, T)\right], \quad \overset{\text{emp. loss}}{\Longrightarrow} \quad \hat{L} = \sup_{g \in \mathcal{G}_\phi} \frac{1}{n} \sum_{i=1}^{n} \phi\left(g, X_i, Y_i, Z_i, T\right). \tag{2}$$

Here $\mathcal{G}_\phi$ denotes some class of functions, usually from $\mathcal{X} = \mathbb{R}^{d_{\mathbf{x}}}$ to $\mathbb{R}$, and we will refer to it as the "critic class" following the convention used in the generative adversarial network (GAN) literature (Goodfellow et al., 2014). In practice, this class can be represented by machine learning models such as neural networks, or more analytically tractable classes such as RKHSs. Our theoretical results in the next two subsections explore the trade-offs involved in choosing more expressive $\mathcal{G}_\phi$ and approximation guarantees. For different choices of $(\phi, \mathcal{G}_\phi)$, the above formulation recovers various popular loss functions such as relative entropy (and more generally, the f-divergence family), kernel-MMD (and more generally, the IPM family), energy distance, and the Wasserstein metric. Besides the conceptual unification, this approach also allows us to obtain theoretical guarantees on the approximation performance of INNs trained via empirical risk minimization strategy, as we discuss in Section 3.2 for INNs (and in Appendix B for NFs).

Before proceeding further, we recall two important family of distance or divergence measures that are realizations of (2). The first class of distance metrics are the IPMs, which for some function class $\mathcal{G}$ that is closed under negation (i.e., if $g \in \mathcal{G}$, then so does $-g$), are defined as

$$\text{IPM}(P_{Y,Z}, P_{Y,T_{\mathbf{z}}(X)}) = \sup_{g \in \mathcal{G}} \mathbb{E}\left[g(Y, Z) - g(Y, T_{\mathbf{z}}(X))\right]$$

which corresponds to $\phi_{\text{IPM}}(g, x, y, z, T) = g(y, z) - g(y, T_{\mathbf{z}}(x))$. Perhaps the most important element of the IPM family is the kernel-MMD metric, where the critic class (also known

as the witness class) is the unit norm ball RKHS associated with a positive-definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. An immediate consequence of the reproducing property of such $\mathcal{G}$ is that for square integrable kernels, we have

$$\text{MMD}(P_{Y,Z}, P_{Y,T_{\mathbf{z}}(X)}) = \|\mu_{P_{Y,Z}} - \mu_{P_{Y,T_{\mathbf{z}}(X)}}\|_k, \quad \text{where} \quad \mu_P = \int k(x,\cdot)dP(x),$$

is referred to as the kernel mean-embedding of a distribution P, $\|\cdot\|_k$ denotes the RKHS norm, and $\int$ above is a Bochner integral. This leads to a very natural interpretation of MMD metric between two distributions P and Q: it is the distance (as measured by the RKHS norm) between the two representative elements μ_P and μ_Q associated with the two distributions. Other important IPMs include the total variation distance, and the Wasserstein 1 metric (also known as the earth-mover distance), which by using Kantorovich-Rubinstein duality (Theorem 1.14 Villani (2021), and Theorem 11.8.2. Dudley (2018)) can be represented as an IPM associated with the class $\mathcal{G}$ of 1-Lipschitz functions.

The second important class of divergence measures captured by (2) is the family of f-divergences. For any convex, lower semicontinuous $f : (0,\infty) \to (0,\infty)$ with $f(1) = 0$, the f-divergence between $P_{Y,Z}$ and $P_{Y,T_{\mathbf{z}}(X)}$, assuming that $P_{Y,Z} \ll P_{Y,T_{\mathbf{z}}(X)}$, is defined as

$$D_f(P_{Y,Z} \parallel P_{Y,T_{\mathbf{z}}(X)}) = \int q_T(y, T_{\mathbf{z}}(x)) f\left(\frac{p_{Y,Z}(y,z)}{p_T(y, T_{\mathbf{z}}(x))}\right) \nu(x,y,z)dxdydz,$$

where we have assumed that both distributions admit densities p_T and $p_{Y,Z}$, with respect to some common dominating measure ν, for all choices of T in the model class. These divergence measures are also known to admit the following variational representation

$$D_f(P_{Y,Z} \parallel P_{Y,T_{\mathbf{z}}(X)}) = \sup_g \left(\mathbb{E}[g(Y,Z) - f^*(g(Y,T_{\mathbf{z}}(X)))]\right), \tag{3}$$

which corresponds to $\phi_f(g,x,y,z,T) = g(y,z) - f^*(g(y,T_{\mathbf{z}}(x))$. Here, the supremum is over all measurable $g : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ make the right-hand side (RHS) well-defined (by avoiding $\infty - \infty$). This implies that if we restrict our attention to any smaller class of functions $\mathcal{G}$, get a lower bound on the divergence, and this gap (that we later refer to as the variational gap) decreases by enlarging the critic class $\mathcal{G}$. This fact will play a central role in our analysis in the next two subsections.

The canonical member of the f-divergence family is the (forward) relative entropy (or Kullback-Leibler (KL) divergence), with generator function $f(u) = u\log u$. The corresponding ϕ function in (2) is $\phi_{KL}(g,x,y,z,T) = \log g(y,z) - g(T(x)) + 1$. Other important f-divergences are the squared Hellinger divergence with $f(u) = (\sqrt{u}-1)^2$ and $\phi_H(g,x,y,z,T) = g(y,z) - g(T(x))/(1-g(T(x)))$, and the Jensen Shannon divergence with $f(u) = -u(+1)\log((1+u)/2) + u\log u$ and $\phi_{JS}(g,x,y,z,T) = g(y,z) + \log\left(2 - e^{g(T(x))}\right)$.

We now proceed to a discussion of INNs defined using a variational objective as introduced in (2).

3.2 Variational invertible neural networks (V-INN)

Let $\mathcal{X} = \mathbb{R}^{d_{\mathbf{x}}}$, $\mathcal{Y} = \mathbb{R}^{d_{\mathbf{y}}}$ and $\mathcal{Z} = \mathbb{R}^{d_{\mathbf{z}}}$, with $d_{\mathbf{x}} = d_{\mathbf{y}} + d_{\mathbf{z}}$, and suppose we have a dataset $\{(X_i, Y_i, Z_i) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} : 1 \le i \le n\}$ drawn i.i.d. from the joint distribution P_{XYZ}. Our

goal is to learn an invertible neural network (INN) from a family of homeomorphisms (i.e., each $T : \mathcal{X} \to \mathcal{Y} \times \mathcal{Z}$ is a continuous bijective map with a continuous inverse) using this training dataset. Each candidate INN T can be decomposed into $T_{\mathbf{y}} : \mathcal{X} \to \mathcal{Y}$ and $T_{\mathbf{z}} : \mathcal{X} \to \mathcal{Z}$, and consequently, the INN training process involves population loss functions:

$$L_{\mathbf{y}}(T) = \mathbb{E}_{P_{XY}}\left[\|T_{\mathbf{y}}(X) - Y\|_2^2\right], \quad \text{and} \quad L_{\mathbf{z}}(T) \leq D(P_{Y,Z}, P_{Y,T_{\mathbf{z}}(X)}),$$

where $D(P, Q)$ is an appropriate notion of distance or divergence between probability distributions $P, Q \in \mathcal{P}(\mathcal{Y} \times \mathcal{Z})$. In this section, we will focus on the case of D being an f-divergence and employ its variational definition stated in (3) within the empirical risk minimization (ERM) framework. The use of an "$\leq$" instead of "$=$" when introducing $L_{\mathbf{z}}(T)$ above is a consequence of the variational definition with a restricted critic class as we make precise in **(INN4)** in Assumption 3.1. More explicitly, we work with a function class $\mathcal{G}_n$ (the "critic class"), and define $L_{\mathbf{z}}(T)$ as

$$L_{\mathbf{z}}(T) = \sup_{g \in \mathcal{G}_n} \{\mathbb{E}_{P_{YZ}}[g(Y, Z)] - \mathbb{E}_{P_{XY}}[f^*(g(Y, T_{\mathbf{z}}(X)))]\} \leq D_f(P_{Y,Z}, P_{Y,T_{\mathbf{z}}(X)}). \tag{4}$$

To train an INN from a family $\mathcal{T}$ (i.e., $\mathcal{T}$ might represent all INNs with a fixed architecture), we use the empirical analogs of the population loss terms,

$$\widehat{L}_{\mathbf{y},n}(T) = \frac{1}{n}\sum_{i=1}^{n} \|T_{\mathbf{y}}(X_i) - Y_i\|_2^2, \; \widehat{L}_{\mathbf{z},n}(T) = \sup_{g \in \mathcal{G}_n} \left\{ \frac{1}{n}\sum_{i=1}^{n} g(Y_i, Z_i) - \frac{1}{n}\sum_{i=1}^{n} f^* \circ g(Y_i, T_{\mathbf{z}}(X_i)) \right\},$$

where f^* denotes the convex conjugate of f. The model learned by ERM can then be defined as

$$\hat{T}_n \in \arg\min_{T \in \mathcal{T}} \; \widehat{L}_{\mathbf{y},n}(T) + \lambda \widehat{L}_{\mathbf{z},n}(T), \tag{5}$$

for some regularization parameter $\lambda > 0$. The first loss term ($\widehat{L}_{\mathbf{y},n}$) forces the learned map $\hat{T}_n$ to approximate Y from an input X, while the second loss term ($\widehat{L}_{\mathbf{z},n}$) ensures that the joint law of $(Y, \hat{T}_{n,\mathbf{z}}(X))$ matches the true joint distribution (Y, Z). Thus, intuitively, if the sample-size n is large enough, we should expect $\hat{T}_n$ to be a good proxy for T^*, an element of $\mathcal{T}$ that minimizes the population risk:

$$T^* \in \arg\min_{T \in \mathcal{T}} \; L_{\mathbf{y}}(T) + \lambda L_{\mathbf{z}}(T).$$

The performance of $\hat{T}_n$ defined above is governed by two effects: (i) the finite sample effect caused by the uniform deviations between the empirical objectives $\widehat{L}_{\mathbf{y},n}$ and $\widehat{L}_{\mathbf{z},n}$ from their population counterparts; and (ii) the variational or approximation error induced by restricting the divergence representation to a class $\mathcal{G}_n$. To make the variational loss smaller, we need to increase the capacity of $\mathcal{G}_n$. However, that also causes the uniform deviations between the empirical and population losses to increase. Thus, the crucial challenge is to find the right trade-offs between these two effects to simultaneously drive the overall error to zero.

For the rest of this section, we follow a three-step roadmap. *First*, we state a general result (Theorem 1) that obtains an upper bound on the quality of the posterior approximation provided by $\hat{T}_n$ under a set of high-level assumptions (stated in Assumption 3.1).

Second, we present more concrete, verifiable, sufficient conditions for satisfying the requirements of Assumption 3.1 in Section 3.2.1. *Finally*, we specialize the general theorem to a specific choice of INN architecture (iRes-Nets in Definition 9) and critic class (RKHS over truncated domain in Definition 10) using Jensen-Shannon divergence in Section 3.2.2.

We now present the assumptions required for stating the main result of this section.

Assumption 3.1 *To analyze the performance of the INN model $\hat{T}_n$ defined in* (5)*, we place the following assumptions:*

- ***(INN1): Realizability and Bi-Lipschitz.*** *There exists a $T^* \in \mathcal{T}$ such that $Y = T^*_{\mathbf{y}}(X)$ almost surely (a.s.) and $P_{Y,Z} = P_{Y,T^*_z(X)}$, and furthermore*

$$\sup_{T\in\mathcal{T}} \left\{\mathrm{Lip}(T),\, \mathrm{Lip}(T^{-1})\right\} \leq J < \infty.$$

- ***(INN2): Uniform Convergence.*** *For any confidence level $\delta \in (0,1)$, there exist deterministic sequences $\{r_n \equiv r_n(\delta) : n \geq 1\}$ and $\{u_n \equiv u_n(\delta) : n \geq 1\}$, with $r_n, u_n \to 0$, such that the following conditions hold with probability at least $1-\delta$:*

$$\sup_{T\in\mathcal{T}} |\widehat{L}_{\mathbf{y},n}(T) - L_{\mathbf{y}}(T)| \leq u_n, \quad \textit{and} \quad \sup_{T\in\mathcal{T}} |\widehat{L}_{\mathbf{z},n}(T) - L_{\mathbf{z}}(T)| \leq r_n.$$

- ***(INN3): Moment Bounds.*** *There exists an $R < \infty$, such that the following holds for some $a > 0$:*

$$\max\left\{\mathbb{E}[\|(Y,Z)\|^{1+a}],\, \sup_{T\in\mathcal{T}} \mathbb{E}[\|(Y, T_{\mathbf{z}}(X)\|^{1+a}]\right\} \leq R.$$

  Here $\|\cdot\|$ to denotes the ℓ_2 norm, and we use $\|(y,z)\|$ as a shorthand for $\sqrt{\|y\|^2 + \|z\|^2}$.

- ***(INN4): Variational Approximation Gap.*** *The function class $\mathcal{G}_n$ contains the $\mathbf{0}$ function for all $n \geq 1$, and there exists a vanishing deterministic sequence $\eta_n \to 0$, such that*

$$\sup_{T\in\mathcal{T}} D_f(P_{Y,Z} \parallel P_{Y,T_{\mathbf{z}}(X)}) - L_z(T) \leq \eta_n.$$

  In other words, we assume that the capacity of the critic class $\mathcal{G}_n$ grows with n to approximate the likelihood ratios of all distributions modeled by elements of $\mathcal{T}$ and the true data distribution $P_{Y,Z}$.

Theorem 1 *Suppose Assumption 3.1 holds, and D_f satisfies the Pinsker-type inequality $c_f\sqrt{D_f(P \parallel Q)} \geq TV(P,Q)$ for all distributions P, Q, and for some constant $c_f > 0$. Then, for every measurable $A \subset \mathcal{Y}$ with $P_Y(A) > 0$, on the $(1-\delta)$ probability event of **(INN2)**, we have*

$$W_1\left(P^{(A)}_X, P^{(A)}_{\hat{T}_n^{-1}(Y,Z)}\right) \lesssim (r_n + u_n + \eta_n)^{\frac{a}{2(1+a)}},$$

where W_1 denotes the 1-Wasserstein metric, $P^{(A)}_X = P_{X|Y\in A}$, $P^{(A)}_{\hat{T}_n^{-1}(Y,Z)} = P_{\hat{T}_n^{-1}(Y,Z)|Y\in A}$, and $\lesssim$ suppresses the constant factors and lower order terms. The exact expression of the upper bound is in (32)*.*

The proof of this result is in Appendix C.1. Note that the Pinsker-type inequality $TV(P,Q) \leq c_f\sqrt{D_f(P \parallel Q)}$ is satisfied by several divergences, such as relative entropy, Jensen Shannon, Hellinger, and chi-squared divergence.

Remark 2 *The result of Theorem 1 is obtained by combining two ingredients: (i) using the Pinsker-type inequality, we convert the control over the given f-divergence into control over total variation, and (ii) a truncation argument (Lemma 25) that converts the total variation into a bound on W_1 under just finite $(1+a)$ moment condition.*

One crucial advantage of our result over Hagemann and Neumayer (2021, Theorem 2) is that we do not impose the bounded support condition on the random variables (X,Y,Z). The mild $(1+a)$ moment requirement significantly broadens the applicability of our result, and in particular, allows us to consider more realistic models used in practice. Additionally, the dependence on the parameter a quantifies how the heavy tails affect the approximation quality. We empirically verify this insight in Section 4.1.6.

Remark 3 *Although Theorem 1 characterizes the distance between the true (and unknown) posterior and the INN posterior in terms of W_1 metric, we can use existing inequalities to translate it into other distances. For example, by Sriperumbudur et al. (2010, Theorem 21), we know that $\mathrm{MMD}(P_X^{(A)}, P^{(A)}_{\hat{T}_n^{-1}(Y,Z)}) \lesssim W_1(P_X^{(A)}, P^{(A)}_{\hat{T}_n^{-1}(Y,Z)})$ for commonly used kernels, such as the Gaussian and Matérn kernels, and thus Theorem 1 also implies a bound on the kernel-MMD distance between the true and estimated posteriors.*

Remark 4 *A closer look at the proof of Theorem 1 reveals several reasons why W_1 metric is the appropriate choice to present the result: (i) By Kantorovich-Rubinstein duality, the W_1 metric is defined via Lipschitz critic functions. This fact coupled with the assumed bi-Lipschitz structure **(INN1)** allows us to transfer bounds from posterior distributions to the forward distributions. (ii) Since W_1 is a metric, it satisfies the triangle inequality which leads to a natural decomposition in the proof into terms that correspond directly to the supervised and unsupervised training losses. (iii) The Lipschitz formulation of W_1 is exactly what makes our truncation argument(Lemma 25) effective for handling heavy tailed distributions under a mild $(1+a)$ moment assumption.*

Remark 5 *While stating Theorem 1, we only needed to explicitly place a $(1+a)$ moment requirement, for some $a>0$. However, since $L_{\mathbf{y}}$ consists of squared losses, the existence of the uniform convergence rates in **(INN2)** implicitly places a stronger $(2+\beta)$ moment conditions, for some $\beta>0$. We will make this explicit when identifying verifiable sufficient conditions of these assumptions in Proposition 6.*

3.2.1 Verifiable Sufficient Conditions for Assumption 3.1

Theorem 1 gives us a general result that translates uniform convergence and the variational approximation gap guarantees on the two losses, $L_{\mathbf{y}}(\hat{T}_n)$ and $L_{\mathbf{z}}(\hat{T}_n)$ into a bound on the W_1 metric between the true posterior, and the INN posterior. We now present more easily verifiable sufficient conditions for the assumptions **(INN2)-(INN4)** required by Theorem 1, working under the realizability part of the assumption **(INN1)**. We begin with the usual characterization of uniform convergence in terms of the Rademacher complexities (see Definition 13 in Appendix A) of the associated function classes.

Proposition 6 *Assume that the critic class $\mathcal{G}_n$ consists of uniformly bounded functions; that is, $\sup_{g\in\mathcal{G}_n}\|g\|_\infty \le b_n$, for some $b_n < \infty$. Let $\mathcal{F}_{\mathbf{y}} = \{x \mapsto \langle u, T_{\mathbf{y}}(x)\rangle : T \in \mathcal{T}, \|u\| \le 1\}$, and define its Rademacher complexity as*

$$\mathfrak{R}_n(\mathcal{F}_{\mathbf{y}}) = \mathbb{E}_{\epsilon^n, X^n}\left[\sup_{f\in\mathcal{F}_{\mathbf{y}}} \sum_{i=1}^n \epsilon_i f(X_i)\right], \quad \textit{with} \quad \epsilon^n \overset{i.i.d.}{\sim} \text{Rademacher}(\{-1,+1\}).$$

Introduce the two critic class complexities

$$\mathfrak{R}_n^{(1)}(\mathcal{G}_n) = \mathbb{E}_{\epsilon^n, Y^n, Z^n}\left[\sup_{g\in\mathcal{G}_n} \frac{1}{n}\sum_{i=1}^n \epsilon_i g(Y_i, Z_i)\right],$$
$$\mathfrak{R}_n^{(2)}(\mathcal{G}_n, \mathcal{T}) = \mathbb{E}_{\epsilon^n, X^n, Y^n}\left[\sup_{T,g} \frac{1}{n}\sum_{i=1}^n \epsilon_i g(Y_i, T_{\mathbf{z}}(X_i))\right].$$

Suppose $\max\{\mathbb{E}[\|Y\|^{2+\beta}], \sup_{T\in\mathcal{T}}\mathbb{E}[\|T_{\mathbf{y}}(X)\|^{2+\beta}]\} \le R_{\mathbf{y}} < \infty$ *for some $\beta > 0$, and $\Pi_{K_n} : \mathbb{R}^{d_{\mathbf{y}}} \to \{y \in \mathbb{R}^{d_{\mathbf{y}}} : \|y\| \le K_n\}$ denoting the projection on a ball of radius K_n in $\mathcal{Y} = \mathbb{R}^{d_{\mathbf{y}}}$, define the projected empirical and population supervised losses:*

$$\widehat{L}^{K_n}_{\mathbf{y},n}(T) = \frac{1}{n}\sum_{i=1}^n \|\Pi_{K_n}(T_{\mathbf{y}}(X_i)) - \Pi_{K_n}(Y_i)\|^2, \quad \textit{and} \quad L^{K_n}_{\mathbf{y}}(T) = \mathbb{E}\left[\|\Pi_{K_n}(T_{\mathbf{y}}(X)) - \Pi_{K_n}(Y)\|^2\right].$$

Assume that the function f (in fact, its convex conjugate f^*) satisfies

$$A_{1,n} \coloneqq \sup_{|u|\le b_n} |(f^*)'(u)| < \infty, \quad \textit{and} \quad A_{2,n} \coloneqq \sup_{|u|\le b_n} |f^*(u)| < \infty. \tag{6}$$

Then, the following two conditions hold with probability at least $(1-\delta)$, for a given δ in$(0,1)$:

$$\sup_{T\in\mathcal{T}} |\widehat{L}_{\mathbf{z},n}(T) - L_{\mathbf{z}}(T)| \lesssim \mathfrak{R}_n^{(1)} + A_{1,n}\mathfrak{R}_n^{(2)} + (b_n + A_{2,n})n^{-1/2}, \tag{7}$$
$$\sup_{T\in\mathcal{T}} |\widehat{L}_{\mathbf{y},n}(T) - L_{\mathbf{y}}(T)| \lesssim K_n\mathfrak{R}_n(\mathcal{F}_{\mathbf{y}}) + K_n^2 n^{-1/2} + 4R_{\mathbf{y}}K_n^{-\beta}, \tag{8}$$

*where $\lesssim$ suppresses constants that depend on δ and $R_{\mathbf{y}}$. Thus, a sufficient condition to satisfy **(INN2)** with vanishing sequences of $\{r_n, u_n : n \ge 1\}$, is to select an appropriate sequence of K_n, b_n, and $\mathcal{G}_n$ to drive both these terms to 0.*

The proof of (7) relies on some standard arguments from empirical process theory (Shorack and Wellner, 2009), while the justification of (8) requires a combination of symmetrization and vector contraction along with a truncation argument, and we present the details in Appendix C.2.

Next, we present a simple result stating that the uniform moment bound required in **(INN3)** can be satisfied if the model class $\mathcal{T}$ is uniformly Lipschitz. As we will observe in Section 3.2.2, this condition holds for an important class of INNs.

Proposition 7 *Suppose there exist constants $J_0, J < \infty$, such that with $\|\cdot\|$ denoting the ℓ_2 norm, we have $\|T(x)\| \le J_0 + J\|x\|$, for all $x \in \mathcal{X}$, and for all $T \in \mathcal{T}$. Then, assuming*

that $\mathbb{E}[\|X\|^{1+a}] < \infty$ *for some* $a > 0$*, the following uniform moment bounds hold under the realizability assumption:*

$$\max\left\{\mathbb{E}\left[\|(Y,Z)\|^{1+a}\right], \sup_{T\in\mathcal{T}} \mathbb{E}\left[\|(Y,T_{\mathbf{z}}(X)\|^{1+a}\right]\right\} \leq R, \quad \textit{with}$$
$$R = 2^{2a+1}\left(J_0^{1+a} + J^{1+a}\mathbb{E}[\|X\|^{1+a}]\right).$$

Proof Since we are working under the realizability assumption, we have $(Y,Z) \overset{d}{=} T^*(X)$. Using the fact that $(x+y)^p \leq 2^{p-1}(x^p + y^p)$, with $p = 1+a$, and the condition that $\|T(x)\| \leq J_0 + J\|x\|$, we have $\mathbb{E}\left[\|(Y,Z)\|^p\right] \leq 2^{p-1}\left(J_0^p + J^p\mathbb{E}[\|X\|^p]\right)$. Next, we look at the term $(Y, T_{\mathbf{z}}(X))$ and observe that

$$\|(Y,T_{\mathbf{z}}(X))\|^p \leq 2^{p-1}\left(\|Y\|^p + \|T_{\mathbf{z}}(X)\|^p\right) \leq 2^{p-1}\left(\|(Y,Z)\|^p + \|T(X)\|^p\right)$$

which implies that

$$\begin{aligned}\mathbb{E}\left[\|(Y,T_{\mathbf{z}}(X))\|^p\right] &\leq 2^{p-1}\mathbb{E}[\|(Y,Z)\|^p] + 2^{2(p-1)}\left(J_0^p + J^p\mathbb{E}[\|X\|^p]\right)\\ &\leq 2\times 2^{2(p-1)}\left(J_0^p + J^p\mathbb{E}[\|X\|^p]\right).\end{aligned}$$

Taking the maximum of the two bounds gives us the required expression for R. ■

Proposition 8 *Introduce the notation* $P \equiv P_{YZ}$ *and* $Q_T \equiv P_{Y,T_{\mathbf{z}}(X)}$*, and assume that for all* $T \in \mathcal{T}$*, we have* $P \ll Q_T$*. For each* $T \in \mathcal{T}$*, let* g_T^* *denote a maximizer in the definition of the population* $L_{\mathbf{z}}(T)$ *in* (4)*, and let* $\bar{g}_T = \mathrm{clip}(g_T^*, -b_n, b_n)$ *denote its clipped version for* $b_n > 0$*. For some* $K_n > 0$*, let* B_{K_n} *denote the ball* $\{u \in \mathbb{R}^{d_{\mathbf{y}}+d_{\mathbf{z}}} : \|u\| \leq K_n\}$*, and define the following terms:*

- *An approximation error term* $\delta_n \equiv \delta_n(b_n, K_n)$ *defined as*

$$\delta_n \equiv \delta_n(b_n,K_n) \coloneqq \sup_{T\in\mathcal{T}}\inf_{g\in\mathcal{G}_n}\left\{\mathbb{E}_P[|\bar{g}_T - g|\mathbf{1}_{B_{K_n}}] + \mathbb{E}_{Q_T}[|\bar{g}_T - g|\mathbf{1}_{B_{K_n}}]\right\}.$$

- *The tightness term* $\tau_1 \equiv \tau_1(K_n)$*, and the clipping error term* $\tau_2 \equiv \tau_2(b_n)$*, as*

$$\begin{aligned}\tau_1(K_n) &= \sup_{T\in\mathcal{T}} P(B_{K_n}^c) + Q_T(B_{K_n}^c),\\ \tau_2(b_n) &= \sup_{T\in\mathcal{T}}\left\{\mathbb{E}_P[(|g_T^*| - b_n)^+] + A_{1,n}\mathbb{E}_{Q_T}\left[(|g_T^*| - b_n)^+\right]\right\},\end{aligned}$$

  where we use $(a)^+$ *to denote* $\max\{0,a\}$ *for any* $a \in \mathbb{R}$*.*

Suppose there exists a sequence of $\{K_n, b_n : n \geq 1\}$ *such that*

$$\lim_{n\to\infty}\left\{\tau_2 + A_{1,n}\delta_n + (b_n + A_{2,n})\,\tau_1\right\} = 0, \tag{9}$$

where $A_{1,n}$ *and* $A_{2,n}$ *were defined in* (6)*. Then the variational gap* η_n *also converges to zero; that is,*

$$\textit{if (9) is true,} \quad \Longrightarrow \quad \eta_n = \sup_{T\in\mathcal{T}}\left\{D_f(P_X \parallel Q_T) - L_{\mathbf{z}}(T)\right\} \overset{n\to\infty}{\longrightarrow} 0.$$

The proof of this result is in Appendix C.3. The structure of (9) is worth further discussion. The term δ_n is a purely approximation property of the critic class $\mathcal{G}_n$, and it enforces the requirement that every optimal critic function can be approximated in L^1 by a bounded element in $\mathcal{G}_n$ with vanishing error. The terms τ_1 and τ_2 can be interpreted as tail approximation terms: τ_1 is controlled by the moment bounds from **(INN3)**, while τ_2 depends on how quickly the tails of the optimal critic function decays with the clipping level b_n.

To summarize, the results of this section provide a concrete recipe to apply the more abstract Theorem 1 to a particular instantiation of the variational INN pipeline. In particular, for a chosen model and critic class pair, it suffices to show that the Rademacher complexities of the associated function classes decay sufficiently quickly to 0 (Proposition 6), to establish a uniform bi-Lipschitz property, and through it the moment bounds (Proposition 7), and show that the critic class and well-approximate the optimal witness function on high probability regions (Proposition 8). In the next subsection, we illustrate these steps for a specific variational INN architecture with Gaussian RKHS critic class.

3.2.2 INN Example

In this section, we instantiate our abstract assumptions and conditions from Section 3.2.1 and Assumption 3.1 with a concrete class of INN models and critic function classes. The main goal of this section is demonstrate the the conditions derived in the previous sections are satisfied by a non-trivial, reasonably expressive, and practically relevant pair of model and critic classes. In particular, we show that a residual INN with bounded weights and biases, trained with a Jensen-Shannon based loss and a Gaussian RKHS critic class, fits into our framework developed in the previous sections. We begin with a formal description of the model class.

Definition 9 *As before, we assume $d_{\mathbf{x}} = d_{\mathbf{y}} + d_{\mathbf{z}}$, and for an integer $M \geq 1$ and $s \in (0,1)$, define $T : \mathbb{R}^{d_{\mathbf{x}}} \to \mathbb{R}^{d_{\mathbf{x}}}$ as follows, with $I_{d_{\mathbf{x}}}$ denoting the $d_{\mathbf{x}} \times d_{\mathbf{x}}$ identity matrix:*

$$T = (I_{d_{\mathbf{x}}} + F_M) \circ (I_{d_{\mathbf{x}}} + F_{M-1}) \circ \cdots \circ (I_{d_{\mathbf{x}}} + F_1), \quad F_j(x) \coloneqq W_{j,2} \tanh(W_{j,1}x + b_{j,1}) + b_{j,2},$$

with the following constraints for all $j \in [M] = \{1, \ldots, M\}$:

$$\|W_{j,1}\|_{op} \leq s, \quad \|W_{j,2}\|_{op} \leq s, \quad \max\{\|b_{j,1}\|, \|b_{j,2}\|\} \leq B.$$

For every $j \in [M]$, the matrices $W_{j,1}$ and $W_{j,2}$ have dimensions $d_j \times d_{\mathbf{x}}$ and $d_{\mathbf{x}} \times d_j$ respectively, and let $H \coloneqq \max_{j \in [M]} d_j$. We split the output coordinates to write $T(x) = (T_{\mathbf{y}}(x), T_{\mathbf{z}}(x))$, with $T_{\mathbf{y}}(x) \in \mathcal{Y} = \mathbb{R}^{d_{\mathbf{y}}}$ and $T_{\mathbf{z}}(x) \in \mathcal{Z} = \mathbb{R}^{d_{\mathbf{z}}}$. It is easy to verify that this function class is bi-Lipschitz and there exist constants J_0, J (depending on M, B, s, H) such that for all $x \in \mathcal{X}$, we have $\|T(x)\| \leq J_0 + J\|x\|$ and $\|T^{-1}(y,z)\| \leq J_0 + J\|(y,z)\|$.

This choice of the INN model class $\mathcal{T}$ above is driven by the need for a global bi-Lipschitz bounds on both the forward and inverse maps. The residual architecture with tanh activations and the norm constraints on the weight and bias terms provide a convenient way of controlling global Lipschitz constants. In contrast the popular coupling-based architectures such as RealNVP of Dinh et al. (2014) generally do not admit such global control of Lipschitz constants. We now present the details of our critic class construction.

Definition 10 *Let $k_n \equiv k_{\gamma_n}$ denote the Gaussian kernel on $\mathbb{R}^{d_\mathbf{y}+d_\mathbf{z}}$,*

$$k_n((y,z),(y',z')) = \exp\left(-\gamma_n\|(y,z)-(y',z')\|^2\right) = \exp\left(-\gamma_n\left(\|y-y'\|^2+\|z-z'\|^2\right)\right),$$

and let $\mathcal{H}_{k_n}$ denote its RKHS. Fix a radius parameter $K_n > 0$ and a uniform upper bound $b_n > 0$ and define the critic class as

$$\mathcal{G}_n = \left\{g = h\mathbf{1}_{C_{K_n}} \;:\; h \in \mathcal{H}_{k_n},\ \|h\|_{k_n} \le b_n\right\}, \quad \textit{with} \quad C_{K_n} = [-K_n, K_n]^{d_\mathbf{x}}.$$

Note that $\mathbf{0} \in \mathcal{G}_n$, and each $g \in \mathcal{G}_n$ also satisfies $\|g\|_\infty \le b_n$ due to the reproducing property and the fact that $\sup_x \sqrt{k_n(x,x)} = 1$.

This particular choice of the critic class $\mathcal{G}_n$ is chosen as a compromise between analytical tractability and representation power. The RKHS associated with Gaussian kernels admit clean bounds on their Rademacher complexity, and are also known to be *universal* in the sense that they are approximate any continuous bounded function over compact domains arbitrarily well in sup norm. These are the exact properties that we need in Proposition 6 to ensure uniform convergence, and in Proposition 8 for controlling the variational approximation gap. The additional truncation on cube C_{K_n} by the indicator function is a purely technical choice: it ensures the uniform boundedness of the critic class while also decoupling the approximation on C_{K_n} from tail events, which can be handled separately via moment bounds. Furthermore, since the ℓ_2-ball B_{K_n} used by Proposition 8 is contained in C_{K_n}, any approximation bound C_{K_n} is also valid on B_{K_n}. We now present the choice of the f-divergence to be used for the unsupervised loss.

Definition 11 *We will employ a Jensen-Shannon divergence based unsupervised loss in our INN training. This corresponds to $f(t) = x\log\frac{2x}{x+1} + \log\frac{2}{x+1}$, and $f^*(t) = \log(1+e^t) - \log 2$ with*

$$A_{1,n} = \sup_{|t|\le b_n} |(f^*)'(t)| \le 1, \quad \textit{and} \quad A_{2,n} = \sup_{|t|\le b_n} |f^*(t)| \le \log((1+e^{b_n})/2) \le b_n.$$

Having introduced all the main components, we can now present the main result of this section.

Theorem 12 *Let $(\mathcal{T}, \mathcal{G}_n)$ denote a pair of model and critic classes as introduced in Definition 9 and Definition 10 respectively, and define the parameters*

$$\begin{aligned} K_n &= M(s\sqrt{H}+B) + \sqrt{d_\mathbf{x}} + \sqrt{2\log n}, \\ \gamma_n &= K_n^{2+\epsilon} \quad \textit{for some fixed } \epsilon > 0, \\ b_n &= C_b\gamma_n^{d_\mathbf{x}/4}K_n^{2+d_\mathbf{x}/2} = C_b K_n^{2+d_\mathbf{x}+\frac{d_\mathbf{x}}{4}\epsilon}. \end{aligned}$$

*Then, under the realizability assumption **(INN1)**, and with these choices of the parameters, the ERM model $\hat{T}_n$ defined in (5) satisfies*

$$W_1\left(P_{X|Y\in A},\, P_{\hat{T}_n^{-1}(Y,Z)|Y\in A}\right) = o(1), \quad \textit{w.p. at least } 1-\delta.$$

*In other words, with these parameters, the models introduced above satisfy the sufficient conditions for **(INN2)**-**(INN4)** to hold, as derived in Section 3.2.1.*

The proof of this result is in Appendix C.4.

As mentioned earlier, the purpose of this theorem is to illustrate that our abstract theoretical results are applicable to a non-trivial, reasonably realistic variational INN pipeline. In particular, Theorem 12 demonstrates that one can work with architectures that are expressive enough to model high-dimensional data and still verify the technical assumptions required for our posterior accuracy guarantees. Extending our arguments to other classes of invertible architectures, such as coupling-based flows equipped with appropriate spectral or Jacobian regularization, is an interesting direction for future work.

4 Empirical Results

In this section, we first present a series of small-scale probing experiments to understand the effects of various design choices involved in training invertible models (§ 4.1), and then apply the insights gained to a practically relevant ocean-acoustic inversion problem (§ 4.2). Our objective in Section 4.1 is two-fold:

- The first objective is to provide empirical support for several theoretical claims derived from the results of the previous section on INNs and from Appendix B on NFs. In particular, we investigate the use of the variational USL framework Section 3.1 for training invertible architectures, and we compare the performance of different f-divergences with that of IPMs. We further analyze the behavior of the forward and backward training losses when variational f-divergences are employed. Additionally, we empirically examine the main theoretical result for NFs presented in Appendix B by studying the relationship between the number of finite moments of X and the W_1 distance between the true and the estimated distributions. The observed polynomial convergence rates are consistent with our theoretical predictions and analogous to the behavior is established for INNs in Theorem 1.

- The other direction is empirically explore the effects of certain design choices in practical performance of INNs. This includes the prior loss or reconstruction loss, the effect of dimensionality of the latent space, and role of the support set, and the effect of entropy regularization in Wasserstein metric. This empirical analysis complements the existing literature and providing insights for applied implementations.

The software accompanying this paper is available on GitHub, https://github.com/ananya-ac/INN-Project.

4.1 Effect of design choices on INN training

As mentioned earlier, in this section, we study the effects of various design choices (such as prior distribution, latent dimension, entropic regularization parameter, etc.) on the training of invertible neural networks. Before presenting the empirical studies in detail, we first clarify two general aspects. First, unless otherwise stated, all case studies were trained for 10 epochs using a batch size of 512. The model parameters were optimized with the Adam optimizer with a learning rate of 10^{-4} and ℓ_2 regularization. Second, since the NLL is a widely adopted loss function for trainingINNs (Dinh et al., 2016; Ren et al., 2020; Kruse et al., 2021), we incorporate the NLL in our case studies, complementing the

USL described in Section 3.1 (the relevant definitions are reviewed in Appendix A.2). Also, as discussed in Section 2, while the Gaussian assumption underlying the computation of the NLL can be restrictive, we avoid such distributional assumptions by leveraging the variational formulation of f-divergences.

4.1.1 Effect of prior on training

In this case study, we use the inverse kinematics (IK) framework described next.

Inverse kinematics example: IK is a core problem in robotics, where the objective is to determine the joint configurations required to achieve a desired end-effector position in space (Niku, 2020). Efficiently solving this problem is essential for tasks like motion control, path planning, and real-time manipulation. Traditional methods such as analytical solutions and numerical optimization often face challenges as the robot configurations become complicated and number of degrees of freedom (DOF) increases. Machine learning algorithms are increasingly being employed for solving IK problems (Toquica et al., 2021), with the 4 DOF robot example becoming a standard benchmark in analyzing the analyzing the practical aspect of training invertible architecture INN (Ardizzone et al., 2018; Kruse et al., 2021; Hagemann and Neumayer, 2021).

We consider an articulated robotic arm that moves vertically along a rail and rotates at three joints. These four degrees of freedom constitute the parameter vector $X = [X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)}]^\top$. The dataset is generated using Gaussian priors defined as $X^{(i)} \sim \mathcal{N}(0, \sigma_i)$, where the standard deviations are specified as $\sigma_1 = 0.25$ and $\sigma_2 = \sigma_3 = \sigma_4 = 0.5$, and the forward kinematic process of the robotics arm is modeled as

$$
\begin{aligned}
Y^{(1)} &= X^{(1)} + l_1 \sin(X^{(2)}) + l_2 \sin(X^{(3)} - X^{(2)}) + l_3 \sin(X^{(4)} - X^{(2)} - X^{(3)}); \\
Y^{(2)} &= l_1 \cos(X^{(2)}) + l_2 \cos(X^{(3)} - X^{(2)}) + l_3 \cos(X^{(4)} - X^{(2)} - X^{(3)}),
\end{aligned}
$$

where the arm segment lengths are given by $l_1 = 0.5$, $l_2 = 0.5$, and $l_3 = 1.0$.

We trained an INN to model the relationship between the joint angles of a robotic arm and a point in its 2D workspace. All models were trained using the NLL loss formulation for the INN. The input consisted of 4 variables corresponding to the joint angles of the robotic arm, and the output was a 2-dimensional vector representing a point in the 2D workspace.

For this case study, we use two different regularization priors: a Gaussian prior and a uniform prior. The Gaussian prior represents the correct prior (Normal prior with $\mu = [0, 0, 0, 0]$ and $\sigma = [0.25, 0.5, 0.5, 0.5]$) as the data-generation process follows the same Gaussian distribution. The corresponding prior loss can be computed using the negative log-likelihood, up to a constant, as follows

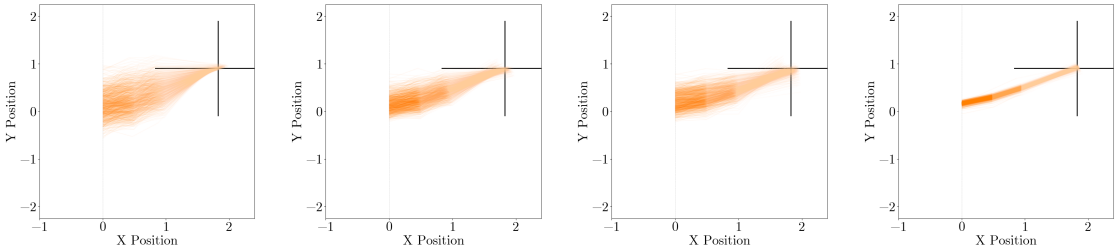
$$
{}^{p}L_x = \frac{1}{n} \sum_{i=1}^{n} (T^{-1}(Y_i, Z_i) - \tilde{X}_i)^2 \,.
$$

Here, $\{\tilde{X}_i\}_{i=1}^{N}$ is the data generated from the input. A uniform prior (Andrle et al., 2021) also represents an inconsistent with the true distribution. If the input variable ranges from

a to b ($b > a$), then the loss term is as follows:

$$ {}^{p}L_x = \frac{1}{n}\sum_{i=1}^{n}\left(\max(0, T^{-1}(Y_i, Z_i) - b) + \max(0, a - T^{-1}(Y_i, Z_i))\right). $$

Here, $a = 0$ and $b = 1$.



(a) No prior (b) Correct Prior, $\lambda' = 1$ (c) Misspecified Prior, $\lambda' = 1$ (d) Correct Prior, $\lambda' = 100$

Figure 2: The four plots illustrate the results of the inverse kinematics case study with different prior regularizations. (a) shows samples from a model trained without any prior loss (i.e., with prior weight $\lambda' = 0$). (b) and (c) display samples from models trained with a prior weight of $\lambda' = 1$. In (b), the model assumes a Gaussian prior over the joint configuration, resulting in a scaled L_2 loss between the ground truth and the generated samples. In contrast, (c) uses a uniform prior $\mathcal{U}(0, 1)$, which is inconsistent with the true distribution. (d) illustrates results from a model trained with a Gaussian prior, using $\lambda' = 100$. The cross indicates the true end-effector position. All models were trained using the NLL loss formulation for the INN.

To demonstrate the effect of incorporating a prior objective during training, the prior weight λ' was set to values of 0, 1, and 100. Figure 2 illustrates the effect of incorporating the prior loss ${}^{p}L_x$ into the training objective. We observe that using a suitable prior, along with a well-chosen prior weight λ', leads to more plausible samples. In the context of the inverse kinematics case study, this corresponds to configurations that keep joint angles close to zero, resulting in "straighter" arm trajectories, as shown in Figure 2b. Since the prior had a mean of 0 and a small variance, it led to samples that favored configurations where the arm positions were straight, with all joint angles close to 0 degrees. However, when the assumed prior is incorrect, incorporating it into the training objective leads to inaccurate joint configurations, as seen in Figure 2c. Finally, Figure 2d demonstrates that when the prior weight λ' is set too high, even with a suitable prior, the prior loss dominates the training, resulting in degraded performance and unrealistic outputs. Increasing the coefficient for ${}^{p}L_x$ highlights this effect further.

We conclude that using prior distribution knowledge in training INNs acts as a form of regularization by constraining the estimated posterior to align with known distributions. By incorporating such priors, the model is guided toward more plausible solutions, preventing it from learning overly complex or irrelevant patterns. Essentially, the prior acts as a form of bias that encourages the model to explore a more reasonable subset of possible solutions,

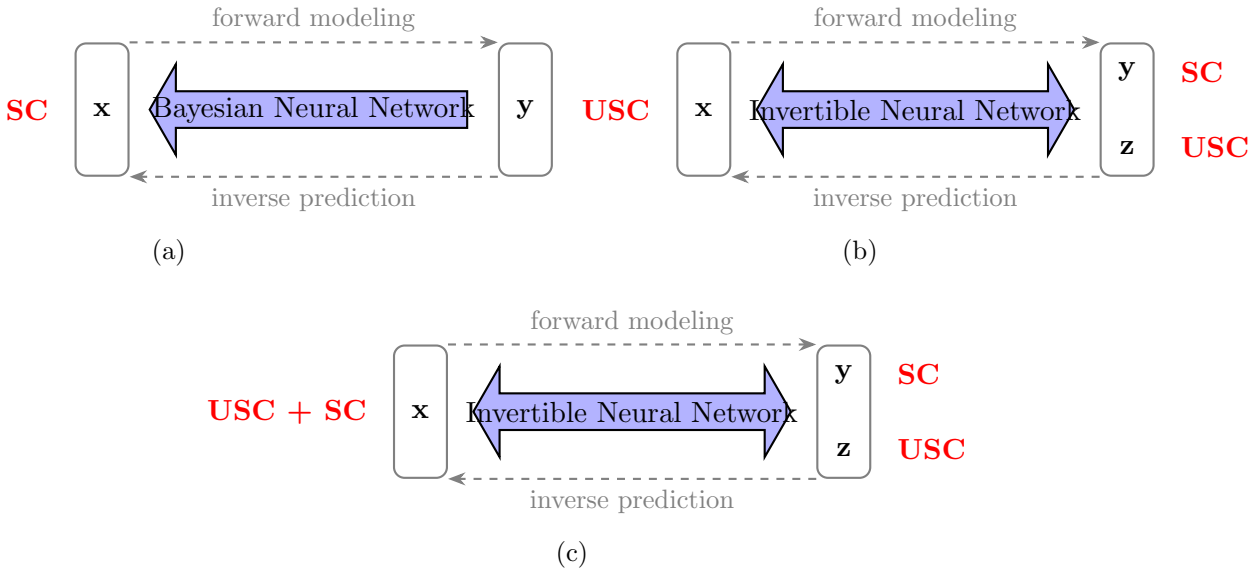


Figure 3: Schematic illustration of how different training paradigms are used to solve inverse problems in the Gaussian setting. (a) Bayesian neural network for solving inverse problem. There is supervised cost (SC) between predicted and true $\mathbf{x}$. (b) INN Trained without ${}^{p}L_x$. There is unsupervised cost (USC) between predicted and true $\mathbf{x}$. (c) Shows an INN model trained with ${}^{p}L_x$, which serves as a reconstruction loss.

leading to robustness in generation tasks. The designers need to navigate the fine line between Bayesian wisdom and risky assumptions. A well-chosen prior can improve the INN performance by effectively regularizing the model, and poorly chosen priors can dominate the results, leading to biased or misleading conclusions.

We conclude this section by discussing the reason behind the importance of prior cost on training by focusing on the Gaussian case. Fig. 3 compares the training procedure of Bayesian neural network (BNN), INN, and INN with prior for solving an inverse problem. The standard direct formulation in BNN depends on a discriminative supervised loss term on X, which is one motivation for introducing INN for inverse problems (Ardizzone et al., 2018). In the case of Gaussian distributions, we have demonstrated that ${}^{p}L_x$ can serve as an effective regularizer by imposing a supervised cost on X. In this case, as demonstrated in Fig. 3c, training with prior propose a hybrid approach applying both supervised and unsupervised costs on X. This approach leverages the regularization with prior to enhance the performance of INN.

4.1.2 Comparison of architectures

This observation compares the performance of INNs on an inverse problem using coupling-based and iResNet architectures, trained with different variational f-divergences as loss functions, considering both forward and backward losses. In general, it is challenging to optimize over all measurable function in the the critic class $\mathcal{G}$, defined in Section 3.1. Hence, in practice, finite-parameter family of functions are used to approximate the optimal critic

function. We adopt the method proposed in Nowozin et al. (2016), to approximate the critic function by a neural network. In this case, we replace the function g in equation (3) with $V_\omega(x) = g_f(V'_\omega(x))$, where $V'_\omega : \mathcal{X} \to \mathbb{R}$ denotes a neural network with parameters ω, and g_f is a monotonic activation function (mapping into the domain of f^*). Here, V'_ω is implemented as a fully connected network with two hidden layers of size 64 with ReLU activations (Goodfellow et al., 2016). The function g_f is determined by the specific choice of the f-divergence (Nowozin et al., 2016). For the Kullback–Leibler (KL) divergence, g_f employs an identity activation, i.e., $g_f(v) = v$. In the case of the Reverse KL divergence, the activation is given by $g_f(v) = -\exp(-v)$, whereas for the Jensen–Shannon (JS) divergence, the activation takes the form $g_f(v) = \log(2) - \log(1 + \exp(-v))$.

In this numerical study, we consider the USL in both the backward and forward directions. In the backward direction, we wish to minimize $D_f(P_{T^{-1}(Y,Z)} \parallel P_X)$ using the following joint objective function

$$\mathcal{F}(\theta, \omega) = \mathbb{E}[V_\omega(T_\theta^{-1}(Y, Z)] - \mathbb{E}[f^*(V_\omega(X))],$$

where T^{-1} and V are parametrized by θ and ω, respectively. In particular, in this case study, we refer to the following regularized empirical estimation as the *backward loss*

$$\mathcal{L}_{\text{backward}}(\theta, \omega) = \frac{1}{N}\sum_{i=1}^{N} V_\omega(T_\theta^{-1}(Y_i, Z_i)) - \frac{1}{N}\sum_{i=1}^{N} f^*(V_\omega(X_i)) + \frac{1}{N}\sum_{i=1}^{N} \|T_\theta^{-1}(Y_i, Z_i) - X_i\|^2 \tag{10}$$

Similarly, in the forward direction, we wish to minimize $D_f(P_{Y,Z} \parallel P_{T(X)})$, using the the objective function

$$\mathcal{F}(\theta, \omega) = \mathbb{E}[V_\omega(Y, Z)] - \mathbb{E}[f^*(V_\omega(T_\theta(X)))]. \tag{11}$$

Here, the following empirical estimation is called the *forward loss*

$$\mathcal{L}_{\text{forward}}(\theta, \omega) = \frac{1}{N}\sum_{i=1}^{N} V_\omega(Y_i, Z_i) - \frac{1}{N}\sum_{i=1}^{N} f^*(V_\omega(T_\theta(X_i))) + \frac{1}{N}\sum_{i=1}^{N} \|T_\theta(X_i) - (Y_i, Z_i)\|^2 .$$

The training scheme, which incorporates f-divergences, differs from the configuration described in the previous section. Unlike models trained with the other cost functions, this setup includes both a discriminator network and a generator network (cf. Nowozin et al. (2016)). Both backward and forward losses are obtained by finding the saddle point of $\mathcal{F}(\theta, \omega)$:

$$\theta^*, \omega^* = \arg\min_\theta \max_\omega \mathcal{F}(\theta, \omega).$$

Each training epoch includes gradient updates for both the discriminator and the generator, and both networks were trained using the Adam optimizer with a learning rate of 2×10^{-4}.

To evaluate the impact of different f-divergences we consider the inverse kinematics problem, describe in Section 4.1.1, aiming to recover joint parameters from the end-effector position. We train INN models using each of the previously discussed f-divergence objectives to investigate their influence on the quality of generated samples. Training performance depends on multiple factors, including the network architecture, and, in this case study, we consider both coupling-based and residual invertible architectures. Both coupling-based

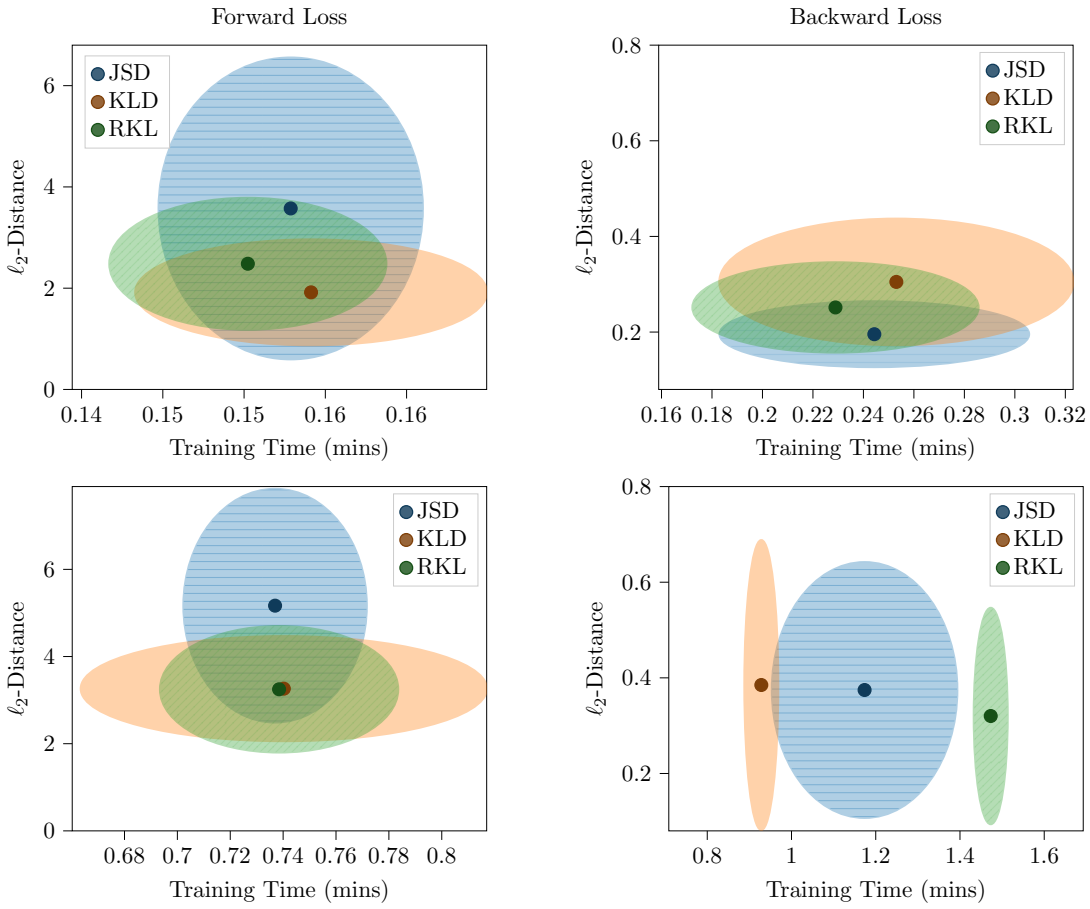


Figure 4: Comparison of f-divergence performance for coupling-based and iResNet architectures in the inverse kinematics case study. The x-axis shows training time and the y-axis shows ℓ_2-distance between true and generated end-effector positions. Ellipses are centered at the mean values, with horizontal and vertical diameters indicating the standard deviations in training time and ℓ_2-distance. The right and left columns correspond to forward loss (11) and backward losses (10), respectively. The top row corresponds to the coupling-based architecture whereas the figures on the bottom row generated based on iResNet layers.

and iResNet architecture comprise four layers, each with affine transformations parameterized by feedforward neural networks with a hidden size of 128. The initial weights of the network were initialized by sampling values from a standard normal distribution. The latent dimension was chosen to be 14 for both configurations.

Figure 4 compares the mean square error and the computation times of the f-divergence losses. We take the mean error over 20 runs of the f-divergences under the given task. As depicted in this figure, training performance is influenced by many factors, including the network architecture, forward and backward training, as well as the capacity of the critic class. Also, different variational formulations of f-divergences may be better suited to

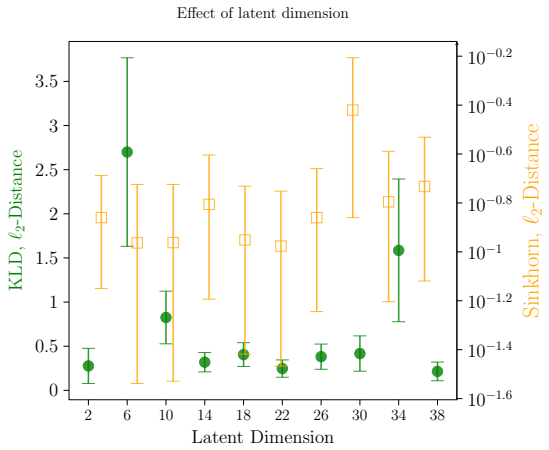


Figure 5: This diagram illustrates the impact of varying latent dimensions on the posterior samples. The circular dots represent the variational formulation of the forward KL divergence and the hollow squares correspond to the Sinkhorn divergence. The x-axis shows latent dimension and the y-axis shows ℓ_2-distance between true and generated end-effector positions.

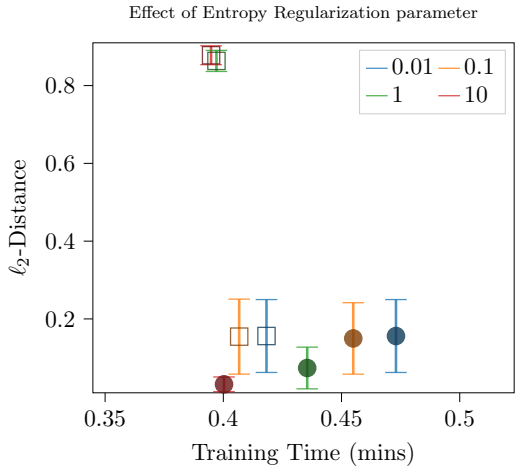


Figure 6: This diagram illustrates the effect of entropy regularization on the Wasserstein distance. The circular dots represent the debiased Sinkhorn divergence, while the hollow squares indicate the Sinkhorn approximation of the W_2 distance. The x-axis shows training time and the y-axis shows the ℓ_2-distance between true and generated end-effector positions.

different settings. As shown in Figure 4, the backward loss in equation (10) yields better performance, and coupling-based designs are more efficient.

4.1.3 Effect of latent dimension in invertible neural networks

To investigate the impact of latent dimension size on training and inference, we train models with different latent dimensions (which can be tuned by appropriately changing $d_{\mathbf{x}}$ by padding). The loss functions employed are the Sinkhorn divergence and the variational formulation of the forward KL divergence. Also, similar to previous examples, the IK problem described in Section 4.1.1 is used for this case study. As illustrated in Figure 5, we observe an approximately local U-shaped relationship between latent dimension and inference performance, where the overall trend is multimodal. Additionally, training with the Sinkhorn divergence exhibits greater stability compared to the forward KL divergence. While in this case study we focus on the dimension of latent variable, the work in Hagemann and Neumayer (2021) investigated the number of modes of P_Z, demonstrating that parameterizing P_Z in dependence on the labels Y is helpful for obtaining stable INNs with reasonable Lipschitz constants.

4.1.4 Effect of entropic regularization on Sinkhorn divergence and Wasserstein distance approximation

To demonstrate the effect of entropic regularization on the debiased Sinkhorn divergence and the Sinkhorn approximation of the Wasserstein distance, we vary the entropic regularization parameter across experiments. The model is trained for the inverse kinematics problem same training configuration described in Section 4.1.1. As illustrated in Figure 6, a decrease in entropic regularization tends to improves sample quality but increases training time. Additionally, training with the Sinkhorn divergence exhibits greater stability compared to the Sinkhorn approximation of the Wasserstein distance.

4.1.5 Effect of support set of the latent distribution and the data distribution

We investigate the effect of the support set of the latent distribution when training with the variational formulation of the KL divergence and the Sinkhorn divergence. In this experiment, the data distribution is fixed as $\mathcal{U}(0,1)$. The latent distributions considered for comparison are $\mathcal{U}(0,1)$, $\mathcal{U}(3,4)$, $\mathcal{U}(5,6)$, $\mathcal{U}(10,11)$ and $\mathcal{U}(15,16)$. As illustrated in Figure 7 and Table 2, the quality of the generated samples deteriorates as the support of the latent distribution becomes increasingly disjoint from that of the true distribution. This effect is observed for both cost functions; however, the model trained using the f-divergence formulation of the KL divergence exhibits a more pronounced degradation compared to the model trained with the Sinkhorn divergence. Thus, our empirical results support the hypothesis that models trained with the Sinkhorn divergence are more robust to mismatches between the supports of the latent and prior distributions.

Divergence / P_Z	KL Divergence	Sinkhorn Divergence
$\mathcal{U}(0,1)$	0.003	0.001
$\mathcal{U}(3,4)$	1.22	0.24
$\mathcal{U}(5,6)$	1.36	0.91
$\mathcal{U}(10,11)$	3.07	1.28
$\mathcal{U}(15,16)$	1.43	1.33

Table 2: This table shows the MMD between samples from the true distribution $\mathcal{U}(0,1)$ and those generated by the trained model.

4.1.6 Effect of moments of X on the performance of NF

In this experiment, we investigate how the number of finite moments of the P_X affects the performance of the NF. To ensure that the moment of X can be adjusted, we assumed X has a Pareto distribution (Arnold, 2014). That is, the dataset was generated by sampling from a Pareto distribution characterized by a shape parameter α and a scale parameter x_m.

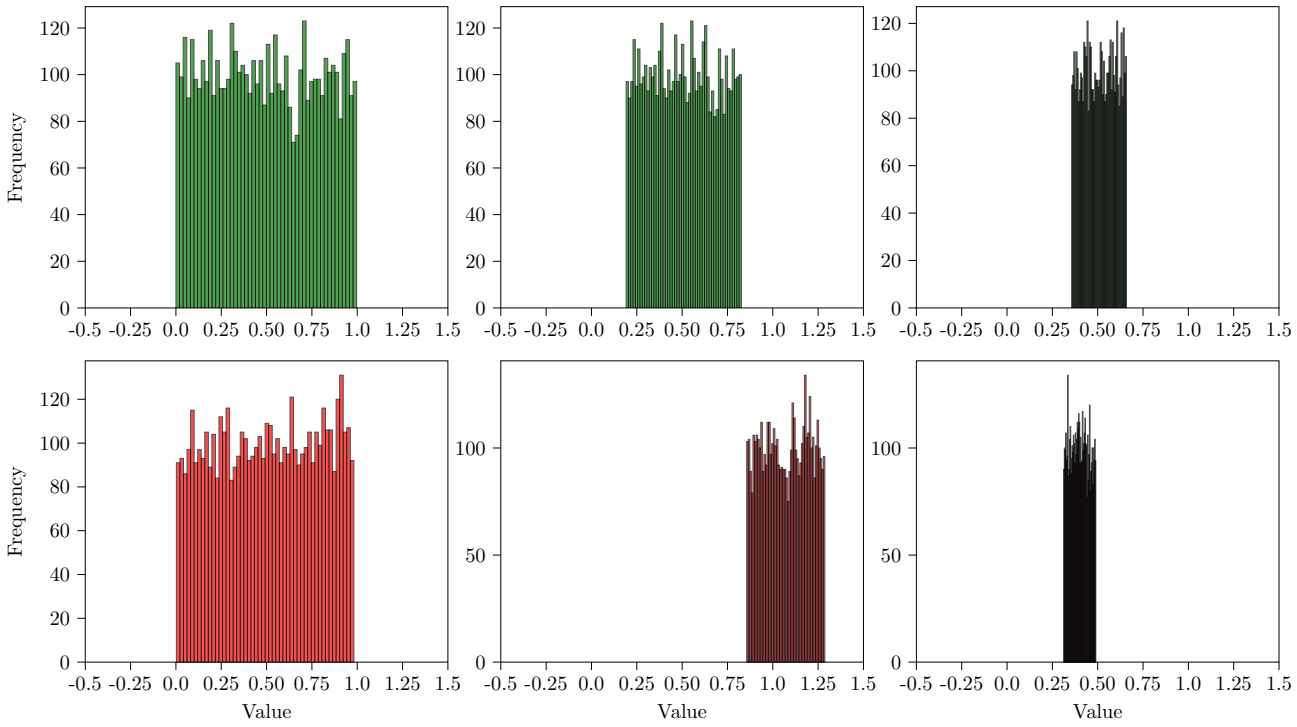


Figure 7: **Top Row:** Samples from NF models trained using the Sinkhorn divergence. The left image corresponds to a latent distribution $\mathcal{U}(0,1)$ used to simulate the data, the middle image uses $\mathcal{U}(3,4)$, and the right image uses $\mathcal{U}(5,6)$. **Bottom Row:** Samples from NF models trained using an f-divergence approximation of the forward KL divergence. The latent distributions mirror those used in the Sinkhorn divergence experiments.

In this case the probability density function of X is given by

$$f(X; x_m, \alpha) = \begin{cases} \dfrac{\alpha x_m^\alpha}{X^{\alpha+1}}, & X \geq x_m, \\ 0, & X < x_m, \end{cases}$$

where $\alpha > 0$ and $x_m > 0$. Also, the NF trained with the variational formulation of KL-Divergence. The latent space was defined by a 10-dimensional random vector drawn from a standard normal distribution.

Here, the parameter α is varied from 1 to 10, and, as illustrated in Fig. 8, the Wasserstein-1 distance between the true and generated distributions decreased. This is consistent with our theoretical results in Appendix B that as the number of finite moment of X increases, the W_1 distance between the true and estimated distributions decreases.

4.2 Geoacoustic Inversion

To demonstrate the practical relevance of our approach, we employ the insights gained from the previous exploratory experiments to the GI setting of the SWellEx-96 experiment (Yardim et al., 2010; Meyer and Gemba, 2021) conducted off the coast of San Diego

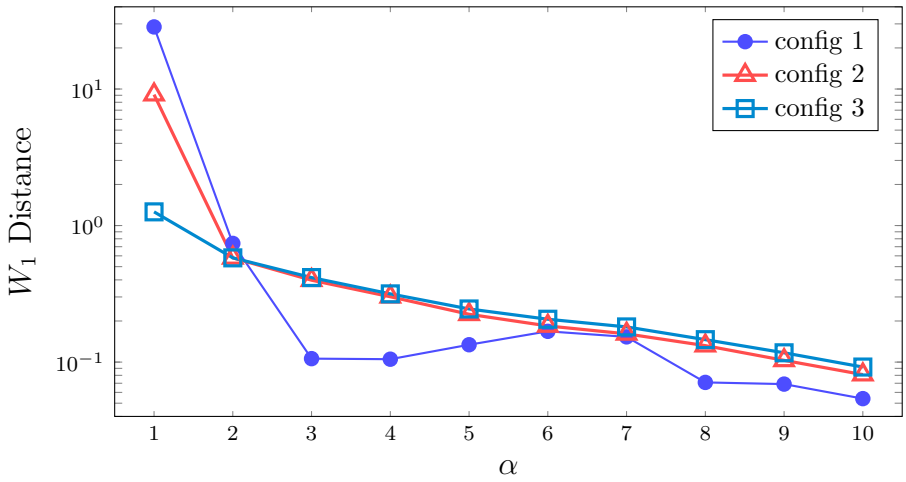


Figure 8: This figure illustrates the Wasserstein-1 distance for increasing values of α in the Pareto distribution, which corresponds to increasing number of finite moments. The y-axis is set to a logarithmic (base 10) scale. Configurations 1–3 specify INNs with 2, 6, 8 coupling layers and subnetworks of hidden sizes 128, 64, 128, respectively.

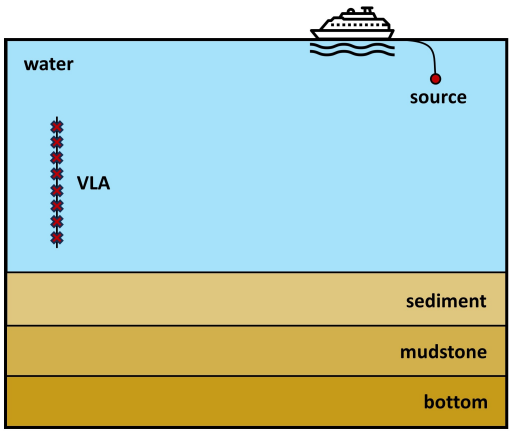


Figure 9: The SWellEx-96 experiment environment. The acoustic source is towed by a research vessel and transmits signals at various frequencies. The acoustic sensor consists of a vertical line array (VLA). Based on the measurements collected at the VLA, the objective is to estimate posterior distributions over parameters of interest (e.g. water depth, sound speed at the water-sediment interface, source range and depth, etc.).

near Point Loma. This experimental setting is one of the most used, documented, and understood studies in the undersea acoustics community.[1] Here, synthetic data corresponding to this experiment is used to simplify the task for the initial application of an INN-based framework in ocean acoustics.

1. see `http://swellex96.ucsd.edu/`

As depicted in Figure 9, the data is collected via a vertical line array (VLA). The specification of the 21 hydrophones of the VLA and sound speed profile (SSP) in the water column is provided in the SWellEx-96 documentation. The SSP and sediment parameters are considered to be range-independent. Water depth refers to the depth of the water at the array. The source is towed by a research vessel which consists of a comb signal comprising frequencies of $49, 79, 112, 148, 201, 283,$ and 388 Hz. While in the SWellEx-96 experiment the position of the source changes with time, for this task we consider the instant when the source depth is 60 m and the distance (or range) between the source and the VLA is 3 km.

The sediment layer is modeled with the following properties. The seabed consists initially of a sediment layer that is 23.5 meters thick, with a density of 1.76 g/cm^3, and an attenuation of 0.2 dB/kmHz. The sound speed at the bottom of this layer is assumed to be 1593 m/s. The second layer is mudstone that is 800 meters thick, possessing a density of 2.06 g/cm^3, and an attenuation of 0.06 dB/kmHz. The top and bottom sound speeds of this layer are 1881 m/s and 3245 m/s respectively. The description of the geoacoustic model of the SWellEx-96 experiment is complemented by a half-space featuring a density of 2.66 g/cm^3, an attenuation of 0.020 dB/kmHz, and a sound speed of 5200 m/s.

Based on the measurements at the VLA, the objective of this task is to infer the posterior distribution over the water depth as well as the sound speed at the water-sediment interface. For this task, we assume all the quantities above to be known. The unknown parameters m_1 (the water depth) and m_2 (the sound speed at the water-sediment interface) follow a uniform prior in $[200.5, 236.5]$ m and $[1532, 1592]$ m/s, i.e. $m_1 \sim \mathcal{U}([200.5, 236.5])$ and $m_2 \sim \mathcal{U}([1532, 1592])$, where $\mathcal{U}(\Omega)$ denotes a uniform distribution in the domain Ω.

The received pressure $\mathbf{y}$ on each hydrophone and for each frequency is a function of unknown parameters $\mathbf{m}$ (e.g. water depth, sound speed at the water-sediment interface, etc.) and additive noise $\boldsymbol{\epsilon}$ as follows

$$\mathbf{y} = s(\mathbf{m}, \boldsymbol{\epsilon}) = F(\mathbf{m}) + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of data noise. Here, $s(\mathbf{m}, \boldsymbol{\epsilon})$ is a known forward model that, assuming an additive noise model, can be rewritten as $F(\mathbf{m}) + \boldsymbol{\epsilon}$, where $F(\mathbf{m})$ represents the undersea acoustic model (Jensen et al., 2011). The SWellEx-96 experiment setup involves a complicated environment and no closed-form analytical solution is available for $F(\mathbf{m})$. In this case, $F(\mathbf{m})$ can only be evaluated numerically, and we use the normal-modes program KRAKEN (Porter, 1992) for this purpose. The signal-to-noise ratio is 15 dB. Inspired by Zhang and Curtis (2021), for the invertible architectures, we include data noise $\boldsymbol{\epsilon}$ as additional model parameters to be learned. In this context, the input of the network is obtained by augmenting the unknown parameters $\mathbf{m}$ with additive noise $\boldsymbol{\epsilon}$.

The pressures received on the hydrophones are considered in the frequency domain, and hence they can be complex numbers. While the invertible architectures can be constructed to address complex numbers, in this case study, we stack the real and imaginary parts of the pressure field at the network's output. That is, the pressure $y = \mathrm{Re}\{y\} + i\,\mathrm{Im}\{y\}$ will be represented as $\big[\mathrm{Re}\{y\}, \mathrm{Im}\{y\}\big]$ at the network's output.

This case study is analyzed noise and signal data collected from five hydrophones. Each hydrophone captured complex measurements across seven frequencies, which we processed by separating into real and imaginary components and then concatenating into a single

vector. This processing approach resulted in input data tensors containing 70 signal measurements. We combined the input parameters with noise measurements, creating tensors of sizes 72.

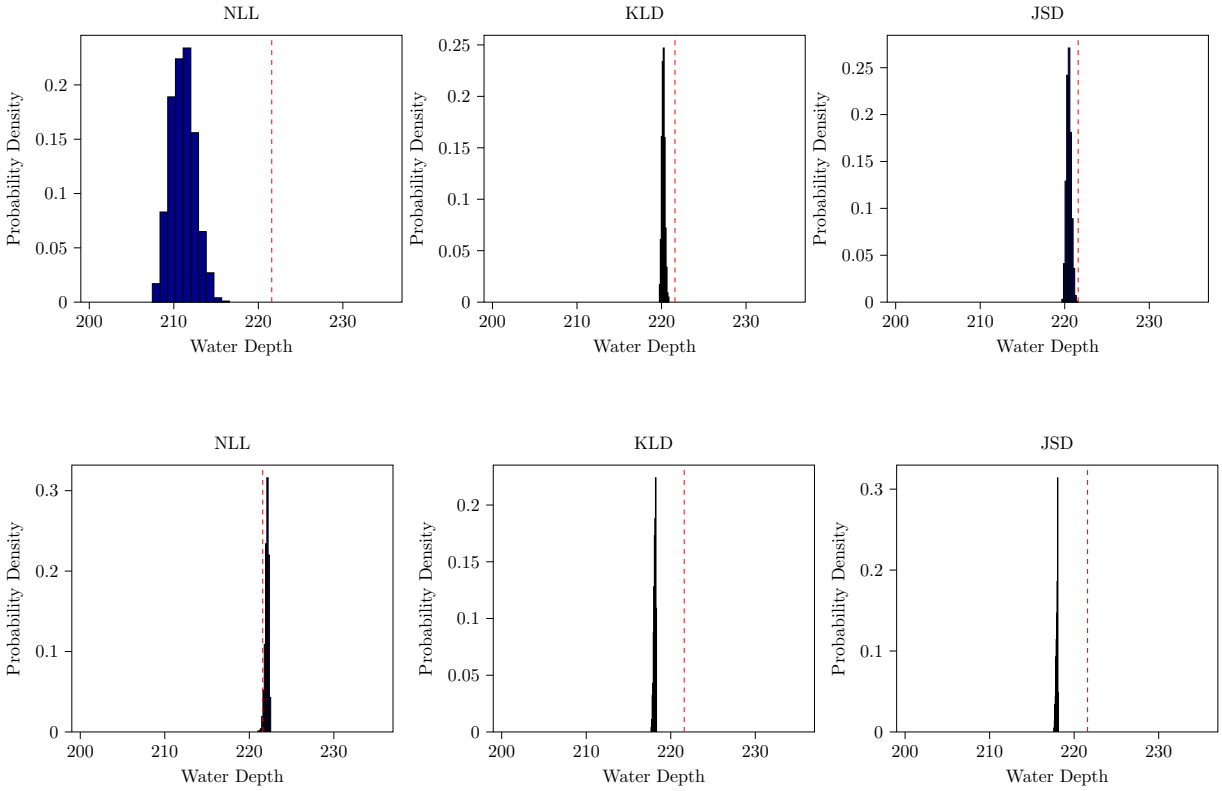


Figure 10: Posterior comparisons between models trained with different loss functions for the GI dataset, using signal measurements from **five** hydrophones. The inferred parameter is **water depth**, with the red dotted line indicating the true value. The INN configuration for the upper row uses Configuration 1 whereas the configuration used in the bottom row is Configuration 2. The y-axis shows un-normalized probability scores.

For the GI dataset, training is conducted using a batch size of 128 over 10 epochs with an 80/20 train–validation split. The optimization settings vary by objective: the INN trained with the NLL loss uses a learning rate of 3×10^{-4}, while the INN and critic network V_ω trained with the f-divergence objective use a learning rate of 2×10^{-4} together with L_2 regularization of 2×10^{-5}. All models employ the Adam optimizer with betas (0.8, 0.9). Two different coupling-based configuration are considered.

Configuration 1 *corresponds to a model with sub-networks of hidden size 64, a latent dimension of 18, and 3 coupling layers.*

Configuration 2 *corresponds to a model with sub-networks of hidden size 403, a latent dimension of 2, and 4 coupling layers.*

The results are presented in Figs 10 and 11, and in Tables 3, 4 and 5. The histograms, in Figs 10 and 11, are based on 1,000 samples drawn from the trained INN. Also, while training times are compared in Table 5 the inference time is also provided in its caption.

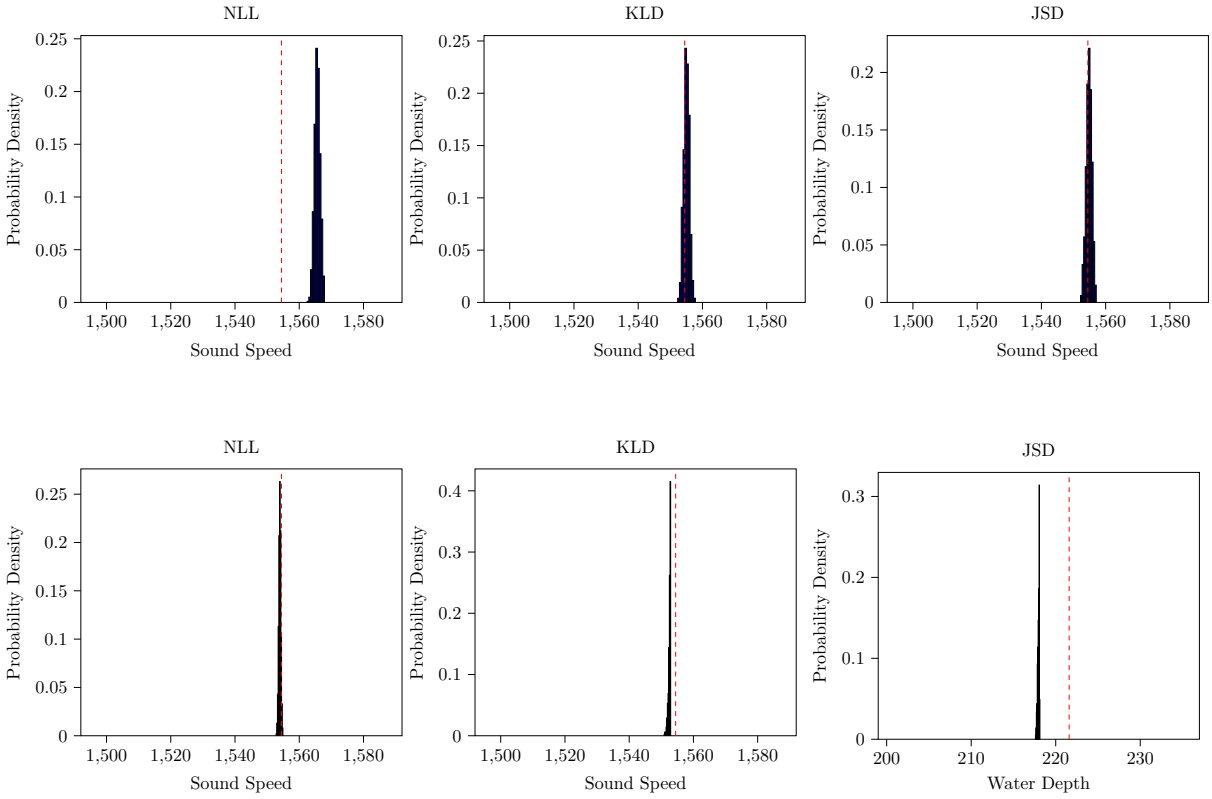


Figure 11: Posterior comparisons between models trained with different loss functions for the GI dataset, using signal measurements from **five** hydrophones. The inferred parameter is **sound speed**, with the red dotted line indicating the true value. The INN configuration for the upper row uses Configuration 1 whereas the configuration used in the bottom row is Configuration 2.

Method \ Epochs	1	5	10
NLL	6.19	22.90	22.48
KL Divergence	7.06	3.46	3.27
JS Divergence	6.63	3.21	3.26

(a) Configuration 1

Method \ Epochs	1	5	10
NLL	25.16	2.32	0.75
KL Divergence	10.56	10.66	11.56
JS Divergence	11.20	11.04	12.26

(b) Configuration 2

Table 3: Comparison of the L_2 loss for water depth using two different model configurations. The parameter being recovered is water depth.

Method \ Epochs	1	5	10
NLL	365.26	211.71	193.99
KL Divergence	31.36	0.92	2.24
JS Divergence	27.40	0.61	7.26

(a) Configuration 1

Method \ Epochs	1	5	10
NLL	9.33	0.73	1.91
KL Divergence	9.26	1.25	6.31
JS Divergence	8.73	2.29	10.24

(b) Configuration 2

Table 4: Comparison of the L_2 loss for sound speed using two different model configurations introduced. The parameter being recovered is sound speed.

Method \ Epochs	1	5	10
NLL	4.87	16.78	30.50
KL Divergence	1.37	5.44	10.91
JS Divergence	1.34	5.58	11.08

(a) Configuration 1

Method \ Epochs	1	5	10
NLL	4.22	14.19	26.02
KL Divergence	1.71	6.76	13.36
JS Divergence	1.86	6.64	12.45

(b) Configuration 2

Table 5: Comparison of **training times** for two model configurations. The configuration used for the left table achieves an **inference time** of 0.0060 seconds per 1000 samples, while the configuration used for the right table achieves 0.0036 seconds for the same number of samples.

For GI, standard likelihood-evaluation-based MCMC is computationally intensive, as each likelihood evaluation necessitates solving the forward model with an acoustic propagation simulator KRAKEN (Porter, 1992). Alternately, a pre-trained INN facilitates rapid, near-real-time posterior inference by eliminating the need for repeated forward model evaluations. This speedup (inference time of 0.0060 seconds per 1000 samples) comes from shifting computation offline into training. Posterior sampling is often computationally challenging, especially in high dimensions (Montanari and Wu, 2023), and there is no universally best configuration for these problems. As demonstrated in our case studies, INNs can be sensitive to architectural and optimization choices; therefore, automated hyperparameter search methods such as the tree-structured Parzen estimator (Bergstra et al., 2011) can be useful in practice.

5 Conclusion

In this work, we have presented a unified theoretical and practical framework that connects INNs and NFs through a common variational formulation, motivated by related works of Nowozin et al. (2016); Zhang et al. (2019); Grover et al. (2018). By analyzing these architectures under a shared perspective, we derived new theoretical guarantees on the approximation quality of both posterior and generative distributions under weaker, more realistic assumptions than those used in prior studies. Our findings bridge a key gap in the

literature by providing a principled understanding of when and why these models succeed in both generative and inverse problem settings. Beyond the theoretical results, our empirical investigations yield general design principles that can guide practitioners in the effective implementation of INN- and NF-based systems. Finally, the application of our framework to a ocean-acoustic inversion task demonstrates its practical value and robustness.

Acknowledgments and Disclosure of Funding

M. J. Khojasteh acknowledges support from the Gleason Endowment at RIT and also thanks Prof. Florian Meyer and Dr. Mohsen Sadr for their constructive discussions and suggestions. K. Youcef-Toumi acknowledges the support from the Center for Complex Engineering Systems (CCES) at King Abdulaziz City for Science and Technology (KACST) and Massachusetts Institute of Technology (MIT).

References

Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50:5–43, 2003.

Anna Andrle, Nando Farchmin, Paul Hagemann, Sebastian Heidenreich, Victor Soltwisch, and Gabriele Steidl. Invertible Neural Networks versus MCMC for Posterior Reconstruction in Grazing Incidence X-Ray Fluorescence . In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 528–539. Springer, 2021.

Lynton Ardizzone, Jakob Kruse, Sebastian Wirkert, Daniel Rahner, Eric W Pellegrini, Ralf S Klessen, Lena Maier-Hein, Carsten Rother, and Ullrich Köthe. Analyzing Inverse Problems with Invertible Neural Networks. *arXiv preprint arXiv:1808.04730*, 2018.

Lynton Ardizzone, Carsten Lüth, Jakob Kruse, Carsten Rother, and Ullrich Köthe. Guided Image Generation with conditional Invertible Neural Networks. *arXiv preprint arXiv:1907.02392*, 2019.

Martin Arjovsky and Leon Bottou. Towards Principled Methods for Training Generative Adversarial Networks. In *International Conference on Learning Representations*, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Barry C Arnold. Pareto distribution. *Wiley StatsRef: Statistics Reference Online*, pages 1–10, 2014.

Yves F Atchadé and Jeffrey S Rosenthal. On Adaptive Markov Chain Monte Carlo algorithms. *Bernoulli*, 11(5):815–828, 2005.

Mark A Beaumont. Approximate bayesian computation. *Annual review of statistics and its application*, 6(1):379–403, 2019.

Jens Behrmann, Will Grathwohl, Ricky TQ Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible Residual Networks. In *International conference on machine learning*, pages 573–582. PMLR, 2019.

Jens Behrmann, Paul Vicol, Kuan-Chieh Wang, Roger Grosse, and Jörn-Henrik Jacobsen. Understanding and mitigating exploding inverses in Invertible Neural Networks. In *International Conference on Artificial Intelligence and Statistics*, pages 1792–1800. PMLR, 2021.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.

Harry Bevins, Will Handley, and Thomas Gessey-Jones. Piecewise Normalizing Flows. *arXiv preprint arXiv:2305.02930*, 2023.

Vladimir Igorevich Bogachev, Aleksandr Viktorovich Kolesnikov, and Kirill Vladimirovich Medvedev. Triangular transformations of measures. *Sbornik: Mathematics*, 196(3):309, 2005.

James Brofos, Marylou Gabrié, Marcus A Brubaker, and Roy R Lederman. Adaptation of the Independent Metropolis-Hastings Sampler with Normalizing Flow Proposals . In *International Conference on Artificial Intelligence and Statistics*, pages 5949–5986. PMLR, 2022.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.

N Ross Chapman and Er Chang Shang. Review of geoacoustic inversion in underwater acoustics. *Journal of Theoretical and Computational Acoustics*, 29(03):2130004, 2021.

Ricky TQ Chen, Jens Behrmann, David K Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for Invertible Generative Modeling. *Advances in neural information processing systems*, 32, 2019.

Florentin Coeurdoux, Nicolas Dobigeon, and Pierre Chainais. Sliced-Wasserstein normalizing flows: beyond maximum likelihood training. *arXiv preprint arXiv:2207.05468*, 2022.

Imre Csiszár. On Information-Type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. *Advances in neural information processing systems*, 26, 2013.

Masoumeh Dashti, Stephen Harris, and Andrew Stuart. Besov priors for Bayesian inverse problems. *Inverse Problems and Imaging*, 6(2):183–200, 2012.

Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear Independent Components Estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. *arXiv preprint arXiv:1605.08803*, 2016.

Stan E Dosso and Jan Dettmer. Bayesian matched-field geoacoustic inversion. *Inverse Problems*, 27(5):055009, 2011.

Arnaud Doucet and Xiaodong Wang. Monte Carlo methods for signal processing: a review in the statistical signal processing context. *IEEE Signal Processing Magazine*, 22(6): 152–170, 2005.

Richard M Dudley. *Real Analysis and Probability*. Chapman and Hall/CRC, 2018.

Samuel G Fadel, Sebastian Mair, Ricardo da S. Torres, and Ulf Brefeld. Principled Interpolation in Normalizing Flows. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II 21*, pages 116–131. Springer, 2021.

Yuwei Fan and Lexing Ying. Solving Inverse Wave Scattering with Deep Learning . *arXiv preprint arXiv:1911.13202*, 2019.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3):707–738, 2015.

Marylou Gabrié, Grant M Rotskoff, and Eric Vanden-Eijnden. Efficient bayesian sampling using normalizing flows to assist markov chain monte carlo methods. *arXiv preprint arXiv:2107.08001*, 2021.

Marylou Gabrié, Grant M Rotskoff, and Eric Vanden-Eijnden. Adaptive Monte Carlo augmented with normalizing flows. *Proceedings of the National Academy of Sciences*, 119 (10):e2109420119, 2022.

John Geweke. Bayesian inference in econometric models using monte carlo integration. *Econometrica: Journal of the Econometric Society*, pages 1317–1339, 1989.

Aidan N Gomez, Mengye Ren, Raquel Urtasun, and Roger B Grosse. The Reversible Residual Network: Backpropagation without Storing Activations. *Advances in neural information processing systems*, 30, 2017.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks . *Advances in neural information processing systems*, 27, 2014.

Aditya Grover, Manik Dhar, and Stefano Ermon. Flow-GAN: Combining Maximum Likelihood and Adversarial Learning in Generative Models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Xiaofei Guan, Xintong Wang, Hao Wu, Zihao Yang, and Peng Yu. Efficient Bayesian Inference using Physics-Informed Invertible Neural Networks for Inverse Problems. *Machine Learning: Science and Technology*, 5(3):035026, 2024.

Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and R Venkatesh Babu. DeLiGAN : Generative Adversarial Networks for Diverse and Limited Data . In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 166–174, 2017.

Paul Hagemann and Sebastian Neumayer. Stabilizing Invertible Neural Networks using Mixture Models. *Inverse Problems*, 37(8):085002, 2021.

Tennessee Hickling and Dennis Prangle. Flexible Tails for Normalizing Flows. *arXiv preprint arXiv:2406.16971*, 2024.

Chen-Fen Huang, Peter Gerstoft, and William S Hodgkiss. Uncertainty analysis in matched-field geoacoustic inversions. *The Journal of the Acoustical Society of America*, 119(1): 197–207, 2006.

Jörn-Henrik Jacobsen, Arnold Smeulders, and Edouard Oyallon. i-RevNet: Deep Invertible Networks. *arXiv preprint arXiv:1802.07088*, 2018.

Finn B Jensen, William A Kuperman, Michael B Porter, Henrik Schmidt, and Alexandra Tolstoy. *Computational ocean acoustics*, volume 2011. Springer, 2011.

Bangti Jin, Zehui Zhou, and Jun Zou. On the Approximation of Bi-Lipschitz Maps by Invertible Neural Networks . *Neural Networks*, 174:106214, 2024.

Jakub M Tomczak JM and Max Welling. Improving Variational Auto-Encoders using convex combination linear Inverse Autoregressive Flow . page 162, 2017.

Galin L Jones and Qian Qin. Markov chain Monte Carlo in practice. *Annual Review of Statistics and Its Application*, 9(1):557–578, 2022.

Liyiming Ke, Sanjiban Choudhury, Matt Barnes, Wen Sun, Gilwoo Lee, and Siddhartha Srinivasa. Imitation Learning as f-divergence minimization. In *Algorithmic Foundations of Robotics XIV: Proceedings of the Fourteenth Workshop on the Algorithmic Foundations of Robotics 14*, pages 313–329. Springer, 2021.

Thomas A Keller, Jorn WT Peters, Priyank Jaini, Emiel Hoogeboom, Patrick Forré, and Max Welling. Self Normalizing Flows. In *International Conference on Machine Learning*, pages 5378–5387. PMLR, 2021.

Yuehaw Khoo and Lexing Ying. SwitchNet: a neural network model for forward and inverse scattering problems. *SIAM Journal on Scientific Computing*, 41(5):A3182–A3201, 2019.

Diederik P Kingma, Max Welling, et al. An Introduction to Variational Autoencoders . *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

Durk P Kingma and Prafulla Dhariwal. GLOW: Generative Flow with Invertible 1x1 Convolutions. *Advances in neural information processing systems*, 31, 2018.

Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing Flows: An Introduction and Review of Current Methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020.

Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *International conference on machine learning*, pages 181–189. PMLR, 2014.

Jakob Kruse, Lynton Ardizzone, Carsten Rother, and Ullrich Köthe. Benchmarking Invertible Architectures on Inverse Problems . *arXiv preprint arXiv:2101.10763*, 2021.

Liam Anthony Kruse, Alexandros Tzikas, Harrison Delecki, Mansur Arief, and Mykel J Kochenderfer. Enhanced Importance Sampling Through Latent Space Exploration in Normalizing Flows. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17983–17989, 2025.

Vyacheslav Kungurtsev, Adam Cobb, Tara Javidi, and Brian Jalaian. Decentralized Bayesian Learning with Metropolis-Adjusted Hamiltonian Monte Carlo. *Machine Learning*, 112(8):2791–2819, 2023.

Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. Copula-Based Normalizing Flows. *arXiv preprint arXiv:2107.07352*, 2021.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. MMD GAN: Towards Deeper Understanding of Moment Matching Network . *Advances in neural information processing systems*, 30, 2017.

Alexander Luce, Ali Mahdavi, Heribert Wankerl, and Florian Marquardt. Investigation of Inverse Design of Multilayer thin-films with conditional Invertible Neural Networks. *Machine Learning: Science and Technology*, 4(1):015014, feb 2023. doi: 10.1088/2632-2153/acb48d. URL `https://dx.doi.org/10.1088/2632-2153/acb48d`.

David JC MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge university press, 2003.

Youssef Marzouk, Tarek Moselhy, Matthew Parno, and Alessio Spantini. An introduction to sampling via measure transport. *arXiv preprint arXiv:1602.05023*, 2016.

Andreas Maurer. A vector-contraction inequality for Rademacher complexities . In *International Conference on Algorithmic Learning Theory*, pages 3–17. Springer, 2016.

Florian Meyer and Kay L. Gemba. Probabilistic focalization for shallow water localization. *J. Acoust. Soc. Am.*, 150(2):1057–1066, 08 2021.

Andrea Montanari and Yuchen Wu. Posterior sampling in high dimension via diffusion processes. *arXiv preprint arXiv:2304.11449*, 2023.

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

Saeed B Niku. *Introduction to Robotics: Analysis, Control, Applications* . John Wiley & Sons, 2020.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization . *Advances in neural information processing systems*, 29, 2016.

Govinda Anantha Padmanabha and Nicholas Zabaras. Solving Inverse Problems using conditional Invertible Neural Networks. *Journal of Computational Physics*, 433:110194, 2021.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked Autoregressive Flow for Density Estimation. *Advances in neural information processing systems*, 30, 2017.

George Papamakarios, David Sterratt, and Iain Murray. Sequential Neural Likelihood: Fast Likelihood-free Inference with Autoregressive Flows. In *The 22nd international conference on artificial intelligence and statistics*, pages 837–848. PMLR, 2019.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing Flows for Probabilistic Modeling and Inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Matthew Parno, Tarek Moselhy, and Youssef Marzouk. A multiscale strategy for Bayesian Inference using transport maps. *SIAM/ASA Journal on Uncertainty Quantification*, 4 (1):1160–1190, 2016.

Gabriel Peyré, Marco Cuturi, et al. Computational Optimal Transport: With applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge university press, 2025.

Michael B Porter. The KRAKEN normal mode program . *Naval Research Laboratory, Washington DC*, 1992.

Stefan T Radev, Frederik Graw, Simiao Chen, Nico T Mutters, Vanessa M Eichel, Till Bärnighausen, and Ullrich Köthe. OutbreakFlow: Model-based Bayesian Inference of disease outbreak dynamics with Invertible Neural Networks and its application to the COVID-19 pandemics in Germany. *PLoS computational biology*, 17(10):e1009472, 2021.

Stefan T Radev, Marvin Schmitt, Lukas Schumacher, Lasse Elsemüller, Valentin Pratz, Yannik Schälte, Ullrich Köthe, and Paul-Christian Bürkner. BayesFlow: Amortized Bayesian workflows with neural networks. *arXiv preprint arXiv:2306.16015*, 2023.

Aaditya Ramdas, Sashank Jakkam Reddi, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the Decreasing Power of Kernel and Distance based Nonparametric Hypothesis Tests in High Dimensions . In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Simiao Ren, Willie Padilla, and Jordan Malof. Benchmarking deep inverse models over time, and the neural-adjoint method. *Advances in Neural Information Processing Systems*, 33: 38–48, 2020.

Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton university press, 1970.

Vivekananda Roy. Convergence diagnostics for Markov chain Monte Carlo. *Annual Review of Statistics and Its Application*, 7(1):387–412, 2020.

Avraham Ruderman, Mark Reid, Darío García-García, and James Petterson. Tighter Variational Representations of f-Divergences via Restriction to Probability Measures . *arXiv preprint arXiv:1206.4664*, 2012.

Christoph Schönle and Marylou Gabrié. Optimizing Markov chain Monte Carlo convergence with normalizing flows and Gibbs sampling. In *NeurIPS 2023 AI for Science Workshop*, 2023.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms* . Cambridge university press, 2014.

Galen R Shorack and Jon A Wellner. *Empirical processes with applications to statistics*. SIAM, 2009.

Jiaming Song, Shengjia Zhao, and Stefano Ermon. A-NICE-MC: Adversarial Training for MCMC. *Advances in neural information processing systems*, 30, 2017.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert Space Embeddings and Metrics on Probability Measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

Vincent Stimper, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Resampling Base Distributions of Normalizing Flows . In *International Conference on Artificial Intelligence and Statistics*, pages 4915–4936. PMLR, 2022.

Andrew M Stuart. Inverse Problems: a Bayesian Perspective. *Acta numerica*, 19:451–559, 2010.

Esteban G Tabak and Cristina V Turner. A Family of Nonparametric Density Estimation Algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

Takeshi Teshima, Isao Ishikawa, Koichi Tojo, Kenta Oono, Masahiro Ikeda, and Masashi Sugiyama. Coupling-based Invertible Neural Networks Are Universal Diffeomorphism Approximators. *Advances in Neural Information Processing Systems*, 33:3362–3373, 2020.

Juan S Toquica, Patrícia S Oliveira, Witenberg SR Souza, José Maurício ST Motta, and Díbio L Borges. An Analytical and a Deep Learning model for solving the Inverse Kinematic problem of an Industrial parallel robot. *Computers & Industrial Engineering*, 151: 106682, 2021.

Alexandre Verine, Benjamin Negrevergne, Muni Sreenivas Pydi, and Yann Chevaleyre. Precision-Recall Divergence Optimization for Generative Modeling with GANs and Normalizing Flows . *Advances in Neural Information Processing Systems*, 36:32539–32573, 2023.

Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2008.

Cédric Villani. *Topics in Optimal Transportation*, volume 58. American Mathematical Soc., 2021.

Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. Learning Likelihoods with Conditional Normalizing Flows. *arXiv preprint arXiv:1912.00042*, 2019.

Steven Winter, Trevor Campbell, Lizhen Lin, Sanvesh Srivastava, and David B Dunson. Machine Learning and the Future of Bayesian Computation . *arXiv preprint arXiv:2304.11251*, 2023.

Sihong Wu, Qinghua Huang, and Li Zhao. Fast Bayesian Inversion of Airborne Electromagnetic Data Based on the Invertible Neural Network. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–11, 2023.

Caglar Yardim, Peter Gerstoft, and William S Hodgkiss. Geoacoustic and Source Tracking using Particle Filtering: Experimental results. *The Journal of the Acoustical Society of America*, 128(1):75–87, 2010.

Lexing Ying. Solving Inverse Problems with Deep Learning. In *Proc. Int. Cong. Math*, volume 7, pages 5154–5175, 2022.

Mingtian Zhang, Thomas Bird, Raza Habib, Tianlin Xu, and David Barber. Variational f-divergence Minimization. *arXiv preprint arXiv:1907.11891*, 2019.

Mingtian Zhang, Peter Hayes, Thomas Bird, Raza Habib, and David Barber. Spread Divergence. In *International Conference on Machine Learning*, pages 11106–11116. PMLR, 2020.

Xin Zhang and Andrew Curtis. Bayesian Geophysical Inversion using Invertible Neural Networks. *Journal of Geophysical Research: Solid Earth*, 126(7):e2021JB022320, 2021.

Appendix A. Additional Background

Definition 13 (Rademacher Complexity) *Let $\mathcal{F}$ denote a class of real-valued functions over the domain $\mathbb{R}^d$, and let $X^n \overset{i.i.d.}{\sim} P_X$ denote an i.i.d. dataset. With ϵ^n denoting n i.i.d. Rademacher random variables (i.e., ± 1 w.p. $1/2$ each), we define the Rademacher complexity of the function class $\mathcal{F}$ as*

$$\mathfrak{R}_n(\mathcal{F}, X^n) = \mathbb{E}_{\epsilon^n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i)\middle| X^n\right]$$

We will denote the expected Rademacher complexity (with respect to the data X^n) as $\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_{X^n}[\mathfrak{R}_n(\mathcal{F}, X^n)]$.

We now recall a result about vector contraction for Rademacher complexities from Maurer (2016).

Fact A.1 *For some $d \geq 1$, let $\phi_i : \mathbb{R}^d \to \mathbb{R}$ denote a collection of J-Lipschitz functions in ℓ_2 norm with $\phi_i(\mathbf{0}) = 0$. Then, for a given set of points $\{x_1, \ldots, x_n\}$ in some domain $\mathcal{X}$, and a function class $\mathcal{H}$ consisting of $h : \mathcal{X} \to \mathbb{R}^d$, we have the following:*

$$\mathbb{E}_{\epsilon^n}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \phi_i(h(x_i))\right] \leq 2J\mathbb{E}_{\sigma^n}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \langle \sigma_i, h(x_i)\rangle\right].$$

Above, we assume that $(\epsilon_1, \ldots, \epsilon_n)$ are i.i.d. Rademacher, and each $(\sigma_1, \ldots, \sigma_n)$ represents i.i.d. $\{-1, 1\}^d$-valued random vectors with i.i.d. Rademacher coordinates.

Finally, we now recall a version of a well-known concentration inequality that we will use to obtain high probability uniform deviation bounds.

Fact A.2 (McDiarmid's Inequality) *Let $\{X_i : 1 \leq i \leq n\}$ denote a stream of independent $\mathcal{X}$-valued observations, and let $h : \mathcal{X}^n \to \mathbb{R}$ be a function satisfying a bounded difference property:*

$$|h(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - h(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c,$$

for all $i \in [n]$, $x^n \in \mathcal{X}^n$ and $x_i' \in \mathcal{X}$. Then we have the following for all $\epsilon > 0$:

$$\mathbb{P}\left(h(X^n) - \mathbb{E}[h(X^n)] > n\epsilon\right) \leq \exp\left(-\frac{2n\epsilon^2}{c^2}\right).$$

A.1 Discussion of Invertible Neural Architectures

This section outlines the key factors that ensure practical invertibility in INNs, which are formally defined in Section 1, including architectural structures, bidirectional training, and padding schemes.

Model Architecture: A key aspect in designing INNs is constructing architectures that are both expressive and invertible. One of the most popular design is constructed by concatenating affine coupling blocks Kingma and Dhariwal (2018); Dinh et al. (2016, 2014).

In this case, invertible architecture T is defined by a sequence of reversible blocks, where each block consists of two complementary affine coupling layers. The the block's input $\mathbf{u} \in \mathbb{R}^{d_\mathbf{u}}$ is split into $\mathbf{u}_1 \in \mathbb{R}^{d_{\mathbf{u}_1}}$ and $\mathbf{u}_2 \in \mathbb{R}^{d_\mathbf{u} - d_{\mathbf{u}_1}}$. Each transformation stage applies a learned mapping to a subset of the input features, while the remaining features are left unchanged, ensuring invertibility of the overall transformation. In other words, we have

$$\begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{u}_1 \odot \exp\big(s_1(\mathbf{u}_2)\big) + t_1(\mathbf{u}_2) \\ \mathbf{u}_2 \end{bmatrix}, \qquad \begin{bmatrix} \mathbf{o}_1 \\ \mathbf{o}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \odot \exp\big(s_2(\mathbf{v}_1)\big) + t_2(\mathbf{v}_1) \end{bmatrix}$$

The mappings s_i and t_i for $i = 1, 2$ are arbitrarily neural network. Each block can be inverted, that is, given the output $\mathbf{o} = [\mathbf{o}_1, \mathbf{o}_2]$ the inverse can be calculated as

$$\mathbf{u}_2 = \big(\mathbf{o}_2 - t_2(\mathbf{o}_1)\big) \oslash \exp\big(s_2(\mathbf{o}_1)\big)$$

and

$$\mathbf{u}_1 = \big(\mathbf{o}_1 - t_1(\mathbf{u}_2)\big) \oslash \exp\big(s_1(\mathbf{u}_2)\big)$$

Note that when the coupling block is inverted, the subnetworks s_i and t_i need not to be invertible they are only evaluated in the forward direction. Also, this structure yields a triangular Jacobian, making the log-determinant computationally tractable.

Another popular approach to achieve invertibility is using residual connections (iResNet models) Gomez et al. (2017); Jacobsen et al. (2018); Behrmann et al. (2019). In this way, the mapping T is constructed by the successive composition of residual blocks, $T = (I + F_M) \circ (I + F_{M-1}) \circ \cdots \circ (I + F_1)$, where " $\circ$ " denotes the composition operation. A sufficient condition for invertibility of each block is that the residual sub-network has a Lipschitz constant less than one. However, unlike coupling-based methods that offer exact solutions for the inverse and the log-determinant of the Jacobian, they need to be approximated in the case of iResNet models Chen et al. (2019).

In summary, coupling layers provide exact inversion and efficient Jacobian computation, but their fixed partitioning limits flexibility and can affect stability. Invertible residual networks support more flexible and expressive architectures, needing only a Lipschitz constraint for stability. However, they have less efficient inversion and Jacobian computation than coupling layers.

Bi-Directional Training: Invertible networks allow applying losses in both the input and output domains, since the mapping can be evaluated in both directions. During training, one may alternate between forward and inverse passes, accumulating gradients from both directions. This bidirectional loss formulation can improve the effectiveness and stability of training Ardizzone et al. (2018).In the forward direction, as discussed in (1), the supervised loss (SL) $L_\mathbf{y}(T)$ and USL $L_\mathbf{z}(T)$ will be minimized. Similarly, in the backward direction an USL $L_\mathbf{x}(T)$ can be considered. For the SL we consider mean square error, and the USL is discussed in the next section.

Padding: Padding adjusts input dimensionality to meet a model's structural requirements Ardizzone et al. (2018). For INNs, it resolves mismatches by adding entries enabling bijective mappings. Common strategies like zero-padding and repetition-padding

keep forward and inverse passes dimensionally consistent and can support stable learning. For instance, if zero-padding is used on either side of the network, additional loss terms, called reconstruction loss, are required to ensure no information is encoded in the padding dimensions and forcing these dimensions to remain inactive.

A.2 Negative Log-Likelihood (NLL)

For a map $T(X;\theta) \mapsto [Y, Z]$ parameterized by θ, and assuming Y and Z are independent, the change-of-variables formula implies that the density $P_{T^{-1}(Y,Z)}$ is

$$P_{T^{-1}(Y,Z)} = P_{T_{\mathbf{y}}(X)}\, P_{T_{\mathbf{z}}(X)} \cdot \left|\det\left(\mathbb{J}_{X \mapsto [Y,Z]}(X)\right)\right|$$

where $\mathbb{J}_{X \mapsto [Y,Z]}(X;\theta)$ denotes the Jacobian of the map T parameterized by θ. As described next, this expression can be used to define an unsupervised training loss. In particular, we aim to minimize the forward KL divergence between the true posterior distribution $P_{X|Y}$ and $P_{T^{-1}(Y,Z)}$, given by

$$\begin{aligned}\mathcal{L}(\theta) &= D_{\mathrm{KL}}\left(P_{X|Y}, P_{T^{-1}(Y,Z)}\right) = -\mathbb{E}_{P_{X|Y}}\left[\log P_{T^{-1}(Y,Z)}\right] + \text{const.} \\ &= -\mathbb{E}_{P_{X|Y}}\left[\log P_{T_{\mathbf{y}}(X)} + \log P_{T_{\mathbf{z}}(X)} + \log\left|\det\left(\mathbb{J}_{X \mapsto [Z,Y]}(X)\right)\right|\right] + \text{const.}\end{aligned}$$

The empirical approximation of the above loss is as follows.

$$\hat{\mathcal{L}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left(\log P_{T_{\mathbf{y}}(X_i)} + \log P_{T_{\mathbf{z}}(X_i)} + \log\left|\det\left(\mathbb{J}_{X \mapsto [Y,Z]}(X_i)\right)\right| + \text{const.}\right) \tag{12}$$

As we can see, minimizing the above Monte Carlo approximation of the KL divergence is equivalent to maximizing likelihood (or minimizing negative log-likelihood). Assuming P_Z is standard Gaussian and P_Y is a multivariate normal distribution around $\mathbf{y}_{\mathrm{gt}}$ and a small standard deviation σ, NLL loss in Eq. (12) becomes

$$\mathcal{L}_{\mathrm{NLL}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\,\frac{\left(T_{\mathbf{y}}(X_i) - y_{\mathrm{gt}}\right)^2}{\sigma^2} + \frac{1}{2}\,T_{\mathbf{z}}(X_i)^2 - \log\left|\det\left(\mathbb{J}_{X \mapsto [Y,Z]}(X_i)\right)\right|\right)$$

Note that we started with the backward loss between $P_{X|Y}$ and $P_{T^{-1}(Y,Z)}$ and end up with a supervised loss and unsupervised loss in the forward direction. Also, the Gaussian assumption on Y can be restrictive and as discussed in Section 2, we eliminate the need for such distributional assumptions by relying solely on samples and leveraging the variational formulation of f-divergences, which yields a lower bound on the true value which is discussed more in Section 4.1.2.

Appendix B. Variational normalizing flow (V-NF)

In this section, we introduce a class of normalizing flows trained using a variational objective. To simplify our presentation, we focus on the concrete example of the variational form of relative entropy, although, similar arguments can be developed for a larger class of f-divergences. Our discussion in this section closely follows the structure of Section 3.2

for INNs, and in fact some of the results we present can be directly inferred from the results of Section 3.2. However, we include all the details to keep this section self-contained and independent of Section 3.2.

Let $\mathcal{X} = \mathbb{R}^d$ denote the observation space and let $\mathcal{Z} = \mathbb{R}^d$ denote the latent space in which the latent variable $Z \sim P_Z$ takes its values. Let $\mathcal{T} : \mathcal{X} \to \mathcal{Z}$ denote a class of diffeomorphisms[2], and for any $T \in \mathcal{T}$, define the model distribution on $\mathcal{X}$ as

$$Q_T = (T^{-1})_{\#} P_Z \quad \iff \quad Q_T(E) = P_Z(\{z \in \mathcal{Z} : T^{-1}(z) \in E\}), \text{ for any measurable } E \subset \mathcal{X}.$$

In other words, for $Z \sim P_Z$, the model T represents a random variable $X_T = T^{-1}(Z) \sim Q_T$ with density $q_T(x) = p_Z(T(x))\det \mathbb{J}T(x)$, with $\mathbb{J}T$ denoting the associated Jacobian.

Training procedure. As mentioned in Section 3.1, we consider a variational objective for training NFs, and in particular, in this section we focus on the concrete case of forward relative entropy loss by using a function class $\mathcal{G}_n$ (critic or witness function class that we allow to grow with n). Formally, the population loss that will be used to train a model is

$$L^{\mathrm{NF}}(T) = \sup_{g \in \mathcal{G}_n} \{\mathbb{E}_{X \sim P_X}[\log g(X)] - \mathbb{E}_{X_T \sim Q_T}[g(X_T)]\} + 1.$$

Here we have used the variational definition of relative entropy presented in Nguyen et al. (2010), and by definition, $L^{\mathrm{NF}}(T) \leq D_{KL}(P_X \parallel Q_T)$ with equality if the true likelihood ratio dP_X/dQ_T is contained in $\mathcal{G}_n$. Given n i.i.d. draws $X_1, \ldots, X_n$ from the data distribution P_X, and with m_n independent latent random variables $Z_1, \ldots, Z_{m_n} \overset{i.i.d.}{\sim} P_Z$, we define the empirical loss associated with $T \in \mathcal{T}$ as

$$\widehat{L}_n^{\mathrm{NF}}(T) = \sup_{g \in \mathcal{G}_n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log g(X_i) - \frac{1}{m} \sum_{i=1}^{m_n} g(T^{-1}(Z_i)) \right\} + 1.$$

Observe that here we use n to denote the size of the true dataset that we are trying to model, while m_n denotes the number of latent data points we use. In general $m \equiv m_n$ can be much larger than the n, as the latent random variables are much easier to generate (a common choice is to select P_Z to be a known multivariate Gaussian distribution). We can now define the empirical NF model $\hat{T}_n$ as

$$\hat{T}_n \in \arg\min_{T \in \mathcal{T}} \widehat{L}_n^{\mathrm{NF}}(T). \tag{13}$$

The performance of the ERM estimator with a variational objective defined above is governed by two factors: the sampling error, and the alproximation error due to the restriction of the critic class to $\mathcal{G}_n$. Classical results from empirical process theory can allow us to control the difference between L^{NF} and $\widehat{L}^{\mathrm{NF}}$ in terms of certain notions of capacity of the class $\mathcal{G}_n$. Informally, smaller $\mathcal{G}_n$ will lead to much better finite sample approximation. On the other hand, reducing the difference $D_{KL}(P_X \parallel Q_T)$, that we refer to as the *variational gap*, requires a larger $\mathcal{G}_n$. This represents the key challenge in our analysis: finding the right tradeoff for the size of the critic class $\mathcal{G}_n$ that simultaneously controls these two terms. We

2. Our theoretical guarantees in Theorem 14 are also valid for homeomorphisms, but due to the specific requirements during the training of NF, here we state the definition based on diffeomorphisms.

first present a general result that assumes the existence of such a sequence of $\{\mathcal{G}_n : n \geq 1\}$ that satisfies the conditions stated in Assumption B.1. Then, in Appendix B.1, we identify sufficient that are easier to verify for Assumption B.1 to hold, and finally in Appendix B.2, we discuss a concrete realistic example satisfying these conditions in Appendix B.2.

Assumption B.1 *To analyze the NF model $\hat{T}_n$ defined in* (13)*, we require the following assumptions:*

- ***(NF1): Realizability.*** *For the P_X, P_Z under consideration, there exists a $T^* \in \mathcal{T}$ with $Q_{T^*} = P_X$, which implies that $L^{\mathrm{NF}}(T^*) = 0$ assuming that $\mathbf{1} \in \mathcal{G}_n$ (note that we allow the critic class $\mathcal{G}_n$ to grow with n, keeping $\mathcal{T}$ fixed).*
- ***(NF2): Uniform Convergence.*** *There exists a deterministic sequence $r_n \to 0$ and a confidence level $\delta \in (0, 1)$, such that with probability $1 - \delta$, we have*

$$\sup_{T \in \mathcal{T}} |\widehat{L}_{NF}(T) - L_{NF}(T)| \leq r_n.$$

  In other words, we assume that the function classes $(\mathcal{G}_n, \mathcal{T})$ are small enough to ensure uniform learnability.

- ***(NF3): Variational Approximation Gap.*** *$P_X \ll Q_T$ for all $T \in \mathcal{T}$, and there exists a deterministic sequence η_n converging to 0 as $n \to \infty$, such that*

$$\sup_{T \in \mathcal{T}} D_{KL}(P_X \parallel Q_T) - L_{NF}(T) \leq \eta_n.$$

  In other words, we assume that the capacity of the critic class $\mathcal{G}_n$ grows with n to approximate the likelihood ratios of all distributions modeled by elements of $\mathcal{T}$ and the true data distribution P_X.

- ***(NF4): Moment Bounds.*** *There exists positive constants R and a, such that*

$$\max\left\{\mathbb{E}[\|X\|^{1+a}],\ \sup_{T \in \mathcal{T}} \mathbb{E}[\|T^{-1}(Z)\|^{1+a}]\right\} \leq R.$$

With these assumptions in place, we can now state the main result of this section.

Theorem 14 *Under Assumption B.1, the empirical risk minimizer $\hat{T}_n$ defined in* (13) *satisfies the following with probability at least $1 - \delta$:*

$$W_1(P_X, Q_{\hat{T}_n}) \lesssim (r_n + \eta_n)^{\frac{a}{2(1+a)}},$$

where $\lesssim$ suppresses a multiplicative constant that depends on a, and W_1 is the 1-Wasserstein metric.

The proof of this statement is given in Appendix B.3. This theorem obtains a Hölder-type transfer from the original empirical variational objective to the Wasserstein error. In particular, it says that the quality of approximation achieved by the NF model $\hat{T}_n$ depends on the sum of uniform empirical-process deviation r_n and the variational gap η_n. The exponent $a/(1 + a)$ encodes the tail behavior from Assumption B.1: distributions with lighter tails (or larger a) yield a stronger, near-$\sqrt{\cdot}$ relation between W_1 and $r_n + \eta_n$, while heavier tails cause a weaker transformation.

Remark 15 *The proof of Theorem 14 proceeds in three steps. (i) On the high-probability event from **(NF2)**, ERM and realizability assumption imply that $L^{\mathrm{NF}}(\hat{T}_n) \le 2r_n$. Combining this with the variational gap assumption **(NF3)** yields $D_{KL}(P_X \parallel Q_{\hat{T}_n}) \le 2r_n + \eta_n$. (ii) The next step is to use Pinsker's inequality to get $\|P_X - Q_{\hat{T}_n}\|_{TV} \le \sqrt{(1/2)(2r_n + \eta_n)}$. (iii) Finally, we employ the truncation argument formalized in Lemma 25 along with the bounded $(1+a)$-moment assumption **(NF4)** to get the required bound in W_1 metric.*

As this outline suggests, this exact argument goes through for any distance or divergence that dominates the total variation metric via a Pinsker-type inequality. Hence, Theorem 14 can be stated more generally for such a class divergence measures, which includes the Jensen Shannon, Hellinger, and chi-squared divergence.

B.1 Verifiable Sufficient Conditions for Assumption B.1

In this section, we work under the realizability assumption **(NF1)**, and derive sufficient conditions on $(\mathcal{T}, \mathcal{G}_n)$ for the other conditions to hold, such that the W_1 metric between the model distribution and the true distribution converges to zero as the sample size n increases.

We begin with the uniform convergence condition **(NF2)**, and identify sufficient conditions in terms of the Rademacher complexities of the function classes involved.

Proposition 16 *Let $(\epsilon_i)_{i\ge 1}$ denote i.i.d. Rademacher random variables, and with $\mathcal{H}_n$ denoting the function class $\{\log g : g \in \mathcal{G}_n\}$, define the expected Rademacher complexities*

$$\mathfrak{R}_n(\mathcal{H}_n) = \mathbb{E}_{\epsilon^n, X^n}\left[\sup_{h\in\mathcal{H}_n} \frac{1}{n}\sum_{i=1}^n \epsilon_i h(X_i)\right], \quad \textit{and}$$
$$\mathfrak{R}_{m_n}(\mathcal{H}_n, \mathcal{T}) = \mathbb{E}_{\epsilon^{m_n}, Z^{m_n}}\left[\sup_{h,T} \frac{1}{m_n}\sum_{i=1}^n \epsilon_i h(T^{-1}(Z_i))\right].$$

If $\|h\|_\infty \le b_n$ for all $h \in \mathcal{H}_n$, for any $\delta > 0$, we have the following:

$$\sup_{T\in\mathcal{T}} \left|\widehat{L}_n^{\mathrm{NF}}(T) - L^{\mathrm{NF}}(T)\right| \le 2\mathfrak{R}_n(\mathcal{H}_n) + 2e^{b_n}\mathfrak{R}_{m_n}(\mathcal{H}_n, \mathcal{T}) + \left(\frac{b_n}{\sqrt{n}} + \frac{e^{b_n}}{\sqrt{m_n}}\right)\sqrt{2\log(4/\delta)}$$

*Thus, to verify **(NF2)**, we need to identify uniformly bounded $\mathcal{H}_n$ (equivalently $\mathcal{G}_n$) whose expected Rademacher complexity can be appropriately controlled to drive the above term to 0 with n.*

This result follows from the standard symmetrization via Rademacher variables argument, and we present the details in Appendix B.4. We next observe that if the model class $\mathcal{T}$ satisfies a uniform Lipschitz property, then the uniform moment control assumption **(NF4)** follows easily.

Proposition 17 *Suppose the function class $\mathcal{T}$ satisfies the following uniform affine growth property:*

$$\sup_{T\in\mathcal{T}} \|T^{-1}(z)\| \le J_0 + J_1\|z\| \quad \textit{for all} \quad z \in \mathcal{Z}.$$

If $\mathbb{E}_{P_Z}[\|Z\|^{1+a}] = M_Z < \infty$ *for an* $a > 0$*, we have the following under the realizability assumption* $(NF1)$*:*

$$\max\left\{\mathbb{E}[\|X\|^{1+a}],\ \sup_{T\in\mathcal{T}} \mathbb{E}\left[\|T^{-1}(Z)\|^{1+a}\right]\right\} \le 2^a\left(J_0^{1+a} + J_1^{1+a} M_Z\right) =: R.$$

The proof of this result is exactly the same as that of Proposition 7, and we omit it to avoid repetition. In other words, assuming realizability, a sufficient condition for **(NF4)** is if the latent variable Z has finite $(1+a)$ moments, and the model class is globally Lipschitz and uniformly bounded at the origin. Finally, we proceed to the condition of vanishing variational approximation gap **(NF3)**.

Proposition 18 *Suppose* $D_{KL}(P_X \parallel Q_T) < \infty$ *for all* $T \in \mathcal{T}$*, and let* $v_T = dP_X/dQ_T$ *and* $\ell_T = \log v_T$ *denote the likelihood ratio and log-likelihood ratio (resp.) of* P_X *and* Q_T*. Fix some sequence* $(K_n)_{n\ge1}$ *such that* $K_n \overset{n\to\infty}{\longrightarrow} \infty$*, and define the following uniform tail expectation terms, with* $X \sim P_X$*,* $X' \sim Q_T$*, and* $\mathcal{H}_n = \{\log g : g \in \mathcal{G}_n\}$*:*

$$\tau_1 \equiv \tau_1(K_n, \mathcal{T}, \mathcal{H}_n) = \sup_{T\in\mathcal{T}} \sup_{h\in\mathcal{H}_n} \mathbb{E}_{P_X}[|\ell_T(X) - h(X)|\, \mathbf{1}_{\|X\|_\infty > K_n}],$$

$$\text{and} \quad \tau_2 \equiv \tau_2(K_n, \mathcal{T}, \mathcal{H}_n) = \sup_{T\in\mathcal{T}} \sup_{h\in\mathcal{H}_n} \mathbb{E}_{Q_T}\left[\left|e^{h(X')} - e^{\ell_T(X')}\right| \mathbf{1}_{\|X'\|_\infty > K_n}\right].$$

Next, introduce a uniform approximation error term computed over a restricted domain:

$$\delta_n \equiv \delta_n(K_n, \mathcal{T}, \mathcal{H}_n) = \sup_{T\in\mathcal{T}} \inf_{h\in\mathcal{H}_n} \sup_{x:\|x\|_\infty \le K_n} |h(x) - \ell_T(x)|.$$

Suppose there exists a sequence of $\{K_n : n \ge 1\}$ *with* $K_n \uparrow \infty$*, such that*

$$\lim_{n\to\infty} \{\delta_n + \tau_1 + \tau_2\} = 0. \tag{14}$$

Then the variational gap η_n *also converges to zero; that is, if* (14) *holds, then*

$$\eta_n = \sup_{T\in\mathcal{T}} \left\{D_{KL}(P_X \parallel Q_T) - L^{\mathrm{NF}}(T)\right\} \overset{n\to\infty}{\longrightarrow} 0.$$

The proof of this result is in Appendix B.5.

B.2 V-NF Example

We now illustrate that the assumptions required by Theorem 14 are satisfied by practically useful models, by constructing specific $(\mathcal{T}, \mathcal{G}_n)$, and verifying this assumptions. Our model class is an iResNet flow with uniformly controlled Jacobians that ensures certain useful global regularity properties. Our critic class is induced by a Gaussian RKHS restricted to a growing cube $C_{K_n} = [-K_n, K_n]^d$, which allows a neat separation between (i) approximation on a compact set, and (ii) tail control outside the compact set, as needed by Proposition 18.

Definition 19 *We will work with i-ResNets of the form* $T = (I_d + F_M) \circ (I_d + F_{M-1}) \circ \cdots \circ (I_d + F_1)$*, where "*$\circ$*" denotes the composition operation, and*

$$F_j(x) = W_{j,2} \tanh\left(W_{j,1}x + b_{j,1}\right) + b_{j,2},$$
$$\text{with} \quad \max\left\{\|W_{j,1}\|_{op}, \|W_{j,2}\|_{op}\right\} \le s \in (0,1), \quad \text{and} \quad \max\{\|b_{j,1}\|,\ \|b_{j,2}\|\} \le B, \quad \forall j \in [M].$$

Each $W_{j,1}$ lies in $\mathbb{R}^{d_j} \times \mathbb{R}^d$ while $W_{j,2}$ lies in $\mathbb{R}^d \times \mathbb{R}^{d_j}$, with d_j denoting the dimension of the hidden layer in block j. Similarly, $b_{j,1} \in \mathbb{R}^{d_j}$ and $b_{j,2} \in \mathbb{R}^d$. Let H denote the maximum value of the hidden layer dimension; that is, $H = \max_{1 \le j \le M} d_j$, $\|\cdot\|_{op}$ denotes the operator norm, and $\|\cdot\|$ denotes the ℓ_2 norm.

Next, we introduce the critic function class $\mathcal{G}_n$.

Definition 20 *Let $k_n(x,y) \equiv k_{\gamma_n}(x,y) = \exp\left(-\gamma_n \|x-y\|^2\right)$ denote a Gaussian kernel with scale parameter $\gamma_n > 0$, and let $\mathcal{H}_{k_n}$ denote the reproducing kernel Hilbert space (RKHS) associated with this kernel. Then, we defile the critic class $\mathcal{G}_n$ as $\mathcal{G}_n = \{e^h : h \in \mathcal{H}_n\}$, where $\mathcal{H}_n = \{h\mathbf{1}_{[-K_n,K_n]^d} : h \in \mathcal{H}_{k_n}, \text{ with } \|h\|_{k_n} \le b_n\}$ for some $b_n, K_n > 0$. By the reproducing property, it follows that $\|h\|_{k_{\gamma_n}} \le b_n$ implies $\|h\|_\infty = \sup_{x \in \mathcal{X}} \langle h, k_{\gamma_n}(x,\cdot)\rangle_{k_{\gamma_n}} \le \sup_{x \in \mathcal{X}} \sqrt{k_{\gamma_n}(x,x)}\|h\| \le b_n$. As a result, every $g \in \mathcal{G}_n$ satisfies*

$$0 < e^{-b_n} \le g(x) \le e^{b_n}.$$

As mentioned above, we work with the i-ResNet model class $\mathcal{T}$ in Definition 19 since this satisfies a global Lipschitz continuity condition that allows us to work with unbounded latent random variables $Z \sim N(0, I_d)$. This global Lipschitz property turns out to be useful in establishing that the $(\mathcal{T}, \mathcal{G}_n)$ pair of Definition 19 and Definition 20 satisfy the conditions required by Assumption B.1. We can also ensure these assumptions are satisfied for the coupling-based architecture Dinh et al. (2016) by restricting the latent distribution P_Z to be supported on a compact domain. We now state the main result of this section.

Theorem 21 *Let $(\mathcal{T}, \mathcal{G}_n)$ be as in the two definitions above, and define the parameters*

$$\begin{aligned} K_n &= M(s\sqrt{H} + B) + \sqrt{d} + \sqrt{2\log n}, \\ \gamma_n &= K_n^{2+\epsilon} \quad \text{for some fixed } \epsilon > 0, \\ b_n &= C_b \gamma_n^{d/4} K_n^{2+d/2} = C_b K_n^{2+d+\frac{d}{4}\epsilon}, \\ m_n &= \lceil e^{4b_n} \rceil. \end{aligned}$$

*Then, under the realizability assumption **(NF1)**, and with these choices of the parameters, the ERM model $\hat{T}_n$ defined in (13) satisfies*

$$W_1(P_X, Q_{\hat{T}_n}) = o(1), \quad \text{w.p. at least } 1-\delta.$$

*In other words, with these parameters, the models in Definition 19 satisfy the sufficient conditions for **(NF2)**-**(NF4)** to hold, as derived in Appendix B.1.*

The proof of this result is in Appendix B.6.

B.3 Proof of Theorem 14

Introduce the uniform approximation event $\mathcal{E}_n = \{\sup_{T \in \mathcal{T}} |\widehat{L}_{NF}(T) - L_{NF}(T)| \le r_n\}$, and note that under the assumption **(NF2)**, we have $\mathbb{P}(\mathcal{E}_n) \ge 1-\delta$. For the rest of this proof, we will work under the event $\mathcal{E}_n$.

By definition of ERM model, we knlw that for any $T \in \mathcal{T}$, we have $\widehat{L}_{NF}(\hat{T}_n) \leq \widehat{L}_{NF}(T)$. This implies the following chain:

$$\begin{aligned} L_{NF}(\hat{T}_n) &\leq \widehat{L}_{NF}(\hat{T}_n) + r_n && \text{(event } \mathcal{E}_n\text{)} \\ &\leq \widehat{L}_{NF}(T^*) + r_n && \text{(ERM property)} \\ &\leq (L_{NF}(T^*) + r_n) + r_n && \text{(event } \mathcal{E}_n\text{)} \\ &= 2r_n && \text{(assumption } \textbf{(NF1)}\text{)}. \end{aligned}$$

Now, by the variational gap assumption **(NF3)**, we know that

$$\left| D_{KL}(P_X \parallel Q_{\hat{T}_n}) - L_{NF}(\hat{T}_n) \right| \leq \sup_{T \in \mathcal{T}} |D_{KL}(P_X \parallel Q_T) - L_{NF}(T)| \leq \eta_n.$$

Thus, combining the two displays above, we obtain the following bound on the relative entropy between the true and the NF distributions, under the $1-\delta$ probability event $\mathcal{E}_n$:

$$D_{KL}(P_X \parallel P_{\hat{T}_n}) \leq \eta_n + 2r_n \quad \overset{\text{Pinsker's}}{\Longrightarrow} \quad \|P_X - P_{\hat{T}_n}\|_{TV} \leq \sqrt{\frac{1}{2}(2r_n + \eta_n)}. \tag{15}$$

We can now conclude the proof with the following chain:

$$W_1(P_X, P_{\hat{T}_n}) \lesssim \left(\|P_X - P_{\hat{T}_n}\|_{TV} \right)^{\frac{a}{1+a}} \lesssim (2r_n + \eta_n)^{\frac{a}{2(1+a)}}.$$

Here the first inequality (modulo constants) is due to Lemma 25, and the second inequality is by (15).

B.4 Proof of Proposition 16

This follows from the standard uniform convergence results for bounded function classes. Observe that

$$\begin{aligned} \sup_{T \in \mathcal{T}} |\widehat{L}^{\mathrm{NF}}(T) - L^{\mathrm{NF}}(T)| &\leq \sup_{h \in \mathcal{H}_n} |(\mathbb{P}_n^{(X)} - \mathbb{P}^{(X)})h| + \sup_{T \in \mathcal{T}} \sup_{h \in \mathcal{H}_n} |(\mathbb{P}_{m_n}^{(Z)} - \mathbb{P}^{(Z)})e^h \circ T^{-1}| \\ &\coloneqq \mathrm{D}_X + \mathrm{D}_Z. \end{aligned}$$

Note that above we have used the notation $\mathbb{P}_n^{(X)} h = \frac{1}{n}\sum_{i=1}^n h(X_i)$, $\mathbb{P}^{(X)} h = \mathbb{E}_{X \sim P_X}[h(X)]$, $\mathbb{P}_{m_n}^{(Z)} e^h \circ T^{-1} = \frac{1}{m_n}\sum_{i=1}^{m_n} e^{h(T^{-1}(Z_i))}$ and $\mathbb{P}^{(Z)} e^h \circ T^{-1} = \mathbb{E}_{Z \sim P_Z}[e^{h(T^{-1}(Z))}]$.

Now observe that by the classical symmetrization technique, we have the following upper bound on the expected value of D_X (Shalev-Shwartz and Ben-David, 2014, Lemma 26.2):

$$\mathbb{E}[\mathrm{D}_X] \leq 2\mathfrak{R}_n(\mathcal{H}_n) = 2\mathbb{E}\left[\mathfrak{R}_n(\mathcal{H}_n, X^n)\right]$$

Here, $\mathfrak{R}_n$ denotes the Rademacher complexity whose definition is recalled in Definition 13 in Appendix A. Since we have assumed that $\|h\|_\infty \leq b_b$, for all $h \in \mathcal{H}_n$, an application of the bounded difference concentration inequality (recalled in Fact A.2 in Appendix A) immediately implies with some constant $C > 0$:

$$\mathbb{P}_{X^n}\left(\mathrm{D}_X \geq \mathbb{E}[\mathrm{D}_X] + Cb_n\sqrt{\frac{\log(2/\delta)}{n}} \right) \leq \mathbb{P}_{X^n}\left(\mathrm{D}_X \geq 2\mathfrak{R}_n(\mathcal{H}_n) + Cb_n\sqrt{\frac{\log(2/\delta)}{n}} \right) \leq \frac{\delta}{2}. \tag{16}$$

A similar argument works for the term D_Z. In particular, we know that the function class $\{e^{h\circ T^{-1}} : h \in \mathcal{H}_n, T \in \mathcal{T}\}$ is uniformly bounded from above by e^{b_n}. Thus, we can again get a high probability upper bound using the expected Rademacher complexity and the bounded difference inequality:

$$\mathbb{P}_{Z^{m_n}}\left(\mathrm{D}_Z \geq \mathbb{E}[\mathrm{D}_Z] + Ce^{b_n}\sqrt{\frac{\log(2/\delta)}{m_n}}\right) \leq \mathbb{P}_{Z^n}\left(\mathrm{D}_Z \geq 2\mathfrak{R}_{m_n}(\mathcal{G}_n \circ \mathcal{T}^{-1}) + Ce^{b_n}\sqrt{\frac{\log(2/\delta)}{m_n}}\right) \leq \frac{\delta}{2}. \tag{17}$$

Together, (17) and (16), along with the observation that $\mathfrak{R}_{m_n}(\mathcal{G}_n \circ \mathcal{T}^{-1}) \leq e^{b_n}\mathfrak{R}_{m_n}(\mathcal{H}_n, \mathcal{T})$ by the contraction result, imply the required uniform convergence bound with probability at least $1-\delta$:

$$\mathrm{D}_X + \mathrm{D}_Z \leq 2\mathfrak{R}_n(\mathcal{H}_n) + 2\mathfrak{R}_{m_n}(\mathcal{H}_n, \mathcal{T}) + C\sqrt{\log(2/\delta)}\left(\frac{b_n}{\sqrt{n}} + \frac{e^{b_n}}{\sqrt{m_n}}\right) =: r_n.$$

Thus, in order to establish a uniform convergence guarantee, it suffices to find a sequence of $\{(n, m_n, b_n, \mathcal{H}_n) : n \geq 1\}$, such that the term r_n in RHS above converges to 0.

B.5 Proof of Proposition 18

Throughout this proof, we assume that $X \sim P_X$ and $X' \sim Q_T$ for $T \in \mathcal{T}$, where Q_T represents the distribution of $T^{-1}(Z)$ for some latent variable $Z \sim P_Z$. Our argument below does not rely on how X' is generated, or on any properties of P_Z. Now, observe that for any $T \in \mathcal{T}$, the relative entropy $D_{KL}(P_X \parallel Q_T)$ is equal to $\mathbb{E}_P[\ell_T(X)]$, which implies that

$$\begin{aligned}\Delta_n(T) &\coloneqq D_{KL}(P_X \parallel Q_T) - \sup_{h\in\mathcal{H}_n}\left\{\mathbb{E}_{P_X}[h(X)] - \mathbb{E}_{Q_T}[e^{h(X')}] + 1\right\} \\ &= \inf_{h\in\mathcal{H}_n}\left\{\mathbb{E}_{P_X}[\ell_T(X) - h(X)] + \mathbb{E}_{Q_T}[e^{h(X')} - e^{\ell_T(X')}]\right\},\end{aligned}$$

where the second equality uses the fact that $E_{Q_T}[e^{\ell_T}] = 1$ (recall that ℓ_T is the log-likelihood ratio between P_X and Q_T). Now, for some $K_n \to \infty$, define the events $E_n = \{\|X\|_\infty \leq K_n\}$ and $F_n = \{\|X'\|_\infty \leq K_n\}$, and observe that

$$\Delta_n(T) \leq \inf_{h\in\mathcal{H}_n}\left\{\mathbb{E}_{P_X}[(\ell_T(X) - h(X))\,\mathbf{1}_{E_n}] + \mathbb{E}_{Q_T}[\left(e^{h(X')} - e^{\ell_T(X')}\right)\mathbf{1}_{F_n}]\right\} + \tau_1 + \tau_2, \tag{18}$$

where the two terms τ_1 and τ_2 are defined as

$$\begin{aligned}\tau_1(K_n, \mathcal{T}, \mathcal{H}_n) &= \sup_{T\in\mathcal{T}}\sup_{h\in\mathcal{H}_n}\mathbb{E}_{P_X}[|\ell_T(X) - h(X)|\,\mathbf{1}_{\|X\|_\infty > K_n}], \\ \text{and} \quad \tau_2(K_n, \mathcal{T}, \mathcal{H}_n) &= \sup_{T\in\mathcal{T}}\sup_{h\in\mathcal{H}_n}\mathbb{E}_{Q_T}\left[\left|e^{h(X')} - e^{\ell_T(X')}\right|\mathbf{1}_{\|X'\|_\infty > K_n}\right].\end{aligned}$$

It remains to analyze the first term in the RHS of (18). We proceed by observing that

$$\left|e^{\ell_T} - e^h\right| = e^{\ell_T}\left|e^{h-\ell_T} - 1\right| \leq e^{\ell_T}\left(|\ell_T - h| + \frac{e^{|\ell_T - h|}}{2}|\ell_T - h|^2\right),$$

where the inequality uses a second-order Taylor approximation of e^x around 0. More specifically, we use $|e^u - 1| \le |u| \le e^{|u|}\frac{u^2}{2}$ with $u \leftarrow \ell_T - h$. Now, let us introduce the following terms:

$$e^{b_n} \coloneqq \sup_{h\in\mathcal{H}_n}\sup_{x:\|x\|_\infty\le K_n} e^{h(x)}, \quad \text{and} \quad \delta_n = \sup_{T\in\mathcal{T}}\inf_{h\in\mathcal{H}_n}\sup_{x:\|x\|_\infty\le K_n}|\ell_T(x) - h(x)|.$$

Using these definitions in (18), we obtain

$$\begin{aligned}
\Delta_n(T) &\le \mathbb{E}_{P_X}[\delta_n \mathbf{1}_{E_n}] + \mathbb{E}_{Q_T}\left[e^{\ell_T(X')}\left(\delta_n + e^{\delta_n}\frac{\delta_n^2}{2}\right)\mathbf{1}_{F_n}\right] + \tau_1 + \tau_2 \\
&\le \delta_n + \left(\delta_n + e^{\delta_n}\frac{\delta_n^2}{2}\right)\mathbb{E}_{Q_T}[e^{\ell_T}] + \tau_1 + \tau_2 \\
&= \delta_n\left(2 + e^{\delta_n}\delta_n/2\right) + \tau_1 + \tau_2.
\end{aligned}$$

Since the RHS is independent of T, taking a supremum over the function class $\mathcal{T}$, gives us the required

$$\eta_n \;=\; \sup_{T\in\mathcal{T}}\Delta_n(T) \;\le\; \left(2 + e^{\delta_n}\delta_n/2\right)\delta_n + \tau_1 + \tau_2,$$

which converges to 0 if δ_n, τ_1, τ_2 converge to 0. This concludes the proof.

B.6 Proof of Theorem 21

To prove this result, we need to verify that the sufficient conditions obtained in Proposition 16 (uniform convergence), Proposition 17 (moment bounds), and Proposition 18 (variational gap), are satisfied by our specific choices in Definition 19 and Definition 20. We verify these conditions in Appendix B.6.1, Appendix B.6.2, and Appendix B.6.3 respectively. Before proceeding to these steps, we first establish certain properties of the function class $\mathcal{T}$ that will be used often.

Suppose P_Z is a Gaussian distribution with identity covariance, and density $p_Z(z) \propto \exp(-\|z\|^2/2)$, which implies that $\psi(z) = \log p_Z(z) = -\|z\|^2/2 + \text{const}$. Under the realizability assumption, there exists $T^* \in \mathcal{T}$ such that $P_X = Q_{T^*}$. Every $T \in \mathcal{T}$ can be represented as

$$T = G_M \circ G_{M-1} \circ \ldots \circ G_1, \quad \text{where} \quad G_j(x) = x + F_j(x), \;\; F_j(x) = W_{j,2}\tanh(W_{j,1}x + b_{j,1}) + b_{j,2}.$$

By assumption, we have $\|W_{j,i}\|_{op} \le s \in (0,1)$ and $\|b_{j,i}\| \le B$ for all j, i, where $\|\cdot\|$ denotes the ℓ_2 norm, and $\|\cdot\|_{op}$ denotes the operator norm induced by $\|\cdot\|$.

B.6.1 Verification of the Uniform Convergence Assumption

To simplify the notation, introduce the function classes $\mathcal{S}_n = \{g \circ T^{-1} : g \in \mathcal{G}_n, T \in \mathcal{T}\}$ and $\mathcal{U}_n = \{(1/b_n)h \circ T^{-1} : h \in \mathcal{H}_n, T \in \mathcal{T}\}$. Recall that $\mathcal{H}_n$ consists of functions of the form $\tilde{h}\mathbf{1}_{[-K_n,K_n]^d}$ for all $\tilde{h}$ lying in the RKHS of a Gaussian kernel k_n (denoted by $\mathcal{H}_{k_n}$), with $\|\tilde{h}\|_{k_n} \le b_n$. We first consider the term $\mathfrak{R}_{m_n}(\mathcal{S}_n)$, since the bound on the other term

follows similarly. Observe that $\mathcal{S}_n = \{e^{b_n u} : u \in \mathcal{U}_n\}$ by definition, leading to the following inequality via the contraction lemma (Shalev-Shwartz and Ben-David, 2014, Lemma 26.9):

$$\mathfrak{R}_{m_n}(\mathcal{S}_n, Z^{m_n}) \leq e^{b_n} \mathfrak{R}_{m_n}\left(\mathcal{U}_n, Z^{m_n}\right).$$

It remains for us to control the Rademacher complexity of the composition class $\mathcal{U}_n$. For any $z^{m_n} = (z_1, \ldots, z_{m_n})$, introduce the terms $\mathcal{I} \equiv \mathcal{I}(z^n, K_n) = \{i \in [m_n] : z_i \in [-K_n, K_n]^d\}$ and $s_n \equiv s_n(z^{m_n}, K_n) = |\mathcal{I}|$. For z^{m_n} such that $s_n > 0$, observe that

$$\begin{aligned}
b_n \mathfrak{R}_{m_n}(\mathcal{U}_n, z^{m_n}) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{T \in \mathcal{T}} \sup_{h \in \mathcal{H}_n} \frac{1}{m_n} \sum_{i=1}^{m_n} \sigma_i h(T^{-1}(z_i))\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{T \in \mathcal{T}} \sup_{\tilde{h}: \|\tilde{h}\|_{k_n} \leq b_n} \frac{1}{m_n} \sum_{i \in \mathcal{I}} \sigma_i \tilde{h}(T^{-1}(z_i))\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{T \in \mathcal{T}} \sup_{\tilde{h}: \|\tilde{h}\|_{k_n} \leq b_n} \left\langle \tilde{h}, \frac{1}{m_n} \sum_{i \in \mathcal{I}} \sigma_i k_n(T^{-1}(z_i), \cdot) \right\rangle_{k_n}\right] \quad \text{(reproducing property)} \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{T \in \mathcal{T}} b_n \left\| \frac{1}{m_n} \sum_{i \in \mathcal{I}} \sigma_i k_n(T^{-1}(z_i), \cdot) \right\|_{k_n}\right] \quad \left(\text{optimal } h \propto \frac{b_n}{m_n} \sum_i k_n(T^{-1}(z_i), \cdot)\right) \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{b_n}{m_n} \sqrt{\sup_{T \in \mathcal{T}} \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{I}} \sigma_i \sigma_j k(T^{-1}(z_i), T^{-1}(z_j))}\right] \\
&\leq \frac{b_n}{m_n} \sqrt{\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{i=1}^{m_n} \sum_{j=1}^{m_n} \sigma_i \sigma_j\right]} \qquad \left(\text{Jensen's} + \sup_{x, x'} k(x, x') \leq 1\right) \\
&= \frac{b_n \sqrt{s_n}}{m_n}.
\end{aligned}$$

As a result, with an i.i.d. sample Z^{m_n}, the expected Rademacher complexity satisfies

$$\begin{aligned}
b_n \mathfrak{R}_{m_n}(\mathcal{U}_n) = b_n \mathbb{E}_{Z^{m_n}}\left[\mathfrak{R}_{m_n}(\mathcal{U}_n, Z^{m_n})\right] &\leq \frac{b_n}{m_n} \mathbb{E}_{Z^{m_n}}\left[\sqrt{\sum_{i=1}^{m_n} \mathbf{1}_{Z_i \in [-K_n, K_n]^d}}\right] \\
&\leq b_n \sqrt{\frac{P_Z([-K_n, K_n]^d)}{m_n}} \leq \frac{b_n}{\sqrt{m_n}}.
\end{aligned}$$

An exactly analogous argument implies $\mathfrak{R}_n(\mathcal{H}_n) \leq b_n/\sqrt{n}$. Combining these two bounds, we see that the uniform convergence rate obtained in Proposition 16 reduces to the following (up to leading constants):

$$r_n \lesssim \frac{b_n}{\sqrt{n}} + \frac{b_n e^{b_n}}{\sqrt{m_n}}. \tag{19}$$

This can be made to converge to 0 with n, by selecting $b_n = o(\sqrt{n})$ and $m_n = \Omega(e^{(2+\epsilon)b_n})$ for any $\epsilon > 0$. Observe that the choices of b_n, m_n in Theorem 21 satisfy these conditions, as we discuss further at the end of this section.

B.6.2 Verification of the Moment Assumption

Lipschitz property. Observe that tanh is 1-Lipschitz as its derivative is $\text{sech}^2 \in [0,1]$, which implies

$$\text{Lip}(F_j) \leq \|W_{j,2}\|_{op}\|W_{j,1}\|_{op} \leq s^2 =: \rho \in (0,1). \tag{20}$$

Hence, for every block $G_j = I_d + F_j$, we have

$$(1-\rho)\|x - x'\| \leq \|G_j(x) - G_j(x')\| \leq (1+\rho)\|x - x'\|$$

Since T is the composition of M such blocks, we immediately obtain

$$(1-\rho)^M\|x - x'\| \leq \|T(x) - T(x')\| \leq (1+\rho)^M\|x - x'\|.$$

This allows us to conclude that every $T \in \mathcal{T}$ is bi-Lipschitz with

$$\text{Lip}(T) \leq (1+s^2)^M, \quad \text{and} \quad \text{Lip}(T^{-1}) \leq (1-s^2)^{-M}.$$

Since these constants are independent of T, they are also uniformly valid over the class $\mathcal{T}$:

$$\sup_{T\in\mathcal{T}} \max\left\{\text{Lip}(T), \text{Lip}(T^{-1})\right\} \leq \max\left\{(1-s^2)^{-M}, (1+s^2)^M\right\} =: J_1. \tag{21}$$

Values at $x = 0$. Fix any $T = G_M \circ \ldots \circ G_1$ in $\mathcal{T}$, and consider $x_0 = 0$, $x_j = G_j(x_{j-1}) = x_{j-1} + F_j(x_{j-1})$ for $j \in [M]$. Observe that by (20), we have for any x and j:

$$\|F_j(x)\| \leq \|F_j(0)\| + s^2\|x\|.$$

Now, $\|F_j(0)\|$ can be bounded as

$$\|F_j(0)\| = \|W_{j,2}\tanh(b_{j,1}) + b_{j,2}\| \leq \|W_{j,2}\|_{op}\left(\|\tanh(b_{j,1})\| + \|b_{j,2}\|\right).$$

We know that $\|W_{j,2}\|_{op} \leq s$ and $\|b_{j,i}\| \leq B$ for $i = 1, 2$. Furthermore, let H denote the maximum dimension of the hidden layer in M blocks; that is, $H = \max_{j\in[M]} d_j$. Then, $\|\tanh(b_{j,1})\| \leq \sqrt{d_j}\tanh(B) \leq \sqrt{H}$. These facts imply that

$$\|F_j(0)\| \leq C_0 \coloneqq s\sqrt{H} + B.$$

This implies that for any $j \in [M]$, due to the Lipschitz property, we have

$$\begin{aligned}\|x_j\| = \|G_j(x_{j-1})\| = \|x_{j-1} + F_j(x_{j-1})\| &\leq \|x_{j-1}\| + \|F_j(x_{j-1})\| \leq \|x_{j-1}\| + \|F_j(0)\| + s^2\|x_{j-1}\| \\ &\leq C_0 + (1+s^2)\|x_{j-1}\|.\end{aligned}$$

Applying this inequality iteratively, with $x_0 = 0$ and $x_M = T(x_0)$, we have

$$\|T(0)\| \leq C_0\sum_{i=0}^{M-1}(1+s^2)^i = C_0\left(\frac{(1+s^2)^M - 1}{s^2}\right). \tag{22}$$

Finally, for any T, let y_T denote $T^{-1}(0)$, and observe that

$$\|y_T\| = \|T^{-1} \circ T(y_T - 0)\| \leq \mathrm{Lip}(T^{-1})\|T(y_T) - T(0)\| = \mathrm{Lip}(T^{-1})\|T(0)\|. \tag{23}$$

Together, (22) and (23) imply

$$\sup_{T\in\mathcal{T}} \max\left\{\|T(0)\|, \|T^{-1}(0)\|\right\} \leq J_1(s\sqrt{H} + B)\left(\frac{(1+s^2)^M - 1}{s^2}\right) =: J_0. \tag{24}$$

Since $Z \sim N(0, I_d)$ has finite moments of all orders and $\|T^{-1}(z)\| \leq J_0 + J_1\|z\|$ uniformly in T, we have that $\sup_{T\in\mathcal{T}} \mathbb{E}[\|T^{-1}(Z)\|^a] < \infty$ for all $a > 0$. Under the realizability assumption, $X \overset{d}{=} (T^*)^{-1}(Z)$, hence this also implies $\mathbb{E}\|X\|^{1+a} < \infty$. Thus, with an arbitrary $a > 0$, and $M_Z = 2^{a/2}\Gamma((a+d)/2)/\Gamma(d/2)$, we get

$$R = 2^a\left(J_0^{1+a} + J_1^{1+a}2^{a/2}\Gamma((a+d)/2)/\Gamma(d/2)\right).$$

Recall that J_0 was defined in (24), and J_1 in (21). This completes the proof.

B.6.3 Verification of the Variational Gap Assumption

To verify this assumption for the model introduced in Definition 19, we will first obtain the Jacobian of the functions in $\mathcal{T}$, which will then allow us to obtain a closed form expression for the log likelihood ratio dP_X/dQ_T and characterize its behavior on a cube $C_{K_n} = [-K_n, K_n]^d$.

Lemma 22 *Consider any $T = G_M \circ G_{M-1} \circ \cdots \circ G_1 \in \mathcal{T}$ of Definition 19, where each block $G_j(x) = (I_d + F_j)(x) = x + W_{j,2}\tanh(W_{j,1}x + b_{j,1}) + b_{j,2}$. Then, we have the following:*

$$\mathbb{J}G_j(x) = I_d + W_{j,2}D_j(x)W_{j,1}, \quad \textit{where} \quad D_j(x) = \mathrm{diag}(\mathrm{sech}^2(W_{j,1}x + b_{j,1}))$$
$$\|\mathbb{J}F_j(x)\|_{op} \leq s^2, \quad \textit{and} \quad \|\mathbb{J}F_j(x) - \mathbb{J}F_j(y)\|_{op} \leq 2s^3\|x - y\|.$$

This result, proved in Appendix D.3, tells us that every G_j is a diffeomorphism with singular values in $[1 - s^2, 1 + s^2]$, and hence, $\det \mathbb{J}G_j(x) > 0$ which implies that $\log\det \mathbb{J}G_j$ is well defined. We now obtain a closed-form expression for the log-likelihood ratio $\ell_T = \log dP_X/dQ_T$.

Lemma 23 *The log-likelihood ratio function $\ell_T(x) = \log dP_X/dQ_T$ for any $T \in \mathcal{T}$ is equal to*

$$\ell_T(x) = \frac{1}{2}\left(\|T(x)\|^2 - \|T^*(x)\|^2\right) + \log\det \mathbb{J}T^*(x) - \log\det \mathbb{J}T(x).$$

Furthermore, when restricted to the domain $[-K, K]^d$, each ℓ_T is Lipschitz with constant $L_1 = \mathcal{O}(K)$ (see (51) for exact expression), which implies the following uniform bound

$$\sup_{T\in\mathcal{T}} \sup_{x\in[-K,K]^d} |\ell_T(x)| \leq L_0 + L_1\sqrt{d}K,$$

for a universal constant L_0 stated in (52)

This result, proved in Appendix D.4, establishes that when restricted to a box $[-K, K]^d$ in the domain $\mathcal{X} = \mathbb{R}^d$, the likelihood ratio function class $\{\ell_T : T \in \mathcal{T}\}$ is well behaved. With these results at hand, we can now proceed towards verifying the sufficient conditions for the variational gap to vanish as n goes to ∞.

Approximation error δ_n. We begin by obtaining a characterization of the approximation error δ_n.

Lemma 24 *Consider the given $\mathcal{H}_n = \{\tilde{h}\mathbf{1}_{[-K_n,K_n]^d} : \tilde{h} \in \mathcal{H}_{k_n},\ \|\tilde{h}\|_{k_n} \leq b_n\}$, where $\mathcal{H}_{k_n}$ is the RKHS associated with the Gaussian kernel $k_n(x,y) = \exp(-\gamma_n\|x-y\|^2)$ for a scale parameter γ_n. For parameters b_n, γ_n, and K_n, we have*

$$b_n \gtrsim \gamma_n^{d/4} K_n^{2+d/2} \quad \implies \quad \delta_n(K_n) \lesssim \frac{K_n}{\sqrt{\gamma_n}}. \tag{25}$$

Here $a_n \lesssim b_n$ represents that there exists a constant c that does not change with n, such that $a_n \leq cb_n$ for all n large enough.

The proof of this statement is in Appendix D.5.

Tail term τ_1. Let us consider the first tail term τ_1, whose definition is recalled below:

$$\tau_1 \equiv \tau_1(K_n, \mathcal{T}, \mathcal{H}_n) = \sup_{T\in\mathcal{T}} \sup_{h\in\mathcal{H}_n} \mathbb{E}_{P_X}[|\ell_T(X) - h(X)|\ \mathbf{1}_{\|X\|_\infty > K_n}].$$

Fix an h and T, and observe that

$$\begin{aligned}\mathbb{E}_{P_X}[|\ell_T(X) - h(X)|\mathbf{1}_{\|X\|_\infty>K_n}] &\leq \mathbb{E}_{P_X}[|\ell_T(X)|\mathbf{1}_{\|X\|_\infty>K_n}] + \mathbb{E}_{P_X}[|h(X)|\mathbf{1}_{\|X\|_\infty>K_n}] \\ &=: \tau_{11}(T,h) + \tau_{12}(T,h).\end{aligned}$$

Now, from Lemma 23, we know that there exist constants c_0, c_1, c_2 independent of n, T, such that we have

$$|\ell_T(x)| \leq c_0 + c_1\|x\| + c_2\|x\|^2 \quad \text{for all} \quad x \in \mathbb{R}^d.$$

If the distribution P_X is such that $\mathbb{E}_{P_X}[\|X\|^2] < \infty$, then, this implies that

$$\tau_{11}(T,h) \leq \mathbb{E}_{P_X}[\left(c_0 + c_1\|X\| + c_2\|X\|^2\right)\mathbf{1}_{\|X\|_\infty>K_n}] \overset{n\to\infty}{\longrightarrow} 0.$$

Thus, the only condition needed to control τ_{11} is that $K_n \to \infty$, and that $\mathbb{E}_{P_X}[\|X\|^2] < \infty$.

Next, we look at the other term τ_{12}. This is much easier to handle, since we know that $\|h\|_\infty \leq b_n$ for all $h \in \mathcal{H}_n$. Hence, we have

$$\tau_{12} \leq b_n \mathbb{P}_{P_X}(\|X\|_\infty > K_n) \leq b_n e^{-cK_n^2},$$

for some constant c (depending on the Lipschitz constant of T^*). Hence, a $K_n = \Omega(\sqrt{\log b_n})$ is sufficient to drive this to 0.

Tail term τ_2. Finally, we consider the remaining tail term τ_2, defined as

$$\tau_2 \equiv \tau_2(K_n, \mathcal{T}, \mathcal{H}_n) = \sup_{T\in\mathcal{T}} \sup_{h\in\mathcal{H}_n} \mathbb{E}_{Q_T}\left[\left|e^{h(X')} - e^{\ell_T(X)}\right| \mathbf{1}_{\|X'\|_\infty>K_n}\right]$$

For fixed $T \in \mathcal{T}$ and $h \in \mathcal{H}_n$, observe that since $h(x) = 0$ for all $x : \|x\|_\infty > K_n$, we get

$$\begin{aligned}\mathbb{E}_{Q_T}\left[\left|e^{h(X')} - e^{\ell_T(X)}\right| \mathbf{1}_{\|X'\|_\infty > K_n}\right] &= \mathbb{E}_{Q_T}\left[|1 - e^{\ell_T}| \mathbf{1}_{\|X'\|_\infty > K_n}\right] \\ &\leq \mathbb{P}_T(\|X'\|_\infty > K_n) + \mathbb{E}_{Q_T}\left[e^{\ell_T} \mathbf{1}_{\|X'\|_\infty > K_n}\right] \\ &= \mathbb{P}(\|X'\|_\infty > K_n) + \mathbb{P}\left(\|X\|_\infty > K_n\right).\end{aligned}$$

Thus, we have

$$\tau_2 \leq \sup_T \mathbb{P}\left(\|T^{-1}(Z)\|_\infty > K_n\right) + \mathbb{P}\left(\|(T^*)^{-1}Z\| > K_n\right),$$

which goes to 0 with n under the $\mathcal{T}$-uniform Lipschitz bounds derived in Lemma 22.

B.6.4 Completing the Proof

To complete the proof, we will show an explicit choice of the free parameters that drives the W_1-metric between the true and the NF distributions to zero. In particular, we observe the following:

- With $m_n = \lceil e^{4b_n} \rceil$, we get $e^{b_n}/\sqrt{m_n} \leq e^{-b_n}$. Plugging this in (19) then gives us

$$r_n \lesssim \frac{b_n}{\sqrt{n}} + b_n e^{-b_n} \lesssim \frac{b_n}{\sqrt{n}}.$$

- Now, set $b_n = C_b \gamma^{d/4} K_n^{2+d/2+\epsilon}$, and $\gamma_n = K_n^{2+\epsilon}$ for some $\epsilon > 0$. This implies that

$$b_n \asymp K_n^{2+d+\epsilon\left(1+\frac{d}{4}\right)} \quad \Longrightarrow \quad r_n \lesssim \frac{K_n^{2+d+\epsilon\left(1+\frac{d}{4}\right)}}{\sqrt{n}}.$$

- The above choice of b_n ensures that $b_n \gtrsim \gamma_n^{d/4} K_n^{2+d/2}$, and hence (25) applies to these parameter choices, which leads to

$$\delta_n(K_n) \lesssim \frac{K_n}{\sqrt{\gamma_n}} \lesssim K_n^{-\epsilon/2}.$$

- Due to the uniform Lipschitz condition on the elements of $\mathcal{T}$ along with the fact that $Z \sim N(0, I_d)$, we can show that $\tau_1 + \tau_2 \lesssim \text{poly}(\text{K}_\text{n}) e^{-cK_n^2}$ for some $c > 0$. With the choice of $K_n \asymp \sqrt{\log n}$ as stated in Theorem 14, this reduces to $\tau_1 + \tau_2 \lesssim \text{poly}(\log n) n^{-c}$. Thus, combining with the previous state, we get

$$\delta_n + \tau_1 + \tau_2 \lesssim (\log n)^{-\epsilon/4} + \text{poly}(\log n) n^{-c}.$$

Thus, with the choice of the parameters in Theorem 21, the estimation error converges to 0 with probability at least $1 - \delta$.

Appendix C. Deferred Proofs from Section 3.2

C.1 Proof of Theorem 1

To simplify the notation, we will use $P^{(A)}$ to represent the conditional probability $P(\cdot \mid Y \in A)$. By the realizability assumption, we have $P^{(A)}_{Y,T^*_z(X)} = P^{(A)}_{YZ}$. The starting point of the proof is to use the realizability and the uniform convergence assumptions to relate the population losses associated with the ERM model $\hat{T}_n$ to the error terms u_n and r_n from Assumption 3.1. For any $T \in \mathcal{T}$, let $L(T) = L_{\mathbf{y}}(T) + \lambda L_{\mathbf{z}}(T)$, and $\widehat{L}_n(T) = \widehat{L}_{\mathbf{y},n}(T) + \lambda \widehat{L}_{\mathbf{z},n}(T)$, and observe the following chain:

$$0 = L(T^*) \geq \widehat{L}_n(T^*) - u_n - \lambda r_n \geq \widehat{L}_n(\hat{T}_n) - u_n - \lambda r_n \geq L(\hat{T}_n) - 2(u_n + \lambda r_n).$$

The equality follows from the realizability assumption **(INN1)**, while the first and third inequalities use the uniform convergence assumption **(INN2)**, and the second inequality follows form the definition of $\hat{T}_n$ as the ERM model. This simple argument allows us to control the individual components of $L(\hat{T}_n)$ as follows:

$$L_{\mathbf{y}}(\hat{T}_n) \leq 2(u_n + \lambda r_n), \quad \text{and} \quad L_{\mathbf{z}}(\hat{T}_n) \leq 2\left(r_n + \frac{u_n}{\lambda}\right). \tag{26}$$

For the first inequality, we use the fact that $L_{\mathbf{z}}(T) \geq 0$ for all T as the critic class $\mathcal{G}_n$ is assumed to contain the $\mathbf{0}$ element, while the second inequality uses the fact that $L_{\mathbf{y}}(T) \geq 0$ by definition of the squared loss.

The next step of the proof is to use the Lipschitz property of the inverse map $\hat{T}_n^{-1}$ to relate the W_1 metric between the posterior distributions $P^{(A)}_{\hat{T}_n(X)}$ and $P^{(A)}_{YZ}$. In particular, to further simplify the notation and use $\mu = P^{(A)}_{\hat{T}_n(X)}$, $\rho = P^{(A)}_{YZ}$, and let F denote the inverse map $\hat{T}_n^{-1} : \mathcal{Y} \times \mathcal{Z} \to \mathcal{X}$. Then, $F_{\#}\mu$ and $F_{\#}\rho$ denote the true and INN posterior distributions on $\mathcal{X}$ by definition (recall that we use $F_{\#}\mu$ to represent the pushforward of measure μ on to the image space of F; that is for any measurable $E \subset \mathcal{X}$, we have $(F_{\#}\mu)(E) = \mu(F^{-1}(E))$). Let $\pi \in \Pi(\mu, \rho)$ denote any coupling on $(\mathcal{Y} \times \mathcal{Z})^2$ with marginals μ and ρ, and let $\pi_X \in \Pi(F_{\#}\mu, F_{\#}\rho)$ denote the coupling by pushing π forward under $F \times F$ (recall that $F = \hat{T}_n^{-1}$); that is, $\pi_X(E_X \times G_X) = \pi\left(\{(y,z) : F(y) \in E_X,\ F(z) \in G_X\}\right)$. With these definitions, we immediately have

$$\begin{aligned}
W_1(F_{\#}\mu, F_{\#}\rho) &= \inf_{\pi' \in \Pi\left(F_{\#}\mu, F_{\#}\rho\right)} \int_{\mathcal{X}\times\mathcal{X}} \|x - x'\| d\pi'(x,x') = \inf_{\substack{\pi_X = (F\times F)_{\#}\pi \\ \pi \in \Pi(\mu,\rho)}} \int_{\mathcal{X}\times\mathcal{X}} \|x - x'\| d\pi_X(x,x') \\
&= \inf_{\pi \in \Pi(\mu,\rho)} \int_{\mathcal{X}\times\mathcal{X}} \|F(y,z) - F(y',z')\| d\pi((y,z),(y',z')) \\
&\leq J \inf_{\pi \in \Pi(\mu,\rho)} \int_{(\mathcal{Y}\times\mathcal{Z})^2} \|(y,z) - (y',z')\| d\pi((y,z),(y',z')) = JW_1(\mu,\rho).
\end{aligned}$$

Putting back the values $\mu \leftarrow P^{(A)}_{\hat{T}_n(X)}$, $\rho \leftarrow P^{(A)}_{YZ}$, and $F \leftarrow \hat{T}_n^{-1}$, the inequality chain above implies

$$W_1\left(P^{(A)}_X, P^{(A)}_{\hat{T}_n^{-1}(Y,Z)}\right) \leq JW_1\left(P^{(A)}_{\hat{T}_n(X)}, P^{(A)}_{YZ}\right).$$

Let us introduce the notation:

$$\nu = P^{(A)}_{\hat{T}_n(X)}, \quad \gamma = P^{(A)}_{Y,\hat{T}_{n,z}(X)}, \quad \text{and} \quad \omega = P^{(A)}_{Y,Z}.$$

Using the fact that W_1 is a metric, we can use triangle inequality to get

$$W_1\left(P^{(A)}_{\hat{T}_n(X)}, P^{(A)}_{YZ}\right) = W_1(\nu,\omega) \le W_1(\nu,\gamma) + W_1(\gamma,\omega). \tag{27}$$

To complete the proof, we need to obtain upper bounds on the two terms in the RHS above. To control the first term in (27), We begin by recalling the explicit definition of the first term

$$W_1(\nu,\gamma) = \inf_{\pi\in\Pi(\nu,\gamma)} \int \|(y,z)-(y',z')\| d\pi((y,z),(y',z'))$$

Let us now construct the natural coupling π between ν and γ in the following steps:

- Generate $(X,Y) \sim P^{(A)}_{XY}$.
- Let ν denote the conditional distribution of $\hat{T}_n(X) = (\hat{T}_{n,y}(X), \hat{T}_{n,z}(X))$
- Let γ denote the conditional distribution of $(Y, \hat{T}_{n,z}(X))$. Thus ν and γ have a common second component.

Using this particular coupling in the definition of $W_1(\nu,\gamma)$, we observe that

$$\begin{aligned} W_1(\nu,\gamma) &\le \int \sqrt{\|\hat{T}_{n,y}(x)-y\|^2 + \|\hat{T}_{n,z}(x)-\hat{T}_{n,z}(x)\|^2}\, dP^{(A)}_{XY}(x,y) = \mathbb{E}^{(A)}_{XY}\left[\|\hat{T}_{n,y}(X)-Y\|\right] \\ &\le \sqrt{\mathbb{E}^{(A)}_{XY}\left[\|\hat{T}_{n,y}(X)-Y\|^2\right]} \overset{(28).1}{\le} \frac{\sqrt{L_{\mathbf{y}}(\hat{T}_n)}}{\sqrt{P_Y(A)}} \le \frac{\sqrt{2(u_n+\lambda r_n)}}{\sqrt{P_Y(A)}}, \end{aligned} \tag{28}$$

where the last inequality follows from (26), and (28).1 uses the fact that $L_{\mathbf{y}}(\hat{T}_n) = \mathbb{E}_{XY}[\|\hat{T}_{n,y}(X)-Y\|^2] = P_Y(A)\mathbb{E}^{(A)}_{XY}[\|\hat{T}_{n,y}(X)-Y\|^2] + P_Y(A^c)\mathbb{E}^{(A^c)}_{XY}[\|\hat{T}_{n,y}(X)-Y\|^2] \ge P_Y(A)\mathbb{E}^{(A)}_{XY}[\|\hat{T}_{n,y}(X)-Y\|^2]$.

It remains for us to obtain a bound on the term $W_1(\gamma,\omega)$. Introduce the conditional total variation (TV) distance between these two measures,

$$\Delta_A = \|P^{(A)}_{YZ} - P^{(A)}_{Y,\hat{T}_{n,z}(X)}\|_{TV} = \|\omega-\gamma\|_{TV},$$

and observe that an application of the truncation argument stated in Lemma 25 gives us

$$W_1(\gamma,\omega) \le C_a \left(\frac{2R}{P_Y(A)}\right)^{\frac{1}{1+a}} \Delta_A^{\frac{a}{1+a}}, \quad \text{where} \quad C_a = 2\left(a^{\frac{1}{1+a}} + a^{\frac{-a}{1+a}}\right). \tag{29}$$

Next, we use the fact that the two distributions γ and ω share the marginal distribution of Y, which can be used to show that

$$D_f\left(P^{(A)}_{Y,\hat{T}_{n,z}(X)} \,\|\, P^{(A)}_{YZ}\right) \overset{(30).1}{\le} \frac{1}{P_Y(A)} D_f\left(P_{Y,\hat{T}_{n,z}(X)} \,\|\, P_{YZ}\right) \overset{(30).2}{\le} \frac{2(r_n+u_n/\lambda)+\eta_n}{P_Y(A)}, \tag{30}$$

where (30).1 is justified in Lemma 26, and (30).2 follows from the bound obtained in (26), along with an application of the variational gap assumption **(INN4)** that relates $L_{\mathbf{z}}(\hat{T}_n)$ to $D_f\big(P_{Y,\hat{T}_{n,z}(X)} \,\|\, P_{YZ}\big)$.

Next, we use the Pinsker-type inequality to obtain

$$\Delta_A \leq c_f \sqrt{\frac{1}{2} D_f\left(P^{(A)}_{Y,\hat{T}_{n,z}(X)} \,\|\, P^{(A)}_{YZ}\right)} \leq c_f \sqrt{\frac{(2r_n + 2u_n/\lambda + \eta_n)}{2P_Y(A)}}.$$

Plugging this into (29), we get

$$W_1(\gamma,\omega) \leq C_a 2^{\frac{2-a}{2(1+a)}} \left(P_Y(A)\right)^{-\frac{2+a}{2(1+a)}} c_f^{\frac{a}{1+a}} (2r_n + 2u_n/\lambda + \eta_n)^{\frac{a}{2(1+a)}} \tag{31}$$

Combining (31) and (28) with (27) gives us the stated upper bound:

$$W_1(P_X^{(A)}, P^{(A)}_{\hat{T}_n^{-1}(Y,Z)}) \leq J\left(\sqrt{\frac{2(u_n + \lambda r_n)}{P_Y(A)}} + K_a (P_Y(A))^{-\frac{2+a}{2(1+a)}} (2r_n + 2u_n/\lambda + \eta_n)^{\frac{a}{2(1+a)}}\right), \tag{32}$$

where $K_a = 2^{\frac{4+a}{2+2a}} \left(a^{\frac{1}{1+a}} + a^{\frac{-a}{1+a}}\right) c_f^{\frac{a}{1+a}}$. For any finite $a > 0$, the dominant term is $\lesssim (r_n + u_n + \eta_n)^{\frac{a}{2(1+a)}}$, as claimed in the statement of Theorem 1. This completes the proof.

C.2 Proof of Proposition 6

C.2.1 Proof of (7)

The uniform convergence of the $\widehat{L}_{\mathbf{z},n}(T)$ follows from the boundedness of the function classes involved, along with some standard symmetrization and concentration techniques. In particular, we can show that with probability at least $1 - \delta/2$, we have

$$\sup_{T \in \mathcal{T}} |\widehat{L}_{\mathbf{z},n}(T) - L_{\mathbf{z}}(T)| \leq 2\mathfrak{R}_n^{(1)} + 2\mathfrak{R}_n^{(2)} + (b_n + A_{2,n})\sqrt{\frac{2\log(4/\delta)}{n}}. \tag{33}$$

We first introduce some notation to simplify the expressions. For any T and g, define

$$\alpha(g) = \frac{1}{n}\sum_{i=1}^n g(Y_i, Z_i) - \mathbb{E}[g(Y,Z)], \quad \psi(T,g) = \frac{1}{n}\sum_{i=1}^n f^*\left(g(Y_i, T_{\mathbf{z}}(X_i))\right) - \mathbb{E}[f^*(g(Y, T_{\mathbf{z}}(X)))],$$

and $\quad s(T,g) = \mathbb{E}[g(Y,Z)] - \mathbb{E}[f^*\left(g(Y, T_{\mathbf{z}}(X))\right)]$.

With these terms, we can write

$$\widehat{L}_{\mathbf{z},n}(T) = \sup_{g \in \mathcal{G}_n} \big[s(T,g) + \alpha(g) - \psi(T,g)\big], \quad \text{and} \quad L_{\mathbf{z}}(T) = \sup_{g \in \mathcal{G}_n} s(T,g).$$

This leads to the following chain:

$$\begin{aligned}
\sup_{T \in \mathcal{T}} \left|\widehat{L}_{\mathbf{z},n}(T) - L_{\mathbf{z}}(T)\right| &= \sup_{T \in \mathcal{T}} \left|\sup_{g \in \mathcal{G}_n} \left(s(T,g) + \alpha(g) - \psi(T,g)\right) - \sup_{g \in \mathcal{G}_n} s(T,g)\right| \\
&\leq \sup_{T \in \mathcal{T}} \sup_{g \in \mathcal{G}_n} |\alpha(g) + \psi(T,g)| \\
&\leq \sup_{g \in \mathcal{G}_n} |\alpha(g)| + \sup_{T \in \mathcal{T}} \sup_{g \in \mathcal{G}_n} |\psi(T,g)|\,.
\end{aligned} \tag{34}$$

It remains to control the two terms in (34). For the first term, standard symmetrization arguments imply

$$\mathbb{E}[\sup_{g\in\mathcal{G}_n}|\alpha(g)|] \le 2\mathbb{E}_{Y^n,Z^n,\epsilon^n}\left[\sup_{g\in\mathcal{G}_n}\frac{1}{n}\sum_{i=1}^n \epsilon_i g(Y_i,Z_i)\right] = 2\mathfrak{R}_n^{(1)}(\mathcal{G}_n).$$

Finally, the uniform boundedness assumption on $\mathcal{G}_n$ implies that the random variable $\sup_{g\in\mathcal{G}_n}\frac{1}{n}\sum_{i=1}^n \epsilon_i g(Y_i,Z_i)$ satisfies the bounded difference property. Hence, an application of McDiarmid's inequality leads to

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}_n}|\alpha(\alpha)| \ge 2\mathfrak{R}_n^{(1)}(\mathcal{G}_n) + b_n\sqrt{\frac{2\log(8/\delta)}{n}}\right) \le \frac{\delta}{4}. \tag{35}$$

Next, to bound the second term in (34), we use the shorthand $m_{T,g}(x,y) = f^*(g(y,T_{\mathbf{z}}(x))$, and let $\mathcal{M}_n = \{h_{T,g} : g\in\mathcal{G}_n,\ T\in\mathcal{T}\}$ denote the associated function, and obtain the following (dropping the subscript from $m_{T,g}$):

$$\begin{aligned}\mathbb{E}[\sup_{T,g}|\psi(T,g)|] &= \mathbb{E}\left[\sup_{m\in\mathcal{M}_n}\left|\frac{1}{n}\sum_{i=1}^n m(X_i,Y_i) - \mathbb{E}[m(X,Y)]\right|\right]\\ &\le 2\mathbb{E}_{X^n,Y^n,\epsilon^n}\left[\sup_{m\in\mathcal{M}_n}\frac{1}{n}\sum_{i=1}^n \epsilon_i m(X_i,Y_i)\right],\end{aligned}$$

by another application of the symmetrization technique. Next, observe that by definition, $u\mapsto f^*(u)$ is Lipschitz with constant $A_{1,n}$, and thus using the vector contraction lemma (Fact A.1), we obtain

$$\mathbb{E}_{X^n,Y^n,\epsilon^n}\left[\sup_{m\in\mathcal{M}_n}\frac{1}{n}\sum_{i=1}^n \epsilon_i m(X_i,Y_i)\right] \le 2A_{1,n}\mathbb{E}_{X^n,Y^n,\epsilon^n}\left[\sup_{T,g}\frac{1}{n}\sum_{i=1}^n \epsilon_i g(Y_i,T_{\mathbf{z}}(X_i))\right] = 2\mathfrak{R}_n^{(2)}(\mathcal{G}_n,\mathcal{T}).$$

Finally, observing that each $|f^*(g(Y_i,T_{\mathbf{z}}(X_i)))| \le A_{2,n}$ by assumption, another application of McDiarmid's bounded difference inequality gives us the required concentration result:

$$\mathbb{P}\left(\sup_{\substack{g\in\mathcal{G}_n\\ T\in\mathcal{T}}}|\psi_{T,g}| \ge 4A_{1,n}\mathfrak{R}_n^{(2)}(\mathcal{G}_n,\mathcal{T}) + A_{2,n}\sqrt{\frac{2\log(8/\delta)}{n}}\right) \le \frac{\delta}{4}. \tag{36}$$

Combining (35) and (36) give us the required $1-\delta/2$ probability bound claimed in (7).

C.2.2 Proof of (8)

The proof of (8) requires a careful truncation idea, as the squared loss function can be unbounded. We begin with the simple decomposition for some $K_n > 0$:

$$\begin{aligned}\sup_T|\widehat{L}_{\mathbf{y},n}(T) - L_{\mathbf{y}}(T)| = &\underbrace{\sup_T|\widehat{L}^{K_n}_{\mathbf{y},n}(T) - L^{K_n}_{\mathbf{y}}(T)|}_{:=\text{Term1}} + \underbrace{\sup_T|\widehat{L}_{\mathbf{y},n}(T) - \widehat{L}^{K_n}_{\mathbf{y},n}(T)|}_{:=\text{Term2}}\\ &+ \underbrace{\sup_T|L^{K_n}_{\mathbf{y}}(T) - L_{\mathbf{y}}(T)|}_{:=\text{Term3}}.\end{aligned} \tag{37}$$

We will bound the three terms in (37) separately.

Bound on Term1 **in** (37)**.** Let us introduce the function $\phi_y(u) = \|\Pi_{K_n}(u) - \Pi_{K_n}(y)\|^2$, and its centered version $\bar{\phi}_y(u) = \phi_y(u) - \phi_y(0)$. Then, observe the following:

$$\begin{aligned}\mathbb{E}_{\epsilon^n}\left[\sup_{T\in\mathcal{T}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\phi_{Y_i}(T_{\mathbf{y}}(X_i))\right] &= \mathbb{E}_{\epsilon^n}\left[\sup_{T}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\bar{\phi}_{Y_i}(T_{\mathbf{y}}(X_i)) + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i\phi_{Y_i}(0)\right]\\ &= \mathbb{E}_{\epsilon^n}\left[\sup_{T}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\bar{\phi}_{Y_i}(T_{\mathbf{y}}(X_i))\right] + \mathbb{E}_{\epsilon^n}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\phi_{Y_i}(0)\right]\\ &= \mathbb{E}_{\epsilon^n}\left[\sup_{T}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\bar{\phi}_{Y_i}(T_{\mathbf{y}}(X_i))\right], \qquad (38)\end{aligned}$$

since each Rademacher random variable ϵ_i is independent of Y_i. Thus, we can conclude via the standard symmetrization argument that

$$\mathbb{E}\left[\text{Term1}\right] \leq 2\mathbb{E}\left[\sup_{T\in\mathcal{T}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\phi_{Y_i}(T_{\mathbf{y}}(X_i))\right] \leq 2\mathbb{E}\left[\sup_{T\in\mathcal{T}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\bar{\phi}_{Y_i}(T_{\mathbf{y}}(X_i))\right].$$

Now, we can verify that the functions $\phi_y(u)$ are $4K_n$ Lipschitz globally, which implies that we can apply the (scalar) contraction lemma to get

$$\mathbb{E}_{\epsilon^n}\left[\sup_{T\in\mathcal{T}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\phi_{Y_i}(T_{\mathbf{y}}(X_i))\right] \leq 4K_n\mathbb{E}_{\epsilon^n}\left[\sup_{\substack{T\in\mathcal{T}\\ \|u_i\|\leq 1}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\langle u_i, T_{\mathbf{y}}(X_i)\rangle\right] = 4K_n\mathfrak{R}_n(\mathcal{F}_{\mathbf{y}}).$$

Plugging this into (38), we get

$$\mathbb{E}[\text{Term1}] \leq 8K_n\mathfrak{R}_n(\mathcal{F}_{\mathbf{y}}).$$

We can translate this expectation result into a high probability deviation bound via the standard bounded difference concentration inequality. In particular, by construction, the functions involved in Term1 satisfy a bounded differences property with parameter $2K_n$, which implies the following

$$\mathbb{P}\left(\text{Term1} \geq 8K_n\mathfrak{R}_n(\mathcal{F}_{\mathbf{y}}) + 4K_n^2\sqrt{2\log\left(\tfrac{8}{\delta}\right)n}\right) \leq \frac{\delta}{4}, \qquad (39)$$

by an application of Mcdiarmid's inequality.

Bound on Term2 **in** (37)**.** From the definitions of $\widehat{L}^{K_n}_{\mathbf{y},n}$ and $\widehat{L}_{\mathbf{y},n}$, we observe that

$$\begin{aligned}\widehat{L}_{\mathbf{y},n} - \widehat{L}^{K_n}_{\mathbf{y},n} &\leq \frac{1}{n}\sum_{i=1}^{n}\|T_{\mathbf{y}}(X_i) - Y_i\|^2\mathbf{1}_{\|T_{\mathbf{y}}(X_i)\|\vee\|Y_i\|>K_n}\\ &\leq \frac{1}{n}\sum_{i=1}^{n}2\|T_{\mathbf{y}}(X_i)\|^2\mathbf{1}_{\|T_{\mathbf{y}}(X_i)\|>K_n} + 2\|Y_i\|^2\mathbf{1}_{\|Y_i\|>K_n}\\ &\leq \frac{2}{n}\sum_{i=1}^{n}\frac{\|T_{\mathbf{y}}(X_i)\|^{2+\beta}}{K_n^{\beta}} + \frac{\|Y_i\|^{2+\beta}}{K_n^{\beta}},\end{aligned}$$

where the last inequality uses the fact that $\mathbf{1}_{A>B} \leq (A/B)^\beta$. Thus, the assumption on the $(2+\beta)$ moment of $T_{\mathbf{y}}(X_i)$ and Y_i together imply that

$$\mathbb{E}[\sup_T |\widehat{L}_{\mathbf{y},n}(T) - \widehat{L}^{K_n}_{\mathbf{y},n}(T)|] \leq \frac{4R_{\mathbf{y}}}{K_n^\beta}. \tag{40}$$

For a given $\delta > 0$, we then apply Markov's inequality to conclude that

$$\mathbb{P}\left(\sup_T |\widehat{L}_{\mathbf{y},n}(T) - \widehat{L}^{K_n}_{\mathbf{y},n}(T)| > \frac{16R_{\mathbf{y}}}{K_n^\beta \delta}\right) \leq \frac{\delta}{4}. \tag{41}$$

Bound on Term3 **in** (37)**.** The same argument that we used to obtain (40) is also applicable to this term, and we can conclude that

$$\begin{aligned}\sup_{T\in\mathcal{T}} |L_{\mathbf{y}}(T) - L^{K_n}_{\mathbf{y}}(T)| &\leq \sup_T \mathbb{E}[\|T_{\mathbf{y}}(X) - Y\|^2 \mathbf{1}_{\|T_{\mathbf{y}}(X)\|^2 \vee \|Y\|^2 > K_n}] \\ &\leq \sup_T 2\left(\mathbb{E}[\|T_{\mathbf{y}}(X)\|^2 \mathbf{1}_{\|T_{\mathbf{y}}(X)\|^2 > K_n}] + \mathbb{E}[\|Y\|^2 \mathbf{1}_{\|Y\|^2 > K_n}]\right) \leq \frac{4R_{\mathbf{y}}}{K_n^\beta}.\end{aligned} \tag{42}$$

Combining (39), (41), and (42) with (37) and (33), we get the required result.

C.3 Proof of Proposition 8

The general outline of the proof is similar to that of Proposition 18. For any $T \in \mathcal{T}$, let γ_T denote $D_f(P \parallel Q_T)$, and observe that for any $g \in \mathcal{G}_n$, we have

$$\begin{aligned}\Delta(T) &= D_f(P \parallel Q_T) - \sup_{g\in\mathcal{G}_n} \{\mathbb{E}_P[g] - \mathbb{E}_{Q_T}[f^*(g)]\} \\ &= (\mathbb{E}_P[g_T^*] - \mathbb{E}_{Q_T}[f^*(g_T^*)]) - \sup_{g\in\mathcal{G}_n} \{\mathbb{E}_P[g] - \mathbb{E}_{Q_T}[f^*(g)]\} \\ &\leq (\mathbb{E}_P[g_T^*] - \mathbb{E}_{Q_T}[f^*(g_T^*)]) - (\mathbb{E}_P[g] - \mathbb{E}_{Q_T}[f^*(g)])\,.\end{aligned}$$

Next, we subtract and add the clipped version of g_T^*, denoted by $\bar{g}_T$, to get

$$\begin{aligned}\Delta(T) &\leq (\mathbb{E}_P[g_T^* - \bar{g}_T] + \mathbb{E}_P[\bar{g}_T]) - (\mathbb{E}_{Q_T}[f^*(g_T^*) - f^*(\bar{g}_T)] + \mathbb{E}_{Q_T}[f^*(\bar{g}_T)]) - (\mathbb{E}_P[g] - \mathbb{E}_{Q_T}[f^*(g)]) \\ &= (\mathbb{E}_P[g_T^* - \bar{g}_T] + \mathbb{E}_{Q_T}[f^*(\bar{g}_T) - f^*(g_T^*)]) + (\mathbb{E}_P[\bar{g}_T - g] + \mathbb{E}_{Q_T}[f^*(g) - f^*(\bar{g}_T)]) \\ &\coloneqq C_1(T) + C_2(T, g).\end{aligned}$$

The first term $C_1(T)$ is the "clipping error" and satisfies the inequality

$$C_1(T) \leq \mathbb{E}_P[(|g_T^*| - b_n)^+] + A_{1,n}\mathbb{E}_{Q_T}[\big(|g_T^*| - b_n)^+\big)].$$

In the above display, we have used the fact that $|g_T^* - \bar{g}_T| = (|g_T^*| - b_n)^+$ and that f^* is $A_{1,n}$ Lipschitz on $[-b_n, b_n]$.

To analyze the term $C_2(T, g)$, we will consider its behavior inside and outside the ball B_{K_n}, which we denote by $C_{21} \equiv C_{2,1}(T, g)$ and $C_{22} \equiv C_{22}(T, g)$ respectively.

$$\begin{aligned}C_{21} &= \mathbb{E}_P[(\bar{g}_T - g)\mathbf{1}_{B_{K_n}}] + \mathbb{E}_{Q_T}[\big(f^*(g) - f^*(\bar{g}_T)\big)\mathbf{1}_{B_{K_n}}], \\ \text{and} \quad C_{22} &= \mathbb{E}_P[(\bar{g}_T - g)\mathbf{1}_{B^c_{K_n}}] + \mathbb{E}_{Q_T}[\big(f^*(g) - f^*(\bar{g}_T)\big)\mathbf{1}_{B^c_{K_n}}].\end{aligned}$$

To bound the term C_{21}, we again appeal to the fact that f^* is $A_{1,n}$ Lipschitz on $[-b_n, b_n]$, and hence we have

$$|C_{21}| \leq \mathbb{E}_P[|\bar{g}_T - g|\mathbf{1}_{B_{K_n}}] + A_{1,n}\mathbb{E}_{Q_T}[|\bar{g}_T - g|\mathbf{1}_{B_{K_n}}].$$

By definition, we have $|\bar{g}_T| \leq b_n$ and $|f^*(\cdot)| \leq A_{2,n}$, and also $|g| \leq b_n$ due to the uniform boundedness of $\mathcal{G}_n$. Together these facts imply

$$|C_{22}| \leq b_n \mathbb{P}_P\left(B_{K_n}^c\right) + A_{2,n}\mathbb{P}_{Q_T}\left(B_{K_n}^c\right).$$

Combining these, we get for any $T \in \mathcal{T}$, and $g \in \mathcal{G}_n$:

$$\Delta(T) \leq |C_1(T)| + \mathbb{E}_P[|\bar{g}_T - g|\mathbf{1}_{B_{K_n}}] + A_{1,n}\mathbb{E}_{Q_T}[|\bar{g}_T - g|\mathbf{1}_{B_{K_n}}] + b_n\mathbb{P}_P\left(B_{K_n}^c\right) + A_{2,n}\mathbb{P}_{Q_T}\left(B_{K_n}^c\right).$$

Taking the infimum over all $g \in \mathcal{G}_n$, and supremum over all $T \in \mathcal{T}$, we get

$$\eta_n \leq \tau_2(b_n) + (1 + A_{1,n})\delta_n(b_n, K_n) + (b_n + A_{2,n})\tau_1(K_n).$$

Hence, a sufficient condition for this to converge to zero is if $\tau_2 + A_{1,n}\delta_n + (b_n + A_{2,n})\tau_1$ converges to 0. This concludes the proof.

C.4 Proof of Theorem 12

The proof of this result follows closely the proof of Theorem 14 as given in Appendix B.3. Since we use the same model class $\mathcal{T}$, we can follow the exact argument we used in Appendix B.6.2 to show that

$$\sup_{T\in\mathcal{T}}\left\{\mathrm{Lip}(T),\, \mathrm{Lip}(T^{-1})\right\} \leq J \coloneqq \max\left\{(1+s^2)^M,\, (1-s^2)^{-M}\right\},$$

$$\text{and} \quad \sup_{T\in\mathcal{T}}\|T(\mathbf{0})\|_2 \leq J_0 \coloneqq \left(s\sqrt{H} + B\right)\frac{(1+s^2)^M - 1}{s^2}.$$

Recall that the terms s, B, H, and M were introduced in Definition 9. Additionally, since our latent variable Z is a multivariate Gaussian, it follows from the realizability assumption that the conditions of Proposition 7 hold with all $\beta > 0$. To complete the proof, it remains to verify that Proposition 6 and Proposition 8 hold.

C.4.1 Verification of Proposition 6

To verify this result, we need to control the supervised and unsupervised loss terms, and in particular, show the existence of $u_n, r_n \to 0$ such that

$$\sup_{T\in\mathcal{T}}|\widehat{L}_{\mathbf{y},n}(T) - L_{\mathbf{y}}(T)| \leq u_n, \quad \text{and} \quad \sup_{T\in\mathcal{T}}|\widehat{L}_{\mathbf{z},n}(T) - L_{\mathbf{z}}(T)| \leq r_n.$$

Proposition 6 implies that it suffices to check the uniform control over $2+\beta$ moment of $T_{\mathbf{y}}(X)$ over all $T \in \mathcal{T}$, and that the Rademacher complexities $\mathfrak{R}_n(\mathcal{F}_{\mathbf{y}}), \mathfrak{R}_n^{(1)}(\mathcal{G}_n)$, and $\mathfrak{R}_n^{(2)}(\mathcal{G}_n, \mathcal{T})$ vanish with n.

Finite $2+\beta$ moment. From the global Lipschitz property of T, there exist finite constants J_0, J such that for all $T \equiv (T_{\mathbf{y}}, T_{\mathbf{z}}) \in \mathcal{T}$, we have

$$\|T_{\mathbf{y}}(x)\| \leq J_0 + J\|x\|, \quad \text{for all} \quad x \in \mathcal{X}.$$

Hence, for any $\beta > 0$, there exists a constant C_β depending on J_0, J and β, such that

$$\|T_{\mathbf{y}}(X)\|^{2+\beta} \leq C_\beta \left(1 + \|X\|^{2+\beta}\right) \quad \Longrightarrow \quad \sup_{T\in\mathcal{T}} \mathbb{E}\|T_{\mathbf{y}}(X)\|^{2+\beta} \leq C_\beta \left(1 + \mathbb{E}[\|X\|^{2+\beta}]\right) < \infty,$$

using the finite $2+\beta$ moment assumption on X.

Rademacher complexity $\mathfrak{R}_n(\mathcal{F}_{\mathbf{y}})$. For any $T \in \mathcal{T}$ and $u : \|u\| \leq 1$, observe that

$$\mathbb{E}\left[\sup_{\substack{T\in\mathcal{T}\\ u:\|u\|\leq 1}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \langle u, T_{\mathbf{y}}(X_i)\rangle\right] \leq 2J\mathbb{E}\left[\sup_{v:\|v\|\leq 1} \frac{1}{n}\sum_{i=1}^n \epsilon_i\langle v, X_i\rangle\right] \leq 2J \sup_{v:\|v\|=1} \|v\|\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_i\right\|\right],$$

where the first inequality follows from the contraction lemma, and the fact that the map $x \mapsto \langle u, T_{\mathbf{y}}(x)\rangle$ is J-Lipschitz, and the second inequality follows from an application of Cauchy-Schwarz. Next, by using Jensen's inequality and the concavity of the map $a \mapsto \sqrt{a}$, we have

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_i\right\|\right] = \mathbb{E}\left[\sqrt{\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_i\right\|^2}\right] \leq \sqrt{\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_i\right\|^2\right]} = \sqrt{\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \epsilon_i\epsilon_j\langle X_i, X_j\rangle\right]}.$$

Since $\epsilon_i \perp \epsilon_j$ for $i \neq j$, we get the bound

$$\mathfrak{R}_n(\mathcal{F}_{\mathbf{y}}) \leq 2J\sqrt{\frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[\|X\|^2]} = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right),$$

which converges to 0 with n.

Rademacher Complexities $\mathfrak{R}_n^{(1)}(\mathcal{G}_n)$ and $\mathfrak{R}_n^{(2)}(\mathcal{G}_n, \mathcal{T})$. For the first term, observe that with $U_i = (Y_i, Z_i)$ and $w_i = \mathbf{1}_{[-K_n,K_n]^{d_{\mathbf{x}}}}(U_i)$ and $g(U_i) = h(U_i)w_i$ for some $h \in \mathcal{H}_k$:

$$\mathfrak{R}_n^{(1)}(\mathcal{G}_n) = \mathbb{E}_{\epsilon^n, U^n}\left[\sup_{h\in\mathcal{H}_k:\|h\|_k\leq b_n} \frac{1}{n}\sum_{i=1}^n \epsilon_i\langle h, k(U_i,\cdot)\rangle_k w_i\right] \leq \frac{b_n}{n}\mathbb{E}_{U^n}\left[\sqrt{\sum_{i=1}^n w_i^2 k(U_i,U_i)}\right] \leq \frac{b_n}{\sqrt{n}},$$

where the first inequality uses a standard argument for RKHSs presented in details in Appendix B.6.1.

Next, for any $T \in \mathcal{T}$ and $g \in \mathcal{G}_n$, introduce the terms $\widetilde{U}_i = (Y_i, T_z(X_i))$ and $\widetilde{w}_i = \mathbf{1}_{[-K_n,K_n]^{d_{\mathbf{x}}}}(\widetilde{U}_i)$, and observe that

$$\mathfrak{R}_n^{(2)}(\mathcal{G}_n, \mathcal{T}) \leq \mathbb{E}_{\epsilon^n, \widetilde{U}^n}\left[\sup_{T,g} \frac{1}{n}\sum_{i=1}^n \epsilon_i \circ g(\widetilde{U}_i)\right] \lesssim \frac{A_{1,n}b_n}{\sqrt{n}},$$

where we used the (scalar) contraction lemma for Rademacher complexities.

C.4.2 Verification of Proposition 8

We now obtain the terms τ_1, τ_2 and δ_n from Proposition 8 for our specific example. Recall that we are using the notation $P \equiv P_{Y,Z}$ and $Q_T \equiv P_{Y,T_z(X)}$

First, we consider the term τ_1, which is equal to $\sup_{T\in\mathcal{T}} Q_T(B^c_{K_n}) + P(B^c_{K_n}) = \sup_{T\in\mathcal{T}} Q_T(\{u : \|u\| > K_n\}) + P(\{u : \|u\| > K_n\})$. Equivalently, with the random vector $U_T = (Y, T_z(X))$ for some $T \in \mathcal{T}$, we can handle $Q_T(\{u : \|u\| > K_n\}) = \mathbb{P}(\|U_T\| > K_n)$ simply by using the Markov's inequality as

$$\mathbb{P}(\|U_T\| > K_n) \leq \frac{\mathbb{E}[\|U_T\|^{2+\beta}]}{K_n^{2+\beta}} \leq \frac{R}{K_n^{2+\beta}},$$

where the term R follows from the uniform moment condition that we verified using Proposition 7. An exact same argument also works for the term $P(B^c_{K_n})$. Hence, τ_1 converges to 0 as long as $K_n \to \infty$ with n.

Next, we look at the term τ_2 which is defined as

$$\tau_2 = \sup_{T\in\mathcal{T}} \left\{\mathbb{E}_P[(|g^*_T| - b_n)^+] + A_{1,n}\mathbb{E}_{Q_T}\left[(|g^*_T| - b_n)^+\right]\right\},$$

where g^*_T is the optimal (unconstrained) critic function (or witness function) in the definition of JS-divergence. A standard calculation shows that

$$g^*_T = \log\left(\frac{2v_T}{1+v_T}\right), \quad \text{where} \quad v_T = \frac{dP}{dQ_T}.$$

Next, observe that we g^*_T is uniformly bounded from above by $\log 2$, and we can use the simple inequality

$$|g^*_T| = |\log 2 + \ell_T - \log(1+v_T)| \leq \log 2 + |\ell_T|, \quad \text{where} \quad \ell_T = \log v_T = \log\frac{dP}{dQ_T}.$$

As a result, for any $b_n > \log 2$, we have

$$(|g^*_T| - b_n)^+ \leq (|\ell_T| - (b_n - \log 2))^+ \leq |\ell_T|\mathbf{1}_{|\ell_T|>b_n-\log 2}.$$

Now, following the exact argument for the NF case, we know from Lemma 23, that there exist constants c_0, c_1, c_2 independent of n, T, such that we have

$$|\ell_T(x)| \leq c_0 + c_1\|x\| + c_2\|x\|^2 \quad \text{for all} \quad x \in \mathbb{R}^d.$$

If the distribution P_X is such that $\mathbb{E}_{P_X}[\|X\|^2] < \infty$, then, this implies that

$$\mathbb{E}_{P_X}[|\ell_T|\mathbf{1}_{|\ell_T|>b_n-\log 2}] \leq \mathbb{E}_{P_X}[c_0 + c_1\|X\| + c_2\|X\|^2 : \|X\|_\infty > b_n - \log 2] \overset{n\to\infty}{\longrightarrow} 0.$$

Thus, the only condition needed to control the above term is that $b_n \to \infty$, and that $\mathbb{E}_{P_X}[\|X\|^2] < \infty$. The same argument also works for $\mathbb{E}_{Q_T}[|\ell_T|\mathbf{1}_{|\ell_T|>b_n-\log 2}] \to 0$ due to the bounded $2+\beta$ moment condition.

It remains to consider the approximation error term on the ball $B_{K_n} = \{u \in \mathbb{R}^{d_{\mathbf{x}}} : \|u\| \leq K_n\}$, defined as

$$\begin{aligned}\delta_n \equiv \delta_n(b_n, K_n) &= \sup_T \inf_g \left\{\mathbb{E}_P\left[|\bar{g}_T - g|\mathbf{1}_{B_{K_n}}\right] + \mathbb{E}_{Q_T}\left[|\bar{g}_T - g|\mathbf{1}_{B_{K_n}}\right]\right\} \\ &\leq 2 \sup_T \inf_g \|\bar{g}_T - g\|_{\infty, C_{K_n}},\end{aligned}$$

where recall that $C_{K_n} = [-K_n, K_n]^{d_{\mathbf{x}}}$, $\bar{g}_T$ is the clipped version of g_T^* at level b_n, and $\|\bar{g}_T - g\|_{\infty, C_{K_n}}$ denotes the sup-norm over the cube C_{K_n}. To complete the proof, it suffices to argue that for every $T \in \mathcal{T}$, there exists a $g_{T,n} \in \mathcal{G}_n$, such that

$$\|\bar{g}_T - g_{T,n}\|_{\infty, C_{K_n}} = \|g_T^* - g_{T,n}\|_{\infty, C_{K_n}} \lesssim \frac{K_n}{\sqrt{\gamma_n}}.$$

This is ensured by Lemma 24.

C.4.3 Completing the proof

Recall that the statement of Theorem 12 makes the following choices of the various parameters:

- $K_n = M(s\sqrt{H} + B) + \sqrt{d_{\mathbf{x}}} + \sqrt{2\log n}$
- $\gamma_n = K_n^{2+\epsilon}$ for some $\epsilon > 0$
- $b_n = C_b \gamma_n^{d_{\mathbf{x}}/4} K_n^{2+d_{\mathbf{x}}/2}$
- $A_{1,n} \leq 1$ and $A_{2,n} \leq b_n$

Now, observe that from (8), these choices imply

$$\sup_{T \in \mathcal{T}} |\hat{L}_y(T) - L_{\mathbf{y}}(T)| \lesssim \sqrt{\frac{\log n}{n}} + \frac{\log n}{\sqrt{n}} + \frac{1}{(\log n)^{2/\beta}} \overset{n\to\infty}{\longrightarrow} 0$$

$$\sup_{T \in \mathcal{T}} |\hat{L}_z(T) - L_{\mathbf{z}}(T)| \lesssim \frac{b_n}{\sqrt{n}} + (b_n + A_{2,n})\sqrt{\frac{\log(1/\delta)}{n}} \lesssim \frac{(\log n)^{d_{\mathbf{x}}\left(\frac{1}{2} + \frac{\epsilon}{8}\right)+1}}{\sqrt{n}} \overset{n\to\infty}{\longrightarrow} 0.$$

This ensures that the conditions of Proposition 6 are satisfied by these parameters. To complete the proof, we will show that the terms τ_1, τ_2, and δ_n from Proposition 8 also converge to 0 for these parameters. As we showed in the previous section,

- $\tau_1 \lesssim \frac{1}{K_n^{2+\beta}}$ for some $\beta > 0$, and hence $\tau_1 \to 0$ with $n \to \infty$ since $K_n \asymp \sqrt{\log n}$.
- $\tau_2 \lesssim \mathbb{E}_P[\|X\|^2 : \|X\|_\infty > b_n - \log 2] + \mathbb{E}_{Q_T}[\|X\|^2 : \|X\|_\infty > b_n - \log 2]$ which converges to 0 as $\lim_{n\to\infty} b_n = \infty$.
- Finally, we also showed that

  $$\delta_n \lesssim \frac{K_n}{\sqrt{\gamma_n}} \asymp \frac{1}{K_n^{\epsilon/2}},$$

  which also vanishes as $n \to \infty$.

This completes the verification of all the conditions for the INN example introduced in Section 3.2.2.

Appendix D. Details of Technical Lemmas

D.1 A Truncation Argument

Lemma 25 *Suppose $U \sim \mu$ and $V \sim \nu$ denote two random variables taking values in $\mathbb{R}^j$ for some $j \geq 1$, and suppose that $\max\left\{\mathbb{E}[\|V\|^{1+a}], \mathbb{E}[\|U\|^{1+a}]\right\} = R < \infty$, for some $a > 0$. Then, we have the following:*

$$W_1(\mu,\nu) = \sup_{\substack{f:\mathbb{R}^j \to \mathbb{R} \\ \mathrm{Lip}(f)\leq 1}} \left| \int f d\mu - \int f d\nu \right| \lesssim \Delta^{\frac{a}{2(1+a)}},$$

where $\Delta = \|\mu - \nu\|_{TV}$ denotes the total variation distance between μ and ν.

Proof Let $M \in \mathbb{R}$ denote a finite positive real number to be specified later, and let us define the set $E = \{x \in \mathbb{R}^j : \|x\| \leq M\}$. Then, we have the following with $g = f - f(0)$:

$$\begin{aligned}
\int f d\mu - \int f d\nu &= \int f d(\mu - \nu) - f(0) \int d(\mu - \nu) = \int g d(\mu - \nu) \\
&= \int_E g d(\mu - \nu) + \int_{E^c} g d(\mu - \nu) \\
&\leq \int_E \|x\| d|\mu - \nu| + \left| \int_{E^c} g(x) d(\mu - \nu) \right|, \qquad (43)
\end{aligned}$$

where the last inequality uses the fact that $g = f - f(0)$ is 1-Lipschitz. The definition of E leads to the following natural upper bound on the first term

$$\int_E \|x\| d|\mu - \nu| \leq M \int_E d|\mu - \nu| \leq M \int_{\mathbb{R}^j} d|\mu - \nu| \leq 2M\Delta. \qquad (44)$$

The last inequality above uses the fact that $2\|\mu - \nu\|_{TV} = |\mu - \nu|(\mathbb{R}^d)$. Next, we consider the second term in (43), and observe that

$$\begin{aligned}
\left| \int_{E^c} g d(\mu - \nu) \right| &\leq \int_{E^c} |g| (d\mu + d\nu) \leq \int_{E^c} \|x\| d(\mu + \nu) = \mathbb{E}\left[\|U\| \mathbf{1}_{\|U\| > M}\right] + \mathbb{E}\left[\|V\| \mathbf{1}_{\|V\| > M}\right] \\
&\leq M^{-a} \mathbb{E}\left[\|U\|^{1+a} \mathbf{1}_{\|U\| > M}\right] + M^{-a} \mathbb{E}\left[\|V\|^{1+a} \mathbf{1}_{\|V\| > M}\right] \\
&\leq M^{-a} \mathbb{E}\left[\|U\|^{1+a}\right] + M^{-a} \mathbb{E}\left[\|V\|^{1+a}\right] \leq 2M^{-a} R. \qquad (45)
\end{aligned}$$

Combining the bounds in (44) and (45), we get

$$W_1(\mu,\nu) = \sup_g \left| \int g d(\mu - \nu) \right| \leq 2M\Delta + 2M^{-a} R. \qquad (46)$$

The function $M \mapsto 2M\Delta + 2M^{-a}R$ is convex in M, and is optimized at $M^* = (aR/\Delta)^{1/(1+a)}$. Plugging this value back in (46) gives us the required

$$W_1(\mu,\nu) \leq 2\left(a^{\frac{1}{1+a}} + a^{\frac{-a}{1+a}}\right) R^{\frac{1}{1+a}} \Delta^{\frac{a}{1+a}} := C_a R^{\frac{1}{1+a}} \Delta^{\frac{a}{1+a}}.$$

As a sanity check, note that if the supports of U and V are bounded and $a \to \infty$, we recover the bound that holds for bounded random variables; that is, $W_1(\mu,\nu) \lesssim \Delta$. ■

D.2 A Conditional f-Divergence Bound

Lemma 26 *Let P, Q denote two probability measures on $\mathcal{Y} \times \mathcal{Z}$, and for any measurable set $A \subset \mathcal{Y}$, let $P_Y(A) = P(A \times \mathcal{Z}) > 0$. Define the measures*

$$P^{(A)}(dy, dz) = \frac{1}{P_Y(A)} \mathbf{1}_{y \in A} P(dy, dz), \quad \text{and} \quad Q^{(A)}(dy, dz) = \frac{1}{P_Y(A)} \mathbf{1}_{y \in A} Q(dy, dz).$$

Then, we have the following: $D_f(P \parallel Q) \leq \frac{1}{P_Y(A)} D_f(P^{(A)} \parallel Q^{(A)})$. Note that in general, $Q^{(A)}$ may not be a probability measure, but the inequality still holds.

Proof Let μ denote a measure that dominates P, Q; for example, we may set $\mu = \frac{1}{2}(P+Q)$, and let p, q denote the densities of P, Q w.r.t. μ. It then follows that

$$\frac{dP^{(A)}}{d\mu}(y, z) = \frac{\mathbf{1}_{y \in A}}{P_Y(A)} p(y, z), \quad \text{and} \quad \frac{dQ^{(A)}}{d\mu}(y, z) = \frac{\mathbf{1}_{y \in A}}{P_Y(A)} q(y, z).$$

As a result, we obtain

$$\frac{dP^{(A)}}{dQ^{(A)}}(y, z) = \frac{p(y, z)}{q(y, z)} = \frac{dP}{dQ}(y, z) \quad \text{for all} \quad (y, z) \in A \times \mathcal{Z}.$$

Now, by the definition of f-divergences, we have

$$\begin{aligned} D_f(P^{(A)} \parallel Q^{(A)}) &= \int_{A \times \mathcal{Z}} f\left(\frac{dP^{(A)}}{dQ^{(A)}}(y, z)\right) Q^{(A)}(dy, dz) = \frac{1}{P_Y(A)} \int_{A \times \mathcal{Z}} f\left(\frac{p(y, z)}{q(y, z)}\right) q(y, z) \mu(dy, dz) \\ &\overset{(47).1}{\leq} \frac{1}{P_Y(A)} \int_{\mathcal{Y} \times \mathcal{Z}} f\left(\frac{p(y, z)}{q(y, z)}\right) q(y, z) \mu(dy, dz) = \frac{1}{P_Y(A)} D_f(P \parallel Q), \end{aligned} \tag{47}$$

where (47).1 uses the nonnegativity of f and q. This completes the proof. ■

D.3 Proof of Lemma 22

Consider any $j \in [M]$, and let $z_j(x) = W_{j,1} x + b_{j,1}$ and $D_j(x) = \operatorname{diag}(\tanh'(z_j(x))) = \operatorname{diag}(\operatorname{sech}^2(z_j(x)))$. Since $\operatorname{sech}^2(z) \in [0, 1]$, it follows that $\mathbf{0} \preccurlyeq D_j(x) \preccurlyeq I_{d_j}$. Then, we can verify that the Jacobians of F_j and G_j, denoted by $\mathbb{J}F_j$ and $\mathbb{J}G_j$ respectively, are equal to

$$\mathbb{J}F_j(x) = W_{j,2} D_j(x) W_{j,1}, \quad \text{and} \quad \mathbb{J}G_j(x) = I_d + W_{j,2} D_j(x) W_{j,1}.$$

By the assumption that $\|W_{j,i}\|_{op} \leq s \in (0, 1)$ for all $j \in [M]$, we obtain

$$\|\mathbb{J}F_j(x)\|_{op} \leq \|W_{j,2}\|_{op} \|D_j(x)\|_{op} \|W_{j,1}\|_{op} \leq s^2 = \rho.$$

Here we used the fact that $\|D_j(x)\|_{op} \leq \|I_{d_j}\|_{op} = 1$. As a result, all the singular values of $\mathbb{J}G_j(x) = (I_d + \mathbb{J}F_j)(x)$ are in the range $[1 - \rho, 1 + \rho]$. Now, by the chain rule of the flow, and with $x_0 = x$, $x_j = G_j x_{j-1}$ for $j \in [M]$, we have

$$\mathbb{J}T(x) = \prod_{j=M}^{1} \mathbb{J}G_j(x_{j-1}) \quad \Longrightarrow \quad \log|\det \mathbb{J}T(x)| = \sum_{j=1}^{M} \log|\det(I_d + W_{j,2} D_j(x_{j-1}) W_{j,1})|.$$

Since for every $j \in [M]$, the singular values of $\mathbb{J}G_j(x_{j-1})$ lie in $[1-\rho, 1+\rho]$, we obtain

$$Md\log(1-\rho) \leq \sum_{j=1}^{M} \log|\det \mathbb{J}G_j(x_{j-1})| \leq Md\log(1+\rho).$$

Note that the above result is uniform in the input x, and uniform over all $T \in \mathcal{T}$.

D.4 Proof of Lemma 23

Let $Z \sim N(0, I_d)$, and denote its density by $\phi(z) = (2\pi)^{-d/2}e^{-\|z\|^2/2}$. Then, for any diffeomorphism $T : \mathbb{R}^d \to \mathbb{R}^d$, we have the following by the change of variables theorem:

$$p_T(x) = \phi(T(x))\det \mathbb{J}T(x).$$

Since $\ell_T = \log p_{T^*}/p_T$, this implies

$$\ell_T(x) = \log\phi(T^*(x)) + \log|\det \mathbb{J}T^*(x)| - \log\phi(T(x)) - \log|\det \mathbb{J}T(x)|.$$

Since $\log\phi(z) = -(1/2)\|z\|^2 - d/2\log(2\pi)$, we obtain the following closed form expression for ℓ_T:

$$\ell_T(x) = \frac{1}{2}\left(\|T(x)\|^2 - \|T^*(x)\|^2\right) + \log|\det \mathbb{J}T^*(x)| - \log|\det \mathbb{J}T(x)|. \tag{48}$$

We now establish the uniform Lipschitz constant and the maximum value of $\{\ell_T : T \in \mathcal{T}\}$ when restricted to the cube $C_K = [-K, K]^d$ for some $K > 0$.

Lipschitz Constant. To derive the Lipschitz constant of ℓ_T, we look at its gradient at any $x \in [-K, K]^d$:

$$\nabla\ell_T(x) = \mathbb{J}T(x)^T T(x) - \mathbb{J}T^*(x)^T T^*(x) + \nabla\log\det \mathbb{J}T^*(x) - \nabla\log\det \mathbb{J}T(x).$$

For each coordinate i, we have

$$\partial_{x_i}\log\det \mathbb{J}T(x) = \operatorname{tr}\left((\mathbb{J}T(x))^{-1}\partial_{x_i}\mathbb{J}T(x)\right), \quad \text{and} \quad \partial_{x_i}\log\det \mathbb{J}T^*(x) = \operatorname{tr}\left((\mathbb{J}T^*(x))^{-1}\partial_{x_i}\mathbb{J}T^*(x)\right) \tag{49}$$

We have already proved in Appendix B.6.2 that

$$\|T(x)\| \leq L\sqrt{d}K + C_0 A, \quad \text{where} \quad L = (1+s^2)^M, \quad C_0 = s\sqrt{H} + B, \quad \text{and} \quad A = \frac{(1+s^2)^M - 1}{s^2}.$$

Hence, we get

$$\max\left\{\|\mathbb{J}T(x)^T T(x)\|,\ \|\mathbb{J}T^*(x)^T T^*(x)\|\right\} \leq L(L\sqrt{d}K + C_0 A).$$

We now need to analyze the norm of the gradient of the log-det terms in (49). We begin with the directional derivative along any direction v:

$$D_v\left(\log\det \mathbb{J}T(x)\right) = \operatorname{tr}\left((\mathbb{J}T(x))^{-1}D_v\mathbb{J}T(x)\right) \leq d\|\mathbb{J}T(x)^{-1}\|_{op}\|D_v\mathbb{J}T(x)\|_{op}.$$

For any residual block G_j, we know that

$$\|\mathbb{J}G_j(u)^{-1}\|_{op} \leq (1-s^2)^{-1}, \quad \text{and} \quad \|D_v \mathbb{J}G_j(u)\|_{op} \leq 2s^3\|v\|.$$

On chaining this for $T = G_M \circ \cdots \circ G_1$, we get

$$\|D_v \log\det \mathbb{J}T(x)\| \leq \sum_{j=1}^{M} \frac{d}{1-s^2} 2s^3 \|\mathbb{J}(G_M \circ \cdots \circ G_1(x)v\| \leq \frac{2ds^3}{1-s^2} A\|v\|.$$

Hence, we get the required bound

$$\max\left\{\|\nabla \log\det \mathbb{J}T(x)\|,\ \|\nabla \log\det \mathbb{J}T^*(x)\|\right\} \leq \frac{2ds^3}{1-s^2} A. \tag{50}$$

Together, (49) and (50) imply the following for $x, y \in [-K, K]^d$:

$$|\ell_T(x) - \ell_T(y)| \leq \left[2L(L\sqrt{d}K + C_0 A) + \frac{4ds^3}{1-s^2} A\right] \|x-y\| =: L_1\|x-y\|. \tag{51}$$

Note that the Lipschitz constant L_1 is independent of T, and hence is valid uniformly over the family $\mathcal{T}$.

Maximum value of ℓ_T on $[-K, K]^d$. From (48), we know that the value of ℓ_T at 0 is

$$\ell_T(0) = \frac{1}{2}\left(\|T(0)\|^2 - \|T^*(0)\|^2\right) + \log\det \mathbb{J}T^*(0) - \log\det \mathbb{J}T(0).$$

From Lemma 22, we know that $Md\log(1-s^2) \leq \log\det \mathbb{J}T(x) \leq Md\log(1+s^2)$, and from Appendix B.6.2, we know that $\|T(0)\| \leq C_0 A$. Combining these facts, we get the required bound

$$\sup_{T\in\mathcal{T}} |\ell_T(0)| \leq (C_0 A)^2 + 2Md\log(1/(1-s^2)) =: L_0. \tag{52}$$

This fact, combined with the (uniform) Lipschitz constant derived in (51), leads to the following uniform bound on the maximum value of ℓ_T:

$$\sup_{T\in\mathcal{T}} \sup_{x\in[-K,K]^d} |\ell_T(x)| \leq L_0 + L_1\sqrt{d}K.$$

This completes the proof.

D.5 Proof of Lemma 24

The high-level idea behind the proof is simple: given the function that we want to approximate (i.e., ℓ_T), we construct h_σ by convolving this function with a Gaussian kernel with a scale parameter σ^2 (to be chosen later). Under certain conditions on σ, we can show that this smoothed function lies in $\mathcal{H}_n$. Hence, getting a bound on δ_n reduces to that of evaluating the discrepancy between ℓ_T and its smoothed version h_σ, which can be obtained via standard arguments.

A compactly supported Lipschitz extension. The first step in our proof is to construct a proxy for ℓ_T that agrees with it on $C_{K_n} = [-K_n, K_n]^d$, and is Lipschitz on the entire domain $\mathbb{R}^d$. To do this, we define

$$\bar{\ell}_T(x) = \begin{cases} \ell_T(x), & \text{if } x \in C_{K_n} = [-K_n, K_n]^d, \\ \inf_{y \in C_{K_n}} \{\ell_T(y) + L\|x - y\|\}, & \text{if } x \notin C_{K_n}. \end{cases}$$

We also need to restrict this Lipschitz extension to a compactly supported domain. To do this introduce the function $\nu_1 : \mathbb{R} \to [0, 1]$, and $\nu_d : \mathbb{R}^d \to [0, 1]$ as

$$\nu_1(x) = \begin{cases} 1, & \text{if } |x| \leq K_n, \\ 1 - |x|/2K_n, & \text{if } |x| \in (K_n, 2K_n], \\ 0, & \text{otherwise,} \end{cases} \quad \text{and} \quad \nu_d(x) = \prod_{i=1}^{d} \nu_1(x_i),$$

where $x = (x_1, \ldots, x_n)$. With this, we introduce $f_\nu(x) = \bar{\ell}_T(x)\nu_d(x)$. and observe that

$$\begin{aligned} &\text{Lip}(f_\nu) \leq \text{Lip}(\bar{\ell}_T)\|\nu_d\|_\infty + \|\bar{\ell}_T\|_\infty \text{Lip}(\nu_d) \leq L_1 + \frac{d(L_0 + L_1 K_n)}{K_n} \lesssim K_n, \\ &\text{and} \quad \|f_\nu\|_{L^2} \leq \|f_\nu\|_\infty (4K_n)^{d/2} \lesssim K_n^{2+d/2}. \end{aligned} \tag{53}$$

The compact support property of f_ν was important for controlling its L^2 norm that will be used later.

Gaussian Convolution. Fix some $\sigma > 0$, and with $\varphi_\sigma(u) = (2\pi\sigma^2)^{-d/2} \exp(-\|u\|^2/2\sigma^2)$, define the function h_σ

$$h_\sigma(x) = (f_\nu * \varphi_\sigma)(x) = \int f_\nu(x - u)\varphi_\sigma(u)du.$$

Hence, for any $x \in C_{K_n} = [-K_n, K_n]^d$, we have

$$\begin{aligned} |f_\nu(x) - h_\sigma(x)| &= \left| \int f_\nu(x)\varphi_\sigma(u)du - \int f_\nu(x - u)\varphi_\sigma(u)du \right| \\ &\leq \int |f_\nu(x) - f_\nu(x - u)|\varphi_\sigma(u) \overset{(i)}{\leq} L_1 \int \|u\|\varphi_\sigma(u)du \\ &\overset{(ii)}{\leq} L_1 \left(\int \|u\|^2 \varphi_\sigma(u)du \right)^{1/2} = L_1\sqrt{d}\sigma \lesssim K_n\sigma. \end{aligned}$$

The inequality (i) uses the fact that the function f_ν is Lipschitz with constant $L_1 \lesssim K_n$ over the entire domain $\mathbb{R}^d$ (and not just over C_{K_n} like ℓ_T), and (ii) uses Jensen's inequality for the concave map $x \mapsto \sqrt{x}$.

RKHS norm of h_σ. We know that the symmetric Fourier transform of the Gaussian kernel is

$$\hat{k}_n(\omega) = \left(\frac{\pi}{\gamma_n} \right)^{d/2} e^{-\|\omega\|^2/4\gamma_n}.$$

Hence, the RKHS norm of any h with $\hat{h} \in L^2$ is

$$\begin{aligned}\|h\|_{k_n}^2 &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \frac{|\hat{h}(\omega)|^2}{|\hat{k}_n(\omega)|} d\omega = \left(\frac{\gamma_n}{\pi}\right)^{d/2} \frac{1}{(2\pi)^{d/2}} \int |\hat{h}(\omega)|^2 e^{\|\omega\|^2/4\gamma_n} d\omega \\ &\lesssim \gamma_n^{d/2} \int |\hat{h}(\omega)|^2 e^{\|\omega\|^2/4\gamma_n} d\omega\end{aligned}$$

Now, returning to h_σ, we know that $\hat{h}_\sigma = \hat{f}_\nu e^{-\sigma^2\|\omega\|^2/2}$, which implies that if $\sigma^2 \geq 1/4\gamma_n$, we have by Plancherel's theorem,

$$\|h_\sigma\|_{k_n}^2 \lesssim \gamma_n^{d/2} \int |\hat{f}_\nu(\omega)|^2 e^{-\|\omega\|^2(\sigma^2 - 1/4\gamma_n)} d\omega \leq \gamma_n^{d/2} \int |\hat{f}_\nu(\omega)|^2 d\omega = \gamma_n^{d/2} \|f_\nu\|_{L^2}^2 \lesssim \gamma_n^{d/2} K_n^{4+d},$$

where the last inequality uses the L^2 norm bound on f_ν obtained in (53). Hence, $h_\sigma \mathbf{1}_{[-K_n, K_n]^d}$ with $\sigma = 1/2\sqrt{\gamma_n}$ lies in our function class $\mathcal{H}_n$, and hence reduces our approximation bound to

$$\sup_{x \in [-K_n, K_n]^d} |f_\nu(x) - h_\sigma(x)| \lesssim \frac{K_n}{\sqrt{\gamma_n}}.$$

This completes the proof.