Deep Regression Representation Learning with Topology

Shihao Zhang [1]   **Kenji Kawaguchi** [1]   **Angela Yao** [1]

Abstract

Most works studying representation learning focus only on classification and neglect regression. Yet, the learning objectives and, therefore, the representation topologies of the two tasks are fundamentally different: classification targets class separation, leading to disconnected representations, whereas regression requires ordinality with respect to the target, leading to continuous representations. We thus wonder how the effectiveness of a regression representation is influenced by its topology, with evaluation based on the Information Bottleneck (IB) principle. The IB principle is an important framework that provides principles for learning effective representations. We establish two connections between it and the topology of regression representations. The first connection reveals that a lower intrinsic dimension of the feature space implies a reduced complexity of the representation $\mathbf{Z}$. This complexity can be quantified as the conditional entropy of $\mathbf{Z}$ on the target $\mathbf{Y}$, and serves as an upper bound on the generalization error. The second connection suggests a feature space that is topologically similar to the target space will better align with the IB principle. Based on these two connections, we introduce PH-Reg, a regularizer specific to regression that matches the intrinsic dimension and topology of the feature space with the target space. Experiments on synthetic and real-world regression tasks demonstrate the benefits of PH-Reg. Code: https://github.com/needylove/PH-Reg.

1. Introduction

Regression is a fundamental task in machine learning in which input samples are mapped to a continuous target

space. Representation learning empowers models to automatically extract, transform, and leverage relevant information from data, leading to improved performance. The information bottleneck (IB) principle (Shwartz-Ziv & Tishby, 2017) provides a theoretical framework and guiding principle for learning effectiveness representations. The IB principle suggests learning a representation $\mathbf{Z}$ with sufficient information about the target $\mathbf{Y}$ but minimal information about the input $\mathbf{X}$. For representation $\mathbf{Z}$, sufficiency keeps all necessary information on $\mathbf{Y}$, while the minimality reduces $\mathbf{Z}$'s complexity and prevents overfitting. The optimal representation, as specified by Achille & Soatto (2018b;a), is the most useful (sufficient) and minimal. Yet, they specified classification and neglect regression.

The IB principle is applicable to both classification and regression in that both learn minimal and sufficient representations. However, there are some fundamental differences. For example, classification shortens the distance between features belonging to the same class while elongating the distance between features of different classes; the shortening and elongating of distances can be interpreted as minimality and sufficiency, respectively (Boudiaf et al., 2020). The two effects lead to disconnected representations (Brown et al., 2022a). By contrast, in regression, the representations are shown to be continuous, connected and form an ordinal relationship with respect to the target (Zhang et al., 2023). The disconnected and connected representations are topologically different, as they have different 0th Betti numbers. The 0th Betti number represents the connectivity in topology, influencing the 'shape' of the feature space[1]. While there are a few works investigating the influence of the representation topology in classification (Hofer et al., 2019; Chen et al., 2019), regression is overlooked. We thus wonder what topology the feature space should have for effective regression and how the topology of the feature space is connected to the IB principle.

In this work, we establish two connections between the topology of the feature space and the IB principle for regression representation learning in deep learning. To establish the connections, we first demonstrate that minimizing the conditional entropies $\mathcal{H}(\mathbf{Y}|\mathbf{Z})$ and $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ can better

---

[1]National Unviersity of Singapore. Correspondence to: Shihao Zhang <zhang.shihao@u.nus.edu>.

---

[1]In this work, the feature space represents the set of projected data points, i.e. the manifold, rather than the entire ambient space.

align with the IB principle. The entropy of a random variable reflects its uncertainty. Specifically, for regression, the conditional entropy $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ is linked to the minimality of $\mathbf{Z}$ and serves as an upper-bound on the generalization error.

The first connection reveals that $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ is bounded by the intrinsic dimension (ID) of the feature space, which suggests encouraging a lower ID feature space for better generalization ability. However, the ID of the feature space should not be less than the ID of the target space to guarantee sufficient representation capabilities. Thus, a feature space with ID equals the target space is desirable. The intrinsic dimension (ID) is a fundamental property of data topology. Intuitively, it can be regarded as the minimal number of dimensions to describe the representation without significant information loss (Ansuini et al., 2019).

The second connection reveals that having a representation $\mathbf{Z}$ homeomorphic to the target space $\mathbf{Y}$ is desirable when both $\mathcal{H}(\mathbf{Y}|\mathbf{Z})$ and $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ are minimal. The homeomorphism between two spaces can be described intuitively as the continuous deformation of one space to the other. From a topological viewpoint, two spaces are considered the same if they are homeomorphic (Hatcher, 2001). However, directly enforcing homeomorphism can be challenging to achieve since the representation $\mathbf{Z}$ typically lies in a high-dimensional space that cannot be modeled without sufficient data samples. As such, we opted to enforce the topological similarity between the target and feature spaces. Here, topological similarity refers to the similarity in topological features, such as clusters and loops, and their localization (Trofimov et al., 2023).

These connections naturally inspire us to learn a regression feature space that is topologically similar to and has the same intrinsic dimension as the target space. To this end, we introduce a regularizer called Persistent Homology Regression Regularizer (PH-Reg). In classification, interest has grown in regulating the intrinsic dimension. For instance, Zhu et al. (2018) explicitly penalizes intrinsic dimension as regularization, while Ma et al. (2018) uses intrinsic dimensions as weights for noise label correction. However, a theoretical justification for using intrinsic dimension as a regularizer is lacking, and they overlook the topology of the target space. Experiments on various regression tasks demonstrate the effectiveness of PH-Reg. Our main contributions are three-fold:

- We are the first to investigate effective feature space topologies for regression. We establish novel connections between the topology of the feature space and the IB principle, which also provides justification for exploiting intrinsic dimension as a regularizer.

- Based on the IB principle, we demonstrate that $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ serves as an upper-bound on the generaliza-

tion error in regression, providing insights for enhancing generalization ability.

- We introduce a regularizer named PH-Reg based on the established connections. Applying PH-Reg achieves significant improvement in coordinate prediction on synthetic datasets and real-world regression tasks such as super-resolution, age estimation, and depth estimation.

2. Related Works

Intrinsic dimension. Raw data and learned data representations often lie on lower intrinsic dimension manifolds but are embedded within a higher-dimensional ambient space (Bengio et al., 2013). The intrinsic dimension of the feature space from the last hidden layer has shown a strong connection with the network generalization ability (Ansuini et al., 2019), and several widely used regularizers like weight decay and dropout effectively reduce the intrinsic dimension (Brown et al., 2022b). Commonly, the generalization ability increases with the decrease of the intrinsic dimension. However, a theoretical justification for why this happened is lacking, and our established connections provide an explanation for this phenomenon in regression.

The intrinsic dimension can be estimated by methods such as the TwoNN (Facco et al., 2017) and Birdal's estimator (Birdal et al., 2021). Among the relevant studies, (Birdal et al., 2021) is the most closely related to ours. This work demonstrates that the generalization error can be bounded by the intrinsic dimension of training trajectories, which possess fractal structures. However, their analysis is based on the parameter space, while ours is on the feature space. Furthermore, we take the target space into consideration, ensuring sufficient representation capabilities.

Topological data analysis. Topological data analysis is a recent field that provides a set of topological and geometric tools to infer robust features for complex data (Chazal & Michel, 2021). It can be coupled with feature learning to ensure that learned representations are robust and reflect the training data's underlying topology and geometric information (Rieck et al., 2020). It has benefitted diverse tasks ranging from fMRI data analysis (Rieck et al., 2020) to and AI-generated text detection (Tulchinskii et al., 2023). It can also be used as a tool to compare data representations (Barannikov et al., 2021a) and data manifolds (Barannikov et al., 2021b). To learn representations that reflect the topology of the training data, a common strategy is to preserve different dimensional topologically relevant distances of the input space and the feature space (Moor et al., 2020; Trofimov et al., 2023). We follow Moor et al. (2020) to preserve topology information. However, unlike classification, regression's target space is naturally a metric space rich in topology induced by the metric, and

crucial for the intended task. Consequently, we leverage the topology of the target space, marking the first exploration of topology specific to effective representation learning for regression.

3. Learning a Desirable Regression Representation

From a topology point of view, what topological properties should a representation for regression have? More simply put, what 'shape' or structure should the feature space have for effective regression? In this work, we suggest a desirable regression representation should (1) have a feature space topologically similar to the target space and (2) the intrinsic dimension of the feature space should be the same as the target space. We arrive at this conclusion by establishing two connections between the topology of the feature space and the Information Bottleneck principle.

Below, we first introduce the notations in Sec. 3.1 and connect the IB principle with two terms $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ and $\mathcal{H}(\mathbf{Y}|\mathbf{Z})$ in Sec. 3.2. We then demonstrate that $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ is the upper-bound on the generalization error in regression in Sec. 3.3. This later provides justification for why lower ID implies higher generalization ability. Subsequently, we establish the first connection in Sec. 3.5, revealing that $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ is bounded by the ID of the feature space. Finally, we establish the second connection, the topological similarity between the feature and target spaces, in Sec. 3.6. Two motivating examples are provided in Sec. 3.4 to enhance understanding of the two connections intuitively.

3.1. Notations

Consider a dataset $S = \{\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i\}_{i=1}^N$ with N samples, sampled from a distribution P with the corresponding label $\mathbf{y}_i \in \mathcal{Y}$. To predict $\mathbf{y}_i$, a neural network first encodes the input $\mathbf{x}_i$ to a representation $\mathbf{z}_i \in \mathbb{R}^d$ before apply a regressor f, i.e. $\hat{\mathbf{y}}_i = f(\mathbf{z}_i)$. The encoder and the regressor f are trained by minimizing a task-specific regression loss $\mathcal{L}_m$ based on a distance between $\hat{\mathbf{y}}_i$ and $\mathbf{y}_i$, i.e. $\mathcal{L}_m = g(||\hat{\mathbf{y}}_i - \mathbf{y}_i||_2)$. Typically, an L2 loss is used, i.e. $\mathcal{L}_m = \frac{1}{N} \sum_i ||\hat{\mathbf{y}}_i - \mathbf{y}_i||_2$, though more robust variants exist such as L1 or the scale-invariant error (Eigen et al., 2014). Note that the dimensionality of $\mathbf{y}$ is task-specific and is not limited to 1. We denote $\mathbf{X}, \mathbf{Y}$, and $\mathbf{Z}$ as random variables representing $\mathbf{x}, \mathbf{y}$, and $\mathbf{z}$, respectively.

3.2. $\mathcal{IB}$ purely between $\mathbf{Y}$ and $\mathbf{Z}$

The IB tradeoff is a practical implementation of the IB principle in machine learning. It suggests that a desirable $\mathbf{Z}$ should contain sufficient information about the target $\mathbf{Y}$ (i.e., maximize the mutual information $\mathcal{I}(\mathbf{Z}; \mathbf{Y})$) and minimal information about the input $\mathbf{X}$ (i.e., minimize

$\mathcal{I}(\mathbf{Z}; \mathbf{X})$). The trade-off between the two aims is typically formulated as a minimization of the associated Lagrangian, $\mathcal{IB} := -\mathcal{I}(\mathbf{Z}; \mathbf{Y}) + \beta \mathcal{I}(\mathbf{Z}; \mathbf{X})$, where $\beta > 0$ is the Lagrange multiplier.

To establish the connections, we first formulate the IB tradeoff into relationships purely between $\mathbf{Y}$ and $\mathbf{Z}$. The following theorem shows that minimizing the conditional entropies $\mathcal{H}(\mathbf{Y}|\mathbf{Z})$ and $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ can be seen as a proxy for optimizing the IB tradeoff when $\beta \in (0, 1)$:

Theorem 1 *Assume that the conditional entropy $\mathcal{H}(\mathbf{Z}|\mathbf{X})$ is a fixed constant for $\mathbf{Z} \in \mathcal{Z}$ for some set $\mathcal{Z}$ of the random variables, or that $\mathbf{Z}$ is deterministic given $\mathbf{X}$. Then, $\min_{\mathbf{Z}} \mathcal{IB} = \min_{\mathbf{Z}} \{(1-\beta)\mathcal{H}(\mathbf{Y}|\mathbf{Z}) + \beta\mathcal{H}(\mathbf{Z}|\mathbf{Y})\}$.*

The detailed proof of Theorem 1 is provided in Appendix A.1. Here, we provide a brief overview by decomposing the terms. The conditional entropy $\mathcal{H}(\mathbf{Y}|\mathbf{Z})$ encourages the learned representation $\mathbf{Z}$ to be informative about the target variable $\mathbf{Y}$. When considering $\mathcal{I}(\mathbf{Z}; \mathbf{Y})$ as a signal, the term $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ in Theorem 1 can be thought of as noise, since $\mathcal{I}(\mathbf{Z}; \mathbf{Y}) = \mathcal{H}(\mathbf{Z}) - \mathcal{H}(\mathbf{Z}|\mathbf{Y})$ and $\mathcal{H}(\mathbf{Z})$ represents the total information. Consequently, minimizing $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ can be seen as learning a minimal representation by reducing noise. The minimality can reduce the complexity of $\mathbf{Z}$ and prevent neural networks from overfitting (Tishby & Zaslavsky, 2015).

It is worth mentioning that the fixed constant assumption given in Theorem 1 holds for most neural networks, as neural networks are commonly deterministic functions. For stochastic representations, we commonly learn a distribution $p(\mathbf{Z}|\mathbf{X})$ approaching a fixed distribution, like the standard Gaussian distribution in VAE. In this case, $\mathcal{H}(\mathbf{Z}|\mathbf{X})$ will tend to be a fixed constant for $\mathbf{Z}$. Discussions about the choice of β, i.e. $\beta \in (0, 1)$ or $\beta > 1$, and more illustrations are given in Appendix B.

3.3. $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ upper-bound on the generalization error

Next, we show $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ upper-bound on the generalization error.

Theorem 2 *Consider dataset $S = \{\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i\}_{i=1}^N$ sampled from distribution P, where $\mathbf{x}_i$ is the input, $\mathbf{z}_i$ is the corresponding representation, and $\mathbf{y}_i$ is the label. Let $d_{max} = \max_{\mathbf{y} \in \mathcal{Y}} \min_{\mathbf{y}_i \in S} ||\mathbf{y} - \mathbf{y}_i||_2$ be the maximum distance of $\mathbf{y}$ to its nearest $\mathbf{y}_i$. Assume $(\mathbf{Z}|\mathbf{Y} = \mathbf{y}_i)$ follows a distribution $\mathcal{D}$ and the dispersion of $\mathcal{D}$ is bounded by its entropy:*

$$\mathbb{E}_{\mathbf{z} \sim \mathcal{D}}[||\mathbf{z} - \bar{\mathbf{z}}||_2] \leq Q(\mathcal{H}(\mathcal{D})), \tag{1}$$

where $\bar{\mathbf{z}}$ is the mean of the distribution $\mathcal{D}$ and $Q(\mathcal{H}(\mathcal{D}))$ is some function of $\mathcal{H}(\mathcal{D})$. Assume the regressor f is L_1-

Lipschitz continuous, then as $d_{max} \to 0$, we have:

$$\mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\}\sim P}[||f(\mathbf{z}) - \mathbf{y}||_2] \tag{2}$$

$$\leq \mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\}\sim S}(||f(\mathbf{z}) - \mathbf{y}||_2) + 2L_1 Q(\mathcal{H}(\mathbf{Z}|\mathbf{Y})) \tag{3}$$

The detailed proof of Theorem 2 is provided in Appendix A.2, and a comparison to a related bound is given in Appendix C. Theorem 2 states that the generalization error $|\mathbb{E}_P[||f(\mathbf{z})-\mathbf{y}||_2] - \mathbb{E}_S[||f(\mathbf{z})-\mathbf{y}||_2]|$, defined as the difference between the population risk $\mathbb{E}_P[||f(\mathbf{z}) - \mathbf{y}||_2]$ and the empirical risk $\mathbb{E}_S[||f(\mathbf{z})-\mathbf{y}||_2]$, is bounded by the $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ in Theorem 1. Theorem 2 suggests minimizing $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ will improve generalization performance.

The tightness of the bound in Theorem 2 depends on the function Q, which aims to bound the dispersion (i.e., $\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[||\mathbf{z} - \bar{\mathbf{z}}||_2]$) of a distribution by its entropy. For a given distribution $\mathcal{D}$, Q exists when its dispersion and entropy are bounded, as we can find a Q to scale its entropy larger than its dispersion in this case. Proposition 1 provides examples of the function Q for various distributions, and the corresponding proof is provided in Appendix A.2.

Proposition 1 *If $\mathcal{D}$ is a multivariate normal distribution $\mathcal{N}(\bar{\mathbf{z}}, \Sigma = k\mathbf{I})$, where $k > 0$ is a scalar and $\bar{\mathbf{z}}$ is the mean of the distribution $\mathcal{D}$. Then, the function $Q(\mathcal{H}(\mathcal{D}))$ in Theorem 2 can be selected as $Q(\mathcal{H}(\mathcal{D})) = \sqrt{\frac{d(e^{2\mathcal{H}(\mathcal{D})})^{\frac{1}{d}}}{2\pi e}}$, where d is the dimension of $\mathbf{z}$. If $\mathcal{D}$ is a uniform distribution, then the $Q(\mathcal{H}(\mathcal{D}))$ can be selected as $Q(\mathcal{H}(\mathcal{D})) = \frac{e^{\mathcal{H}(\mathcal{D})}}{\sqrt{12}}$.*

3.4. Motivating Examples

Encouraging the same intrinsic dimension. Figure 1(a) plots pixel-wise representations of the last hidden layer's feature space, depicted as dots with different colors corresponding to ground truth depth. These representations are obtained from a batch of 32 images from the NYU-v2 test set for depth estimation. A modified ResNet-50 produces these representations, with the last hidden layer changed to dimension 3 for visualization. This figure provides a visualization of the last hidden layer's feature space, where the representations lay on a manifold where the ID varies locally from 1 (blue region) to 3 (green region). The black arrow represents the linear regressor's weight vector θ, and the predicted depth $\hat{\mathbf{Y}} = f(\mathbf{Z}) = \theta^{\mathsf{T}}\mathbf{z}$ is obtained by mapping $\mathbf{Z}$ (represented as dots) to θ. The gray plane represents the solution space of $f(\mathbf{Z}) = \hat{\mathbf{y}}_i$, and the entropy of its distribution in this plane, i.e. $\mathcal{H}(\mathbf{Z}|\hat{\mathbf{Y}} = \hat{\mathbf{y}}_i)$, can be seen as an approximate of $\mathcal{H}(\mathbf{Z}|\mathbf{Y} = \mathbf{y}_i)$.

The target space for depth estimation is one-dimensional; enforcing an intrinsic dimension to match the 1D target space will squeeze the feature space into a line. Under such a scenario, the solution space of $f(\mathbf{z}) = \hat{\mathbf{y}}_i$ is compressed into a point, implying $\mathcal{H}(\mathbf{Z}|\hat{\mathbf{Y}} = \hat{\mathbf{y}}_i) = 0$ (discrete case) and a lower $\mathcal{H}(\mathbf{Z}|\mathbf{Y} = \mathbf{y}_i)$. Lower $\mathcal{H}(\mathbf{Z}|\mathbf{Y} = \mathbf{y}_i)$ for all i implies a lower $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$. *Thus, by controlling the ID, we obtain a lower $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$, implying a higher generalization ability.* Since the ID of the feature space is commonly higher than the ID of the target space, the first connection generally encourages learning a lower ID feature space.

In classification, we tighten clusters for a lower $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$, while in regression, lowering the ID achieves a lower $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$. Lowering the ID of feature space can be intuitively understood as tightening the clusters in classification, where each solution space represents a cluster in classification.

Enforcing topological similarity. Figure 1(b) provides a PCA visualization (from 100 dimension to 3 dimension, t-sne visualization can be found in Figure 3) of the feature space with a 'Mammoth' shape target space (see Sec. 5.1 for details). This feature space is topologically similar to the target space, which indicates regression potentially captures the topology of the target space. The second connection suggests improving such similarity.

3.5. Encouraging the Same Intrinsic Dimension

Now, we can establish our first connection, which reveals that $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ is bounded by the ID of the feature space. Note, intrinsic dimension is not a well-defined mathematical object, and different mathematical definitions exist (Ma et al., 2018; Birdal et al., 2021). We first define Intrinsic Dimension following Ghosh & Motani (2023):

Definition 1 *(Intrinsic Dimension). We define the intrinsic dimension of the manifold $\mathcal{M}$ of a random variable $\mathbf{X}$ as*

$$Dim_{ID}\mathcal{M} = \lim_{\epsilon \to 0^+} \mathbb{E}_{\rho \sim p(\mathbf{X})}[d_\epsilon(\rho)], \tag{4}$$

where $d_\epsilon(\rho) = \min n$ s.t. $(\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_n) \leftrightarrow \mathbf{X}_\epsilon^\rho$ can be regarded as the intrinsic dimension locally at point ρ in the manifold $\mathcal{M}$. $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_n$ represent random random variables, $\leftrightarrow$ means there exist continuous functions f_1, f_2 such that $f_1(\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_n) = X_\epsilon^\rho$ and $f_2(\mathbf{X}_\epsilon^\rho) = (\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_n)$. $\mathbf{X}_\epsilon^\rho$ is a new random variable that follows distribution P_ϵ^ρ given by:

$$P_\epsilon^\rho(\mathbf{X}) = \begin{cases} \frac{P(\mathbf{X})}{c}, & if \, ||\mathbf{X} - \rho|| \leq \epsilon \\ 0, & otherwise \end{cases} \tag{5}$$

where $c = \int_{||\mathbf{X}-\rho||\leq\epsilon} P(\mathbf{X})d\mathbf{X}$.

The manifold is assumed to locally resemble a n-dimensional Euclidean space. Intuitively, we can consider the ID as the expectation of n over the distribution of this manifold.
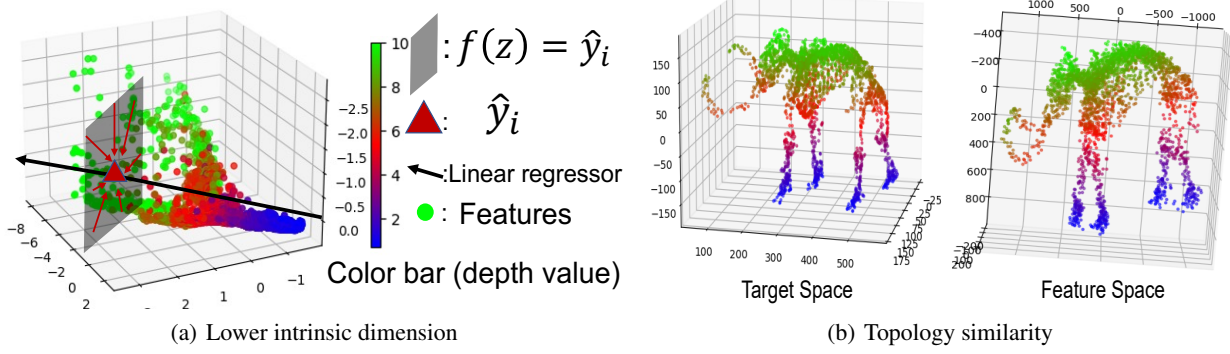
Figure 1. (a) Visualization of the feature space from depth estimation task. Enforcing an ID equal to the target space (1 dimensional) will squeeze the feature space into a line, reducing the unnecessary $\mathcal{H}(\mathbf{Z}|\mathbf{Y} = \mathbf{y}_i)$ corresponding to the solution space of $f(\mathbf{z}) = \hat{y}_i$ (the gray quadrilateral) for all i and implying a lower $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$. (b) Visualization of the feature space (right) and the 'Mammoth' shape target space (left), see Sec. 5.1 for details. The feature space is topologically similar to the target space .

Theorem 3 *Assume that* $\mathbf{z}$ *lies in a manifold* $\mathcal{M}$ *and the* $\mathcal{M}_i \subset \mathcal{M}$ *is a manifold corresponding to the distribution* $(\mathbf{z}|\mathbf{y} = \mathbf{y}_i)$. *Let* $C(\epsilon)$ *be some function of* ϵ:

$$C(\epsilon) = \int_{||\mathbf{z}-\mathbf{z}'|| \leq \epsilon} P'(\mathbf{z})d\mathbf{z}, \qquad (6)$$

where $P'(\mathbf{z})$ *is the probability of* $\mathbf{z}$ *when* $(\mathbf{z}|\mathbf{y} = \mathbf{y}_i)$ *is uniformly distributed across* $\mathcal{M}_i$, *and* $\mathbf{z}'$ *is any point on* $\mathcal{M}_i$. *Then, as* $\epsilon \to 0^+$, *we have:*

$$\mathcal{H}(\mathbf{Z}|\mathbf{Y}) = \mathbb{E}_{\mathbf{y}_i \sim \mathcal{Y}} \mathcal{H}(\mathbf{Z}|\mathbf{Y} = \mathbf{y}_i) \qquad (7)$$

$$\leq \mathbb{E}_{\mathbf{y}_i \sim \mathcal{Y}}[-\log(\epsilon)Dim_{ID}\mathcal{M}_i + \log \frac{K}{C(\epsilon)}], \quad (8)$$

for some fixed scalar K. $Dim_{ID}\mathcal{M}_i$ *is the intrinsic dimension of the manifold* $\mathcal{M}_i$.

Theorem 3 is derived from [(Ghosh & Motani, 2023), Proposition 1]. The detailed proof is provided in Appendix A.3. Theorem 3 states that the conditional entropy $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ is bounded by the IDs of manifolds corresponding to the distribution $(\mathbf{z}|\mathbf{y} = \mathbf{y}_i)$, and the bound is tight when $(\mathbf{z}|\mathbf{y} = \mathbf{y}_i)$ are uniformly distributed across the manifolds.

Since $\mathcal{M}_i \subset \mathcal{M}$, Theorem 3 suggests that reducing the intrinsic dimension of the feature space $\mathcal{M}$ will lead to a lower $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$, which in turn implies a better generalization performance based on Theorem 2. On the other hand, the ID of $\mathcal{M}$ should not be less than the intrinsic dimension of the target space to guarantee sufficient representation capabilities. Thus, a $\mathcal{M}$ with an intrinsic dimension equal to the dimensionality of the target space is desirable.

3.6. Enforcing Topological Similarity

Below, we establish the second connection: topological similarity between the feature and target spaces. We

first define the optimal representation following Achille & Soatto (2018b).

Definition 2 *(Optimal Representation).* *The representation* $\mathbf{Z}$ *is optimal if (1)* $\mathcal{H}(\mathbf{Y}|\mathbf{Z}) = \mathcal{H}(\mathbf{Y}|\mathbf{X})$ *and (2)* $\mathbf{Z}$ *is fully determined given* $\mathbf{Y}$, *i.e.* $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ *is minimal.*

In Definition 2, $\mathcal{H}(\mathbf{Y}|\mathbf{Z}) = \mathcal{H}(\mathbf{Y}|\mathbf{X})$ means $\mathbf{Z}$ is sufficient for the target $\mathbf{Y}$, while $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ is minimal means $\mathbf{Z}$ discards all information that is not relevant to $\mathbf{Y}$, and $\mathbf{Z}$ is fully determined given $\mathbf{Y}$. For continuous entropy, a minimal $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ implies that $\mathcal{H}(\mathbf{Z}|\mathbf{Y}) = -\infty$, as $\mathbf{Z}$ is distributed as a delta function once given $\mathbf{Y}$. In the discrete case, $\mathcal{H}(\mathbf{Z}|\mathbf{Y}) = 0$.

Proposition 2 *Let the target* $\mathbf{Y} = \mathbf{Y}' + \mathbf{N}'$ *where* $\mathbf{Y}'$ *is fully determined by* $\mathbf{X}$ *and* $\mathbf{N}'$ *is the aleatoric uncertainty that is independent of* $\mathbf{X}$. *Assume the underlying mapping* f' *from* $\mathbf{Z}$ *to* $\mathbf{Y}'$ *and* g' *from* $\mathbf{Y}'$ *to* $\mathbf{Z}$ *are continuous, where the continuous mapping is based on the topology induced by the Euclidean distance. Then the representation* $\mathbf{Z}$ *is optimal if and only if* $\mathbf{Z}$ *is homeomorphic to* $\mathbf{Y}'$.

The detailed proof of Proposition 2 is provided in Appendix A.4. Proposition 2 demonstrates that the optimal $\mathbf{Z}$ is homeomorphic to $\mathbf{Y}'$, implying the need to learn a $\mathbf{Z}$ that is homeomorphic to $\mathbf{Y}'$. However, directly enforcing homeomorphism can be challenging to achieve since $\mathbf{Y}'$ is generally unknown, and the representation $\mathbf{Z}$ typically lies in a high-dimensional space that cannot be modeled without sufficient data samples. As such, we opted to enforce the topological similarity between the target and feature spaces, preserving topological features similar to homomorphism. Here, topological similarity refers to the similarity in topological features, such as clusters and loops, and their localization (Trofimov et al., 2023). The two established connections imply that the desired $\mathbf{Z}$ should be

topologically similar to the target space and share the same ID as the target space. More illustrations are given in Appendix D and E.

4. PH-Reg for Regression

Our analysis in Sec. 3 inspires us to learn a feature space that is (1) topologically similar to the target space and (2) with an intrinsic dimension (ID) equal to that of the target space. To this end, we propose a regularizer named Persistent Homology Regression Regularizer (PH-Reg). PH-Reg features two terms: an intrinsic dimension term $\mathcal{L}_d$ and a topology term $\mathcal{L}_t$. $\mathcal{L}_d$ follows Birdal's regularizer (Birdal et al., 2021) to control the ID of feature space. Additionally, it considers the target space to ensure sufficient representation capabilities. $\mathcal{L}_d$ exploit the topology autoencoder (Moor et al., 2020) to encourage the topological similarity. Note the two regularizer terms are mainly introduced to verify our connections, and other ID and topology regularizers can also be considered. However, empirical observations suggest that our $\mathcal{L}_d$ and $\mathcal{L}_t$ effectively align with our established connections, perform well, and do not conflict with each other.

We first introduce some notations. Let $\mathbf{Z}_n$ represent the set of n samples from $\mathbf{Z}$, and $\mathbf{Y}_n$ be the labels corresponding to $\mathbf{Z}_n$. We denote $\mathrm{PH}_0(\mathrm{VR}(\mathbf{Z}_n))$ the 0-dimensional persistent homology. Intuitively, $\mathrm{PH}_0(\mathrm{VR}(\mathbf{Z}_n))$ can be regarded as a set of edge lengths, where the edges are derived from the minimum spanning tree obtained from the distance matrix $\mathbf{A}^{\mathbf{Z}_n}$ of $\mathbf{Z}_n$. We denote $\pi^{\mathbf{Z}_n}, \pi^{\mathbf{Y}_n}$ the set of the index of edges in the minimum spanning trees of $\mathbf{Z}_n$ and $\mathbf{Y}_n$, respectively, and $\mathbf{A}^*[\pi^*]$ the corresponding length of the edges. Let $E(\mathbf{Z}_n) = \sum_{\gamma \in \mathrm{PH}_0(\mathrm{VR}(\mathbf{Z}_n))} |I(\gamma)|$ be the sum of edge lengths of the minimum spanning trees corresponding to $\mathbf{Z}_n$. We define $E(\mathbf{Y}_n)$ similarly. Some topology preliminaries are given in Appendix F.

Birdal et al. (2021) suggests to estimate the intrinsic dimension as the slope between $\log E(\mathbf{Z}_n)$ and $\log n$. Note, the definition of intrinsic dimension used in Birdal et al. (2021) is based on the 0-dimensional persistent homology $\mathrm{PH}_0(\mathrm{VR}(\mathbf{Z}_n))$, which is different from ours (Definition 1, coming from Ghosh & Motani (2023)). However, both definitions define the same object, i.e. the intrinsic dimension, and it is thus reasonable to exploit Birdal et al. (2021)'s method to constrain the intrinsic dimension.

Let $\mathbf{e}' = [\log E(\mathbf{Z}_{n_1}), \log E(\mathbf{Z}_{n_2}), \cdots, \log E(\mathbf{Z}_{n_m})]$, where $\mathbf{Z}_{n_i}$ is the subset sampled from a batch, with size $n_i = |\mathbf{Z}_{n_i}|$. Let $n_i < n_j$ for $i < j$, and $\mathbf{n} = [\log n_1, \log n_2, \cdots, \log n_m]$. Birdal et al. (2021) encourage a lower intrinsic dimension feature space by minimizing the slope between $\mathbf{e}'$ and $\mathbf{n}$, which can be estimated via the least square method:

$$\mathcal{L}'_d = (m \sum_{i=1}^{m} \mathbf{n}_i \mathbf{e}'_i - \sum_{i=1}^{m} \mathbf{n}_i \sum_{i=1}^{m} \mathbf{e}'_i)/(m \sum_{i=1}^{m} \mathbf{n}_i^2 - (\sum_{i=1}^{m} \mathbf{n}_i)^2). \tag{9}$$

Intuitively, the growth rate of $E(\mathbf{Z}_n)$ is proportional to the volume of the corresponding manifold; this volume is proportional to the intrinsic dimension. In fact, there is a classical result on the growth rate of (Steele, 1988), showing that the growth rate (i.e. the slop) can constrain the intrinsic dimension.

$\mathcal{L}'_d$ purely encourage the feature space to have a lower intrinsic dimension; sometimes it may even result in an intrinsic dimension lower than that of the target space (see Figure 3, Swiss Roll, where the target space is two-dimensional and the feature space is almost one-dimensional.). In contrast, we wish to lower the ID of the feature space while preventing it from being lower than that of the target space. We propose to minimize slope between $\mathbf{e} = [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_m]$ and $\mathbf{n}$:

$$\mathcal{L}_d = |(m \sum_{i=1}^{m} \mathbf{n}_i \mathbf{e}_i - \sum_{i=1}^{m} \mathbf{n}_i \sum_{i=1}^{m} \mathbf{e}_i)/(m \sum_{i=1}^{m} \mathbf{n}_i^2 - (\sum_{i=1}^{m} \mathbf{n}_i)^2)|, \tag{10}$$

where $\mathbf{e}_i = \log E(\mathbf{Z}_{n_i})/\log E(\mathbf{Y}_{n_i})$. Compared with $\mathcal{L}'_d$, $\mathcal{L}_d$ further exploits the topological information of the target space through $\log E(\mathbf{Y}_{n_i})$. When the feature and target spaces have the same ID, $E(\mathbf{Z}_{n_i}) = E(\mathbf{Y}_{n_i})$ for all i and $\mathcal{L}_d = 0$ is in its minimal. As shown in Figure 2(c) and Figure 3, $\mathcal{L}_d$ well controls the ID of the feature space while better preserving the topology of the target space.

The topology autoencoder (Moor et al., 2020) enforces the topological similarity between the feature and the target spaces by preserving 0-dimensional topologically relevant distances from the two spaces. We exploit it as the topology part $\mathcal{L}_t$:

$$\mathcal{L}_t = ||\mathbf{A}^{\mathbf{Z}_{n_m}}[\pi^{\mathbf{Z}_{n_m}}] - \mathbf{A}^{\mathbf{Y}_{n_m}}[\pi^{\mathbf{Z}_{n_m}}]||_2^2 \tag{11}$$

$$+ ||\mathbf{A}^{\mathbf{Z}_{n_m}}[\pi^{\mathbf{Y}_{n_m}}] - \mathbf{A}^{\mathbf{Y}_{n_m}}[\pi^{\mathbf{Y}_{n_m}}]||_2^2 \tag{12}$$

As shown in Figure 2(d) and Figure 3, $\mathcal{L}_t$ well preserves the topology of the target space. We define the persistent homology regression regularizer, PH-Reg, as $\mathcal{L}_R = \mathcal{L}_d + \mathcal{L}_t$. As shown in Figure 2(e) and Figure 3, PH-Reg can both encourage a lower intrinsic dimension and preserve the topology of target space. The final loss function $\mathcal{L}_R$ is defined as:

$$\mathcal{L}_R = \mathcal{L}_m + \lambda_t \mathcal{L}_t + \lambda_d \mathcal{L}_d, \tag{13}$$

where $\mathcal{L}_m$ is the task-specific regression loss and λ_d, λ_t are trade-off parameters, and their values are determined by the value of the task task-specific loss $\mathcal{L}_m$, e.g. for a high $\mathcal{L}_m$, λ_d and λ_t should also be set to high values.
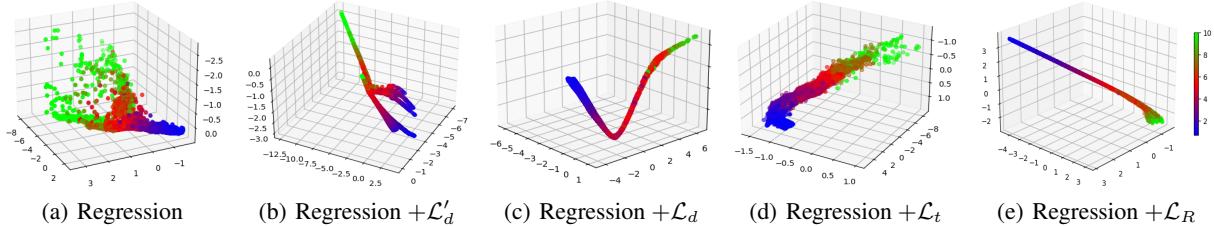
Figure 2. Visualization of the last hidden layer's feature space from the depth estimation task. The representations are obtained through a modified ResNet-50, with the last hidden layer changed to dimension 3 for visualization. The target space is a 1-dimensional line, and colors represent the ground truth depth. (b) $\mathcal{L}'_d$ encourages a lower intrinsic dimension yet fails to preserve the topology of the target space. (c) $\mathcal{L}_d$ takes the target space into consideration and can further preserve its topology. (d) $\mathcal{L}_t$ can enforce the topological similarity between the feature and target spaces. (e) Adding the $\mathcal{L}_t$ to $\mathcal{L}_d$ better preserves the topology of the target space.

5. Experiments

We compare our method with four methods. 1) Information Dropout (InfDrop) (Achille & Soatto, 2018b). InfDrop serves as an IB baseline. It functions as a regularizer designed based on IB, aiming to learn representations that are minimal, sufficient, and disentangled. 2) Ordinal Entropy (OE) (Zhang et al., 2023). OE acts as a regression baseline. It takes advantage of classification by learning higher entropy feature space for regression tasks. 3) Birdal's regularizer (i.e., $\mathcal{L}'_d$) (Birdal et al., 2021) serves as an intrinsic dimension baseline. 4) Topology Autoencoder (i.e., $\mathcal{L}_t$) (Moor et al., 2020) serves as a topology baseline. Note that the proposed PH-Reg is mainly introduced to verify the established connections, and we do not aim for state-of-the-art results.

5.1. Coordinate Prediction on the Synthetic Dataset

To verify the topological relationship between the feature space and target space, we synthesize a dataset that contains points sampled from topologically different objects, including Swiss roll, torus, circle and the more complex object "mammoth" (Coenen & Pearce, 2019). We randomly sample 3000 points with coordinate $\mathbf{y} \in \mathbb{R}^3$ from each object. These 3000 points are then divided into sets of 100 for training, 100 for validation, and 2800 for testing. Each point $\mathbf{y}_i$ is encoded into a 100 dimensional vector $\mathbf{x}_i = [f_1(\mathbf{y}_i), f_2(\mathbf{y}_i), f_3(\mathbf{y}_i), f_4(\mathbf{y}_i), \text{noise}]$, where the dimensions 1-4 are signal and the rest 96 dimensions are noise. The coordinate prediction task aims to learn the mapping $G(\mathbf{x}) = \hat{\mathbf{y}}$ from $\mathbf{x}$ to $\mathbf{y}$, and the mean-squared error $\mathcal{L}_{\text{mse}} = \frac{1}{N} \sum_i ||\hat{\mathbf{y}}_i - \mathbf{y}_i||_2^2$ is adopted as the evaluation metric. We use a two-layer fully connected neural network with 100 hidden units as the baseline architecture. More details are given in Appendix G.

Table 1 shows that encouraging a lower intrinsic dimension while considering the target space $(+\mathcal{L}_d)$ enhances performance, particularly for Swiss Roll and Torus. In contrast,

Table 1. Results ($\mathcal{L}_{\text{mse}}$) on the synthetic dataset. We report results as mean $\pm$ standard variance over 10 runs. **Bold** numbers indicate the best performance.

Method	Swiss Roll	Mammoth	Torus	Circle
Baseline	2.99 ± 0.43	211 ± 55	3.01 ± 0.11	0.154 ± 0.006
+ InfDrop	4.15 ± 0.37	367 ± 50	2.05 ± 0.04	0.093 ± 0.003
+ OE	2.95 ± 0.69	187 ± 88	2.83 ± 0.07	0.114 ± 0.007
$+\mathcal{L}'_d$	2.74 ± 0.85	141 ± 104	1.13 ± 0.06	0.171 ± 0.04
$+\mathcal{L}_d$	0.66 ± 0.08	89 ± 66	0.62 ± 0.12	0.090 ± 0.019
$+\mathcal{L}_t$	1.83 ± 0.70	80 ± 61	0.95 ± 0.05	0.036 ± 0.004
$+\mathcal{L}_d + \mathcal{L}_t$	$\mathbf{0.61 \pm 0.17}$	$\mathbf{49 \pm 27}$	$\mathbf{0.61 \pm 0.05}$	$\mathbf{0.013 \pm 0.008}$

naively lowering the intrinsic dimension $(+\mathcal{L}'_d)$ performs poorly. Enforcing the topology similarity between the feature space and target space $(+\mathcal{L}'_t)$ decreases the $\mathcal{L}_{\text{mse}}$ by more than 60%, except for the Swiss roll. The best gains, however, are achieved by incorporating both $\mathcal{L}_t$ and $\mathcal{L}_d$, which decrease the $\mathcal{L}_{\text{mse}}$ even more than 90% for the circle coordinate prediction task. Figure 3 shows feature space visualization results using t-SNE (100 dimensions $\rightarrow$ 3 dimensions). The feature space of the regression baseline shows a similar structure to the target space, especially for Swiss roll and mammoth, which indicates regression potentially captures the topology of the target space. Regression $+\mathcal{L}_t$ significantly preserves the topology of the target space. Regression $+\mathcal{L}_d$ potentially preserves the topology of the target space, e.g. circle, while it primarily reduces the complexity of the feature space by maintaining the same intrinsic dimension as the target space. Combining both $\mathcal{L}_d$ and $\mathcal{L}_t$ in regression preserves the topology information while also reducing the complexity of the feature space, i.e. lower its intrinsic dimension.

5.2. Real-World Regression Tasks

We conduct experiments on three real-world regression tasks, including depth estimation (Table 2), super-resolution (Table 3) and age estimation (Table 4). The target spaces of the three tasks are topologically different, i.e. a 1- dimensional line for depth estimation, 3-dimensional
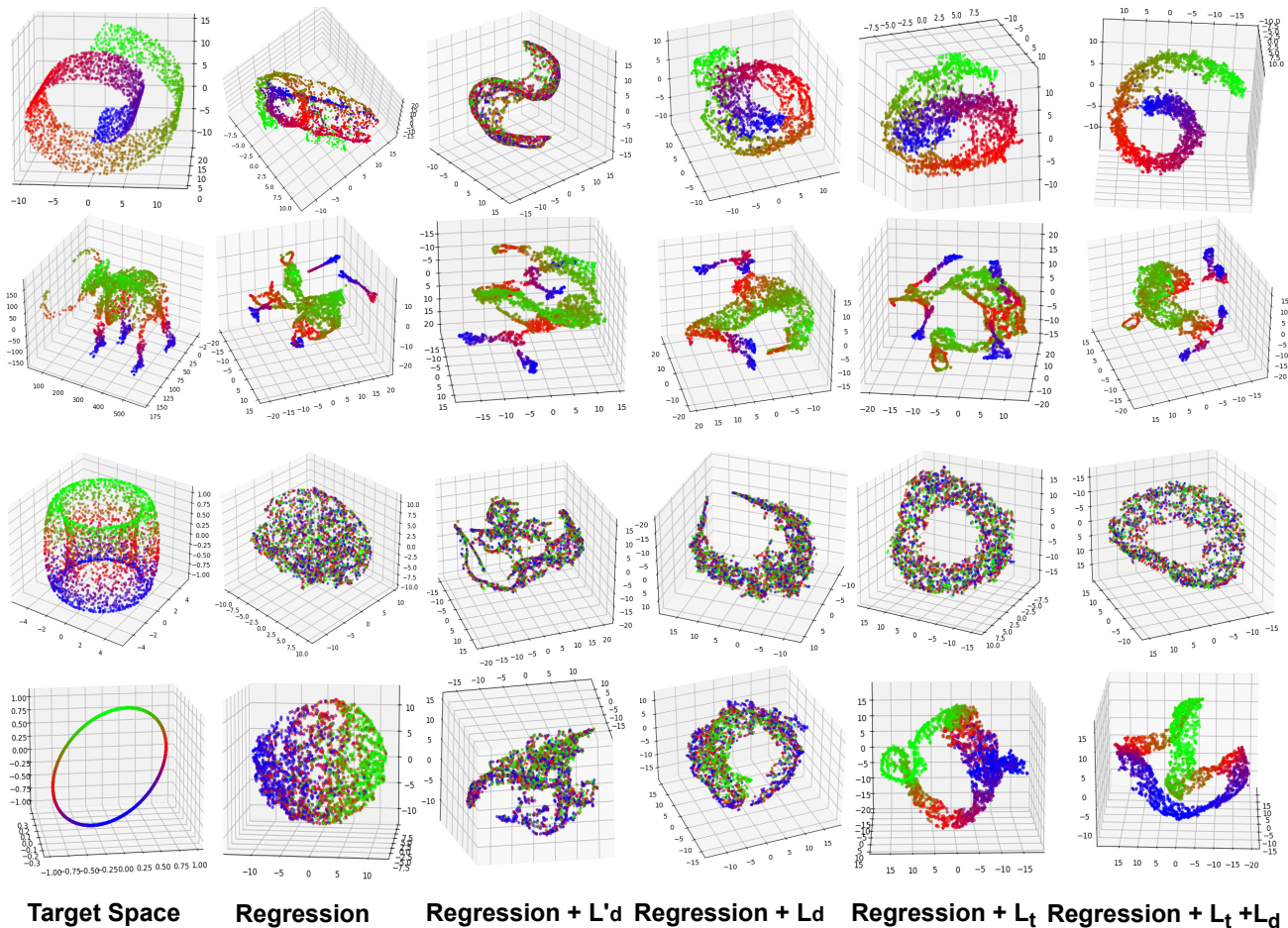
Figure 3. t-sne visualization of the feature spaces (100 dimensions → 3 dimensions) with topological different target spaces.

space for super-resolution and discrete points for age estimation. Detailed settings, related introductions and more discussions are given in Appendix H.

Results on the three tasks demonstrate that both $\mathcal{L}_t$ and $\mathcal{L}_d$ can enhance performance, and combining both further boosts the performance. Specifically, combining both achieves 0.48 overall improvements (i.e. ALL) on age estimation, a PSNR improvement of 0.096 on super-resolution for Urban100, and a reduction of 6.7% δ_1 error on depth estimation.

5.3. Ablation Studies

Hyperparameter λ_t **and** λ_d: We maintain λ_d and λ_t at their default value 10 for Swiss roll coordinate prediction, and we vary one of them to examine their impact. Figure 4(a) shows when $\lambda_t \leq 10$, the MSE decreases consistently as λ_t increases. However, it tends to overtake the original learning objective when set too high, i.e. 1000. Regarding the λ_d, as shown in Figure 4(b), MSE remains relatively stable over a large range of λ_d, with a slight increase in

Table 2. Quantitative comparison (MAE) on AgeDB. We report results as mean ± standard variance over 3 runs. **Bold** numbers indicate the best performance.

Method	ALL	Many	Med.	Few
Baseline	7.80 ± 0.12	6.80 ± 0.06	9.11 ± 0.31	13.63 ± 0.43
+ InfDrop	8.04 ± 0.14	7.14 ± 0.20	9.10 ± 0.71	13.61 ± 0.32
+ OE	7.65 ± 0.13	6.72 ± 0.09	8.77 ± 0.49	13.28 ± 0.73
$+\mathcal{L}'_d$	7.75 ± 0.05	6.80 ± 0.11	8.87 ± 0.05	13.61 ± 0.50
$+\mathcal{L}_d$	7.64 ± 0.07	6.82 ± 0.07	8.62 ± 0.20	12.79 ± 0.65
$+\mathcal{L}_t$	7.50 ± 0.04	6.59 ± 0.03	8.75 ± 0.03	12.67 ± 0.24
$+\mathcal{L}_d + \mathcal{L}_t$	$\mathbf{7.32 \pm 0.09}$	$\mathbf{6.50 \pm 0.15}$	$\mathbf{8.38 \pm 0.11}$	$\mathbf{12.18 \pm 0.38}$

variance when $\lambda_d = 1000$.

Sample Size (n_m): In practice, we model the feature space using a limited number of samples within a batch. For dense prediction tasks, the available No. of samples is very large (No. pixels per image × batch size), while it is constrained to the batch size for image-wise prediction tasks. We investigate the influence of n_m from Eq. 10 and 11 on Swiss roll coordinate prediction. Figure 4(c) shows our PH-Reg performs better with a larger n_m, while maintain-
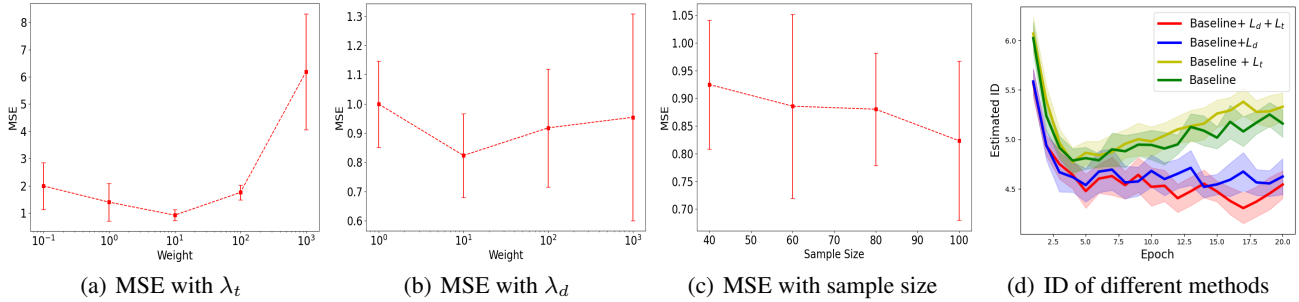
8

Figure 4. Ablation study based on (a-c) the Swiss roll coordinate prediction task and (d) the depth estimation task.

Table 3. Quantitative comparison (PSNR(dB)) of super-resolution results with public benchmark and DIV2K validation set. **Bold** numbers indicate the best performance.

Method	Set5	Set14	B100	Urban100	DIV2K
Baseline	32.241	28.614	27.598	26.083	28.997
+ InfDrop	32.219	28.626	27.594	26.059	28.980
+ OE	32.280	28.659	27.614	26.117	29.005
$+\mathcal{L}'_d$	32.252	28.625	27.599	26.078	28.989
$+\mathcal{L}_d$	32.293	28.644	27.619	26.151	29.022
$+\mathcal{L}_t$	**32.322**	28.673	27.624	26.169	29.031
$+\mathcal{L}_d + \mathcal{L}_t$	32.288	**28.686**	**27.627**	**26.179**	**29.038**

Table 4. Depth estimation results with NYU-Depth-v2. **Bold** and <u>underline</u> numbers indicate the best and second best performance, respectively.

Method	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	REL $\downarrow$	RMS $\downarrow$	$\log_{10} \downarrow$
Baseline	0.792	0.955	0.990	0.153	0.512	0.064
+ InfDrop	0.791	0.960	0.992	0.153	0.507	0.064
+ OE	**0.811**	-	-	**0.143**	**0.478**	**0.060**
$+\mathcal{L}'_d$	0.804	0.954	0.988	0.151	0.502	0.063
$+\mathcal{L}_d$	0.795	**0.959**	**0.992**	0.150	0.497	0.063
$+\mathcal{L}_t$	0.798	0.958	0.990	0.149	0.502	0.063
$+\mathcal{L}_d + \mathcal{L}_t$	<u>0.807</u>	**0.959**	**0.992**	<u>0.144</u>	<u>0.481</u>	<u>0.061</u>

ing stability even with a small n_m.

ID of different methods: Figure 4(d) displays the intrinsic dimension of the last hidden layer, estimated using TwoNN (Facco et al., 2017), for the testing set of NYU-Depth-v2 from different methods throughout training. While our method is based on Birdal's estimator (Birdal et al., 2021), another estimator, TwoNN, captures a decrease in ID when applied $\mathcal{L}_d$. We observe that without $\mathcal{L}_d$, the intrinsic dimension tends to increase after epoch 3, potentially overfitting details, whereas $\mathcal{L}_d$ prevents such a trend.

Efficiency: Efficiency-wise, the computing complexity equals finding the minimum spanning tree from the distance matrix of the samples, which have a complexity of $\mathcal{O}(n_m^2 \log n_m)$ using the simple Kruskal's Algorithm, and it can speed up with some advanced methods (Bauer, 2021). The synthetic experiments (Table 5) use a simple 2-layer MLP, so the regularizer adds significant computing

Table 5. Quantitative comparison of the time consumption and memory usage on the synthetic dataset and NYU-Depth-v2, and the corresponding training times are 10000 and 1 epoch, respectively.

n_m	Regularizer	Coordinate Prediction (2 Layer MLP)		Depth Estimation (ResNet-50)	
		Training(s)	Memory (MB)	Training(s)	Memory (MB)
0	-	8.88	959	1929	11821
100	$\mathcal{L}_t$	175.06	959	1942	11833
100	$\mathcal{L}_d$	439.68	973	1950	12211
100	$\mathcal{L}_t + \mathcal{L}_d$	617.41	973	1980	12211
300	$\mathcal{L}_t + \mathcal{L}_d$	-	-	2370	12211

time. However, the real-world experiments on depth estimation (Table 5) use a ResNet-50 backbone, and the added time and memory are negligible (18.6% and 0.3%, respectively), even with $n_m = 300$. These increases are only during training and do not add demands for inference.

6. Conclusion

In this paper, we establish novel connections between topology and the IB principle for regression representation learning. The established connections imply that the desired **Z** should exhibit topological similarity to the target space and share the same intrinsic dimension as the target space. Inspired by the connections, we proposed a regularizer to learn the desired **Z**. Experiments on synthetic and real-world regression tasks demonstrate its benefits.

Acknowledgement

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Achille, A. and Soatto, S. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018a.

Achille, A. and Soatto, S. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2897–2905, 2018b.

Ansuini, A., Laio, A., Macke, J. H., and Zoccolan, D. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Barannikov, S., Trofimov, I., Balabin, N., and Burnaev, E. Representation topology divergence: A method for comparing neural network representations. *ICML*, 2021a.

Barannikov, S., Trofimov, I., Sotnikov, G., Trimbach, E., Korotin, A., Filippov, A., and Burnaev, E. Manifold topology divergence: a framework for comparing data manifolds. *NeurIPS*, 34:7294–7305, 2021b.

Bauer, U. Ripser: efficient computation of vietoris–rips persistence barcodes. *Journal of Applied and Computational Topology*, 5(3):391–423, 2021.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35 (8):1798–1828, 2013.

Bevilacqua, M., Roumy, A., Guillemot, C., and Alberi-Morel, M. L. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012.

Birdal, T., Lou, A., Guibas, L. J., and Simsekli, U. Intrinsic dimension, persistent homology and generalization in neural networks. *Advances in Neural Information Processing Systems*, 34:6776–6789, 2021.

Boudiaf, M., Rony, J., Ziko, I. M., Granger, E., Pedersoli, M., Piantanida, P., and Ayed, I. B. A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses. In *European conference on computer vision*, pp. 548–564. Springer, 2020.

Brown, B. C., Caterini, A. L., Ross, B. L., Cresswell, J. C., and Loaiza-Ganem, G. Verifying the union of manifolds hypothesis for image data. In *The Eleventh International Conference on Learning Representations*, 2022a.

Brown, B. C., Juravsky, J., Caterini, A. L., and Loaiza-Ganem, G. Relating regularization and generalization through the intrinsic dimension of activations. *arXiv preprint arXiv:2211.13239*, 2022b.

Chazal, F. and Michel, B. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in artificial intelligence*, 4:108, 2021.

Chen, C., Ni, X., Bai, Q., and Wang, Y. A topological regularizer for classifiers via persistent homology. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2573–2582. PMLR, 2019.

Coenen, A. and Pearce, A. Understanding umap, mammoth dataset, 2019. URL https://github.com/MNoichl/UMAP-examples-mammoth-/tree/master.

Eigen, D., Puhrsch, C., and Fergus, R. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

Facco, E., d'Errico, M., Rodriguez, A., and Laio, A. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1): 12140, 2017.

Ghosh, R. and Motani, M. Local intrinsic dimensional entropy. *AAAI*, 2023.

Hatcher, A. *Algebraic topology*. Cambridge University Press, 2001.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hofer, C., Kwitt, R., Niethammer, M., and Dixit, M. Connectivity-optimized representation learning via persistent homology. In *International conference on machine learning*, pp. 2751–2760. PMLR, 2019.

Huang, J.-B., Singh, A., and Ahuja, N. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5197–5206, 2015.

Kawaguchi, K., Deng, Z., Ji, X., and Huang, J. How does information bottleneck help deep learning? *ICML*, 2023.

Lee, J. H., Han, M.-K., Ko, D. W., and Suh, I. H. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

Lim, B., Son, S., Kim, H., Nah, S., and Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144, 2017.

Ma, X., Wang, Y., Houle, M. E., Zhou, S., Erfani, S., Xia, S., Wijewickrema, S., and Bailey, J. Dimensionality-driven learning with noisy labels. In *International Conference on Machine Learning*, pp. 3355–3364. PMLR, 2018.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pp. 416–423. IEEE, 2001.

Moor, M., Horn, M., Rieck, B., and Borgwardt, K. Topological autoencoders. In *International conference on machine learning*, pp. 7045–7054. PMLR, 2020.

Rieck, B., Yates, T., Bock, C., Borgwardt, K., Wolf, G., Turk-Browne, N., and Krishnaswamy, S. Uncovering the topology of time-varying fmri data using cubical persistence. *Advances in neural information processing systems*, 33:6900–6912, 2020.

Shwartz-Ziv, R. and Tishby, N. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

Steele, J. M. Growth rates of euclidean minimal spanning trees with power weighted edges. *The Annals of Probability*, 16(4):1767–1787, 1988.

Timofte, R., Agustsson, E., Van Gool, L., Yang, M.-H., and Zhang, L. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 114–125, 2017.

Tishby, N. and Zaslavsky, N. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pp. 1–5. IEEE, 2015.

Trofimov, I., Cherniavskii, D., Tulchinskii, E., Balabin, N., Burnaev, E., and Barannikov, S. Learning topology-preserving data representations. *ICLR*, 2023.

Tulchinskii, E., Kuznetsov, K., Kushnareva, L., Cherniavskii, D., Barannikov, S., Piontkovskaya, I., Nikolenko, S., and Burnaev, E. Intrinsic dimension estimation for robust detection of ai-generated texts. *NeurIPS*, 2023.

Yang, Y., Zha, K., Chen, Y., Wang, H., and Katabi, D. Delving into deep imbalanced regression. In *International Conference on Machine Learning*, pp. 11842–11851. PMLR, 2021.

Zeyde, R., Elad, M., and Protter, M. On single image scale-up using sparse-representations. In *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, pp. 711–730. Springer, 2012.

Zhang, S., Yang, L., Mi, M. B., Zheng, X., and Yao, A. Improving deep regression with ordinal entropy. *ICLR*, 2023.

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., and Fu, Y. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2472–2481, 2018.

Zhu, W., Qiu, Q., Huang, J., Calderbank, R., Sapiro, G., and Daubechies, I. Ldmnet: Low dimensional manifold regularized neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2743–2751, 2018.

A. Proofs

A.1. Proof of the Theorem 1

Theorem 1 *Assume that the conditional entropy $\mathcal{H}(\mathbf{Z}|\mathbf{X})$ is a fixed constant for $\mathbf{Z} \in \mathcal{Z}$ for some set $\mathcal{Z}$ of the random variables, or $\mathbf{Z}$ is determined given $\mathbf{X}$. Then, $\min_{\mathbf{Z}} \mathcal{IB} = \min_{\mathbf{Z}} \{(1-\beta)\mathcal{H}(\mathbf{Y}|\mathbf{Z}) + \beta\mathcal{H}(\mathbf{Z}|\mathbf{Y})\}$.*

Proof From the definition of the mutual information, we have

$$\mathcal{I}(\mathbf{Z};\mathbf{X}) = \mathcal{H}(\mathbf{Z}) - \mathcal{H}(\mathbf{Z}|\mathbf{X}) = \mathcal{I}(\mathbf{Z};\mathbf{Y}) + \mathcal{H}(\mathbf{Z}|\mathbf{Y}) - \mathcal{H}(\mathbf{Z}|\mathbf{X}).$$

By substituting the right-hand side of this equation into $\mathcal{I}(\mathbf{Z};\mathbf{X})$,

$$\mathcal{IB} = -\mathcal{I}(\mathbf{Z};\mathbf{Y}) + \beta\mathcal{I}(\mathbf{Z};\mathbf{X}) = (\beta-1)\mathcal{I}(\mathbf{Z};\mathbf{Y}) + \beta\mathcal{H}(\mathbf{Z}|\mathbf{Y}) - \beta\mathcal{H}(\mathbf{Z}|\mathbf{X}) \tag{14}$$

Since $\mathcal{I}(\mathbf{Z};\mathbf{Y}) = \mathcal{H}(\mathbf{Y}) - \mathcal{H}(\mathbf{Y}|\mathbf{Z})$,

$$\mathcal{IB} = (\beta-1)(\mathcal{H}(\mathbf{Y}) - \mathcal{H}(\mathbf{Y}|\mathbf{Z})) + \beta\mathcal{H}(\mathbf{Z}|\mathbf{Y}) - \beta\mathcal{H}(\mathbf{Z}|\mathbf{X}) \tag{15}$$

$$= (1-\beta)\mathcal{H}(\mathbf{Y}|\mathbf{Z}) + \beta\mathcal{H}(\mathbf{Z}|\mathbf{Y}) + (\beta-1)\mathcal{H}(\mathbf{Y}) - \beta\mathcal{H}(\mathbf{Z}|\mathbf{X}). \tag{16}$$

1) If $\mathcal{H}(\mathbf{Z}|\mathbf{X})$ is a constant for $\mathbf{Z} \in \mathcal{Z}$. Since $\mathcal{H}(\mathbf{Y})$ is a fixed constant for any $\mathbf{Z}$, this implies that

$$\mathcal{IB} = (1-\beta)\mathcal{H}(\mathbf{Y}|\mathbf{Z}) + \beta\mathcal{H}(\mathbf{Z}|\mathbf{Y}) + C,$$

where C is a fixed constant for $\mathbf{Z} \in \mathcal{Z}$. Thus:

$$\min_{\mathbf{Z}} \mathcal{IB} = \min_{\mathbf{Z}} \{(1-\beta)\mathcal{H}(\mathbf{Y}|\mathbf{Z}) + \beta\mathcal{H}(\mathbf{Z}|\mathbf{Y})\}.$$

2) If $\mathbf{Z}$ is determined given $\mathbf{X}$, then $\mathcal{H}(\mathbf{Z}|\mathbf{X})$ is not a term can be optimized. Since $\mathcal{H}(\mathbf{Y})$ is a fixed constant for any $\mathbf{Z}$:

$$\min_{\mathbf{Z}} \mathcal{IB} = \min_{\mathbf{Z}} \{(1-\beta)\mathcal{H}(\mathbf{Y}|\mathbf{Z}) + \beta\mathcal{H}(\mathbf{Z}|\mathbf{Y})\}.$$

$\square$

A.2. Proof of the Theorem 2 and Proposition 1

Theorem 2 *Consider dataset $S = \{\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i\}_{i=1}^N$ sampled from distribution P, where $\mathbf{x}_i$ is the input, $\mathbf{z}_i$ is the corresponding representation, and $\mathbf{y}_i$ is the label. Let $d_{max} = \max_{\mathbf{y} \in \mathcal{Y}} \min_{\mathbf{y}_i \in S} ||\mathbf{y} - \mathbf{y}_i||_2$ be the maximum distance of $\mathbf{y}$ to its nearest $\mathbf{y}_i$. Assume $(\mathbf{Z}|\mathbf{Y} = \mathbf{y}_i)$ follows a distribution $\mathcal{D}$ and the dispersion of $\mathcal{D}$ is bounded by its entropy:*

$$\mathbb{E}_{\mathbf{z} \sim \mathcal{D}}[||\mathbf{z} - \bar{\mathbf{z}}||_2] \leq Q(\mathcal{H}(\mathcal{D})), \tag{17}$$

where $\bar{\mathbf{z}}$ is the mean of the distribution $\mathcal{D}$ and $Q(\mathcal{H}(\mathcal{D}))$ is some function of $\mathcal{H}(\mathcal{D})$. Assume the regressor f is L_1-Lipschitz continuous, then as $d_{max} \to 0$, we have:

$$\mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\} \sim P}[||f(\mathbf{z}) - \mathbf{y}||_2] \leq \mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\} \sim S}(||f(\mathbf{z}) - \mathbf{y}||_2) + 2L_1 Q(\mathcal{H}(\mathbf{Z}|\mathbf{Y})) \tag{18}$$

Proof For any sample $\{\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i\}$, we define its local neighborhood set N_i as

$$N_i = \{\{\mathbf{x}, \mathbf{z}, \mathbf{y}\} : ||\mathbf{y} - \mathbf{y}_i||_2 < ||\mathbf{y} - \mathbf{y}_j||_2, j \neq i, p(\mathbf{y}) > 0\}. \tag{19}$$

For each set N_i, we have

$$\mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\} \sim N_i}[||f(\mathbf{z}) - \mathbf{y}||_2] = \mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\} \sim N_i}[||f(\mathbf{z}) - f(\mathbf{z}_i) + f(\mathbf{z}_i) - \mathbf{y}_i + \mathbf{y}_i - \mathbf{y}||_2] \tag{20}$$

$$\leq \mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\} \sim N_i}[||f(\mathbf{z}) - f(\mathbf{z}_i)||_2] + \mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\} \sim N_i}[||f(\mathbf{z}_i) - \mathbf{y}_i||_2] + \mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\} \sim N_i}[||\mathbf{y}_i - \mathbf{y}||_2] \tag{21}$$

$$\leq L_1 \mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\} \sim N_i}[||\mathbf{z} - \mathbf{z}_i||_2] + \mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\} \sim N_i}[||f(\mathbf{z}_i) - \mathbf{y}_i||_2] + d_{max} \tag{22}$$

$$= L_1 \mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\} \sim N_i}[||\mathbf{z} - \bar{\mathbf{z}}_i + \bar{\mathbf{z}}_i - \mathbf{z}_i||_2] + \mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\} \sim N_i}[||f(\mathbf{z}_i) - \mathbf{y}_i||_2] + d_{max} \tag{23}$$

$$\leq L_1 \mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\} \sim N_i}[||\mathbf{z} - \bar{\mathbf{z}}_i||_2 + ||\bar{\mathbf{z}}_i - \mathbf{z}_i||_2] + \mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\} \sim N_i}[||f(\mathbf{z}_i) - \mathbf{y}_i||_2] + d_{max} \tag{24}$$

$$= L_1 \mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\} \sim N_i}[||\mathbf{z} - \bar{\mathbf{z}}_i||_2] + L_1 ||\bar{\mathbf{z}}_i - \mathbf{z}_i||_2 + \mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\} \sim N_i}[||f(\mathbf{z}_i) - \mathbf{y}_i||_2] + d_{max} \tag{25}$$

12

We denote the probability distribution over $\{N_i\}$ as P', where $P(N_i) = P(\{\mathbf{x}, \mathbf{z}, \mathbf{y}\} \in N_i)$. Then, we have

$$\mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\}\sim P}[||f(\mathbf{z}) - \mathbf{y}||_2] = \mathbb{E}_{N_i\sim P'}\mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\}\sim N_i}[||f(\mathbf{z}) - \mathbf{y}||_2] \tag{26}$$

$$\leq \mathbb{E}_{N_i\sim P'}[L_1\mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\}\sim N_i}[||\mathbf{z} - \bar{\mathbf{z}}_i||_2] + L_1||\bar{\mathbf{z}}_i - \mathbf{z}_i||_2 + \mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\}\sim N_i}[||f(\mathbf{z}_i) - \mathbf{y}_i||_2] + d_{max}] \tag{27}$$

$$= L_1\mathbb{E}_{N_i\sim P'}\mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\}\sim N_i}[||\mathbf{z} - \bar{\mathbf{z}}_i||_2] + L_1\mathbb{E}_{N_i\sim P'}||\bar{\mathbf{z}}_i - \mathbf{z}_i||_2 + \mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\}\sim S}(||f(\mathbf{z}_i) - \mathbf{y}_i||_2) + d_{max} \tag{28}$$

As $d_{max} \to 0$, we can approximate $\mathbb{E}_{N_i\sim P'}\mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\}\sim N_i}[||\mathbf{z} - \bar{\mathbf{z}}_i||_2]$ as $\mathbb{E}_{\mathbf{y}_i\sim\mathcal{Y}}\mathbb{E}_{\{(\mathbf{x},\mathbf{z},\mathbf{y})|\mathbf{y}=\mathbf{y}_i\}}[||\mathbf{z} - \bar{\mathbf{z}}_i||_2]$. Since $(\mathbf{Z}|\mathbf{Y} = \mathbf{y}_i) \sim \mathcal{D}$, we have $\mathcal{H}(\mathbf{Z}|\mathbf{Y}) = \mathbb{E}_{y\sim\mathcal{Y}}\mathcal{H}(\mathbf{Z}|\mathbf{Y} = y) = \mathcal{H}(\mathbf{Z}|\mathbf{Y} = \mathbf{y}_i) = \mathcal{H}(\mathbf{Z}|\mathbf{Y} = \mathbf{y}_j) = \mathcal{H}(\mathcal{D})$ for all $1 \leq i, j \leq N$, and $\mathbb{E}_{N_i\sim P'}||\mathbf{z}_i - \bar{\mathbf{z}}_i||_2$ can thus be approximate as $\mathbb{E}_{\{(\mathbf{x},\mathbf{z},\mathbf{y})|\mathbf{y}=\mathbf{y}_i\}}||\mathbf{z} - \bar{\mathbf{z}}_i||_2$. We have:

$$\mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\}\sim P}[||f(\mathbf{z}) - \mathbf{y}||_2] \tag{29}$$

$$\leq L_1\mathbb{E}_{N_i\sim P'}\mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\}\sim N_i}[||\mathbf{z} - \bar{\mathbf{z}}_i||_2] + L_1\mathbb{E}_{N_i\sim P'}||\bar{\mathbf{z}}_i - \mathbf{z}_i||_2 + \mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\}\sim S}(||f(\mathbf{z}_i) - \mathbf{y}_i||_2) + d_{max} \tag{30}$$

$$= L_1\mathbb{E}_{\mathbf{y}_i\sim\mathcal{Y}}\mathbb{E}_{\{(\mathbf{x},\mathbf{z},y)|y=\mathbf{y}_i\}}[||\mathbf{z} - \bar{\mathbf{z}}_i||_2] + L_1\mathbb{E}_{\{(\mathbf{x},\mathbf{z},y)|y=\mathbf{y}_i\}}||\mathbf{z}_i - \bar{\mathbf{z}}_i||_2 + \mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\}\sim S}(||f(\mathbf{z}_i) - \mathbf{y}_i||_2) \tag{31}$$

$$\leq L_1\mathbb{E}_{\mathbf{y}_i\sim\mathcal{Y}}[Q(\mathcal{H}(\mathbf{Z}|\mathbf{Y} = \mathbf{y}_i))] + L_1Q(\mathcal{H}(\mathbf{Z}|\mathbf{Y} = \mathbf{y}_i)) + \mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\}\sim S}(||f(\mathbf{z}_i) - \mathbf{y}_i||_2) \tag{32}$$

$$= 2L_1Q(\mathcal{H}(\mathbf{Z}|\mathbf{Y})) + \mathbb{E}_{\{\mathbf{x},\mathbf{z},\mathbf{y}\}\sim S}(||f(\mathbf{z}_i) - \mathbf{y}_i||_2) \tag{33}$$

$\square$

Proposition 1 *If $\mathcal{D}$ is a multivariate normal distribution $\mathcal{N}(\bar{\mathbf{z}}, \Sigma = k\mathbf{I})$, where $k > 0$ is a scalar and $\bar{\mathbf{z}}$ is the mean of the distribution $\mathcal{D}$. Then, the function $Q(\mathcal{H}(\mathcal{D}))$ in Theorem 2 can be selected as $Q(\mathcal{H}(\mathcal{D})) = \sqrt{\frac{d(e^{2\mathcal{H}(\mathcal{D})})^{\frac{1}{d}}}{2\pi e}}$, where d is the dimension of $\mathbf{z}$. If $\mathcal{D}$ is a uniform distribution, then the $Q(\mathcal{H}(\mathcal{D}))$ can be selected as $Q(\mathcal{H}(\mathcal{D})) = \frac{e^{\mathcal{H}(\mathcal{D})}}{\sqrt{12}}$.*

Proof   We first consider the case when $\mathcal{D} \sim \mathcal{N}(\bar{\mathbf{z}}, \Sigma = k\mathbf{I})$. Assume $\mathbf{Z} \sim \mathcal{N}(\bar{\mathbf{z}}, \Sigma)$, then $\mathcal{H}(\mathbf{Z}) = \frac{1}{2}\log(2\pi e)^n|\Sigma|$:

$$\mathcal{H}(\mathbf{Z}) = -\int_{\mathbf{z}} p(\mathbf{z})\log(p(\mathbf{z}))d\mathbf{z} \tag{34}$$

$$= -\int_{\mathbf{z}} p(\mathbf{z})\log\frac{1}{(\sqrt{2\pi})^d|\Sigma|^{\frac{1}{2}}}e^{\frac{-1}{2}(\mathbf{z}-\bar{\mathbf{z}})^\mathsf{T}\Sigma^{-1}(\mathbf{z}-\bar{\mathbf{z}})}d\mathbf{z} \tag{35}$$

$$= -\int_{\mathbf{z}} p(\mathbf{z})\log\frac{1}{(\sqrt{2\pi})^d|\Sigma|^{\frac{1}{2}}}d\mathbf{z} - \int_{\mathbf{z}} p(\mathbf{z})\log e^{\frac{1}{2}(\mathbf{z}-\bar{\mathbf{z}})^\mathsf{T}\Sigma^{-1}(\mathbf{z}-\bar{\mathbf{z}})}d\mathbf{z} \tag{36}$$

$$= \frac{1}{2}\log(2\pi)^d|\Sigma| + \frac{\log e}{2}\mathbb{E}[\sum_{i,j}(\mathbf{z}_i - \bar{\mathbf{z}}_i)(\Sigma^{-1})_{ij}(\mathbf{z}_j - \bar{\mathbf{z}}_j)] \tag{37}$$

$$= \frac{1}{2}\log(2\pi)^d|\Sigma| + \frac{\log e}{2}\mathbb{E}[\sum_{i,j}(\mathbf{z}_i - \bar{\mathbf{z}}_i)(\mathbf{z}_j - \bar{\mathbf{z}}_j)(\Sigma^{-1})_{ij}] \tag{38}$$

$$= \frac{1}{2}\log(2\pi)^d|\Sigma| + \frac{\log e}{2}\sum_{i,j}\mathbb{E}[(\mathbf{z}_i - \bar{\mathbf{z}}_i)(\mathbf{z}_j - \bar{\mathbf{z}}_j)](\Sigma^{-1})_{ij} \tag{39}$$

$$= \frac{1}{2}\log(2\pi)^d|\Sigma| + \frac{\log e}{2}\sum_j\sum_i\Sigma_{ji}(\Sigma^{-1})_{ij} \tag{40}$$

$$= \frac{1}{2}\log(2\pi)^d|\Sigma| + \frac{\log e}{2}\sum_j(\Sigma\Sigma^{-1})_j \tag{41}$$

$$= \frac{1}{2}\log(2\pi)^d|\Sigma| + \frac{\log e}{2}\sum_j I_{jj} \tag{42}$$

$$= \frac{1}{2}\log(2\pi)^d|\Sigma| + \frac{\log e}{2} \tag{43}$$

$$= \frac{1}{2}\log(2\pi e)^d|\Sigma| \tag{44}$$

We have the following:

$$\mathbb{E}[||\mathbf{z} - \bar{\mathbf{z}}||_2^2] = \text{tr}(\Sigma) = dk. \tag{45}$$

The following also holds:

$$|\Sigma| = k^d. \tag{46}$$

Thus, we have:

$$(\mathbb{E}[||\mathbf{z} - \bar{\mathbf{z}}||_2])^2 \leq \mathbb{E}[||\mathbf{z} - \bar{\mathbf{z}}||_2^2] = d|\Sigma|^{\frac{1}{d}} = d\left(\frac{e^{2\mathcal{H}(\mathbf{Z})}}{(2\pi e)^d}\right)^{\frac{1}{d}} = \frac{d(e^{2\mathcal{H}(\mathbf{Z})})^{\frac{1}{d}}}{2\pi e} \tag{47}$$

Finally,

$$\mathbb{E}[||\mathbf{z} - \bar{\mathbf{z}}||_2] \leq \sqrt{\frac{d(e^{2\mathcal{H}(\mathbf{Z})})^{\frac{1}{d}}}{2\pi e}} \tag{48}$$

Thus, $Q(\mathcal{H}(\mathcal{D}))$ in Theorem 2 can be selected as $Q(\mathcal{H}(\mathcal{D})) = \sqrt{\frac{d(e^{2\mathcal{H}(\mathcal{D})})^{\frac{1}{d}}}{2\pi e}}$, when $\mathcal{D} \sim \mathcal{N}(\bar{\mathbf{z}}, \Sigma = k\mathbf{I})$.

Similarly, if $\mathcal{D}$ is a uniform distribution $U(a, b)$, then its variance is given by:

$$\mathbb{E}[||\mathbf{z} - \bar{\mathbf{z}}||_2^2] = \frac{(b-a)^2}{12}, \tag{49}$$

and its entropy is given by:

$$\mathcal{H}(\mathbf{D}) = \log(b - a). \tag{50}$$

We have:

$$(\mathbb{E}[||\mathbf{z} - \bar{\mathbf{z}}||_2])^2 \leq \mathbb{E}[||\mathbf{z} - \bar{\mathbf{z}}||_2^2] = \frac{(b-a)^2}{12} = \frac{e^{2\mathcal{H}(\mathcal{D})}}{12} \tag{51}$$

Finally,

$$\mathbb{E}[||\mathbf{z} - \bar{\mathbf{z}}||_2] \leq \frac{e^{\mathcal{H}(\mathcal{D})}}{\sqrt{12}} \tag{52}$$

Thus, $Q(\mathcal{H}(\mathcal{D}))$ in Theorem 2 can be selected as $Q(\mathcal{H}(\mathcal{D})) = \frac{e^{\mathcal{H}(\mathcal{D})}}{\sqrt{12}}$, when $\mathcal{D}$ is a uniform distribution

$\square$

A.3. Proof of the Theorem 3

We first show a straightforward result of [(Ghosh & Motani, 2023), Proposition 1]:

Lemma 1 *Assume that $\mathbf{z}$ lies in a manifold $\mathcal{M}$ and the $\mathcal{M}_i \subset \mathcal{M}$ is a manifold corresponding to the distribution $(\mathbf{z}|\mathbf{y} = \mathbf{y}_i)$. Assume for all features $\mathbf{z}_i \in \mathcal{M}_i$, the following holds:*

$$\int_{||\mathbf{z} - \mathbf{z}_i|| \leq \epsilon} P(\mathbf{z}) d\mathbf{z} = C(\epsilon), \tag{53}$$

where $C(\epsilon)$ is some function of ϵ. The above imposes a constraint where the distribution $(\mathbf{z}|\mathbf{y} = \mathbf{y}_i)$ is uniformly distributed across $\mathcal{M}_i$. Then, as $\epsilon \to 0^+$, we have:

$$\mathcal{H}(\mathbf{Z}|\mathbf{Y}) = \mathbb{E}_{\mathbf{y}_i \sim \mathcal{Y}} \mathcal{H}(\mathbf{Z}|\mathbf{Y} = \mathbf{y}_i) = \mathbb{E}_{\mathbf{y}_i \sim \mathcal{Y}}[-\log(\epsilon)Dim_{ID}\mathcal{M}_i + \log\frac{K}{C(\epsilon)}], \tag{54}$$

for some fixed scalar K. $Dim_{ID}\mathcal{M}_i$ is the intrinsic dimension of the manifold $\mathcal{M}_i$.

14

Proof   By using the same proof technique as [(Ghosh & Motani, 2023), Proposition 1], we can show

$$\mathcal{H}(\mathbf{Z}|\mathbf{Y} = \mathbf{y}_i) = -\log(\epsilon)\text{Dim}_{\text{ID}}\mathcal{M}_i + \log\frac{K}{C(\epsilon)}, \tag{55}$$

Since $\mathcal{H}(\mathbf{Z}|\mathbf{Y}) = \mathbb{E}_{\mathbf{y}_i \sim \mathcal{Y}}\mathcal{H}(\mathbf{Z}|\mathbf{Y} = \mathbf{y}_i)$, the result follows. $\qquad\square$

Theorem 3   *Assume that $\mathbf{z}$ lies in a manifold $\mathcal{M}$ and the $\mathcal{M}_i \subset \mathcal{M}$ is a manifold corresponding to the distribution $(\mathbf{z}|\mathbf{y} = \mathbf{y}_i)$. Let $C(\epsilon)$ be some function of ϵ:*

$$C(\epsilon) = \int_{||\mathbf{z}-\mathbf{z}'||\leq\epsilon} P'(\mathbf{z})d\mathbf{z}, \tag{56}$$

where $P'(\mathbf{z})$ is the probability of $\mathbf{z}$ when $(\mathbf{z}|\mathbf{y} = \mathbf{y}_i)$ is uniformly distributed across $\mathcal{M}_i$, and $\mathbf{z}'$ is any point on $\mathcal{M}_i$. Then, as $\epsilon \to 0^+$, we have:

$$\mathcal{H}(\mathbf{Z}|\mathbf{Y}) = \mathbb{E}_{\mathbf{y}_i \sim \mathcal{Y}}\mathcal{H}(\mathbf{Z}|\mathbf{Y} = \mathbf{y}_i) \leq \mathbb{E}_{\mathbf{y}_i \sim \mathcal{Y}}[-\log(\epsilon)\text{Dim}_{ID}\mathcal{M}_i + \log\frac{K}{C(\epsilon)}], \tag{57}$$

for some fixed scalar K. $\text{Dim}_{ID}\mathcal{M}_i$ is the intrinsic dimension of the manifold $\mathcal{M}_i$.

Proof   Since the uniform distribution has the largest entropy over all distributions over the support $\mathcal{M}_i$, based on Lemma 1, we thus have:

$$\mathcal{H}(\mathbf{Z}|\mathbf{Y}) = \mathbb{E}_{\mathbf{y}_i \sim \mathcal{Y}}\mathcal{H}(\mathbf{Z}|\mathbf{Y} = \mathbf{y}_i) \leq \mathbb{E}_{\mathbf{y}_i \sim \mathcal{Y}}[-\log(\epsilon)\text{Dim}_{\text{ID}}\mathcal{M}_i + \log\frac{K}{C(\epsilon)}], \tag{58}$$

$$\square$$

A.4. Proof of the Proposition 2

Proposition 2   *Let the target $\mathbf{Y} = \mathbf{Y}' + \mathbf{N}'$ where $\mathbf{Y}'$ is fully determined by $\mathbf{X}$ and $\mathbf{N}'$ is the aleatoric uncertainty that is independent of $\mathbf{X}$. Assume the underlying mapping f' from $\mathbf{Z}$ to $\mathbf{Y}'$ and g' from $\mathbf{Y}'$ to $\mathbf{Z}$ are continuous, where the continuous mapping is based on the topology induced by the Euclidean distance. Then the representation $\mathbf{Z}$ is optimal if and only if $\mathbf{Z}$ is homeomorphic to $\mathbf{Y}'$.*

Proof   If $\mathbf{Z}$ is optimal (optimal $\mathbf{Z} \Rightarrow \mathbf{Z}$ is homeomorphic to $\mathbf{Y}'$):

$$\mathcal{H}(\mathbf{Y}|\mathbf{Z}) = \mathcal{H}(\mathbf{Y}' + \mathbf{N}'|\mathbf{Z}) = \mathcal{H}(\mathbf{Y}'|\mathbf{Z}) + \mathcal{H}(\mathbf{N}'|\mathbf{Z}, \mathbf{Y}') = \mathcal{H}(\mathbf{Y}'|\mathbf{Z}) + \mathcal{H}(\mathbf{N}'|\mathbf{Y}'), \tag{59}$$

$$\mathcal{H}(\mathbf{Y}|\mathbf{X}) = \mathcal{H}(\mathbf{Y}' + \mathbf{N}'|\mathbf{X}) = \mathcal{H}(\mathbf{Y}'|\mathbf{X}) + \mathcal{H}(\mathbf{N}'|\mathbf{X}, \mathbf{Y}') = \mathcal{H}(\mathbf{Y}'|\mathbf{X}) + \mathcal{H}(\mathbf{N}'|\mathbf{Y}'). \tag{60}$$

Since $\mathbf{Z}$ is optimal, we have $\mathcal{H}(\mathbf{Y}|\mathbf{Z}) = \mathcal{H}(\mathbf{Y}|\mathbf{X})$. Based on the two equations above, we have:

$$\mathcal{H}(\mathbf{Y}'|\mathbf{Z}) = \mathcal{H}(\mathbf{Y}'|\mathbf{X}). \tag{61}$$

Since $\mathbf{Y}'$ is fully determined by $\mathbf{X}$ and $\mathcal{H}(\mathbf{Y}'|\mathbf{Z}) = \mathcal{H}(\mathbf{Y}'|\mathbf{X})$, $\mathbf{Y}$ is also fully determined by $\mathbf{Z}$. Thus, for each $\mathbf{z}_i \in \mathbf{Z}$, there exists and only exists one $\mathbf{y}'_i \in \mathbf{Y}'$ corresponding to the $\mathbf{z}_i$, and thus the mapping function f' exists.

$\mathbf{Z}$ is optimal also means $\mathbf{Z}$ is fully determined given $\mathbf{Y}$, Since $\mathbf{N}'$ is independent of $\mathbf{Z}$:

$$\mathcal{H}(\mathbf{Z}|\mathbf{Y}) = \mathcal{H}(\mathbf{Z}|\mathbf{Y}' + \mathbf{N}') = \mathcal{H}(\mathbf{Z}|\mathbf{Y}'), \tag{62}$$

thus, for each $\mathbf{y}'_i$, there exist and only exist one $\mathbf{z}_i$ corresponding to the $\mathbf{y}'_i$. Thus, the mapping function f' is a bijection, and its inverse f'^{-1} is g'. Since f' and f'^{-1} are continuous, $\mathbf{Z}$ is homeomorphic to $\mathbf{Y}$.

If $\mathbf{Z}$ is homeomorphic to $\mathbf{Y}'$: ($\mathbf{Z}$ is homeomorphic to $\mathbf{Y}' \Rightarrow$ optimal $\mathbf{Z}$):

$\mathbf{Z}$ is homeomorphic to $\mathbf{Y}'$ means a continuous bijection exist between $\mathbf{Z}$ and $\mathbf{Y}'$, thus $\mathcal{H}(\mathbf{Z}|\mathbf{Y}) = \mathcal{H}(\mathbf{Z}|\mathbf{Y}') = \mathcal{H}(\mathbf{Y}'|\mathbf{Z})$ and $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ is minimal. We have:

$$\mathcal{H}(\mathbf{Y}|\mathbf{Z}) = \mathcal{H}(\mathbf{Y}'|\mathbf{Z}) + \mathcal{H}(\mathbf{N}'|\mathbf{Y}') = \mathcal{H}(\mathbf{N}'|\mathbf{Y}') = \mathcal{H}(\mathbf{Y}'|\mathbf{X}) + \mathcal{H}(\mathbf{N}'|\mathbf{Y}') = \mathcal{H}(\mathbf{Y}|\mathbf{X}), \tag{63}$$

thus, $\mathbf{Z}$ is optimal. $\qquad\square$

B. Discussions about Theorem 1

Choice of β in Theorem 1: β in $(0, 1)$ means we need to maximize $\mathcal{I}(\mathbf{Z}; \mathbf{Y})$ for sufficiency, while we want to minimize $\mathcal{I}(\mathbf{Z}; \mathbf{X})$ for minimality. When $\beta > 1$, then we value the minimality more than the sufficiency, resulting in the need to maximize $\mathcal{H}(\mathbf{Y}|\mathbf{Z})$. But, in the typical setting, we always value $\mathcal{I}(\mathbf{Z}; \mathbf{Y})$ more than $\mathcal{I}(\mathbf{Z}, \mathbf{X})$ for a good task-specific performance, and $\beta > 0$ will lead $\mathbf{Z}$ compressed to be a single point, as $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ is minimal and $\mathcal{H}(\mathbf{Y}|\mathbf{Z})$ is maximized in this case. The qualitative behavior should change based on $0 < \beta < 1$ or $\beta > 1$, as it controls which we value the more: sufficiency or minimality.

Difference between the target $\mathbf{Y}$ and the predicted $\mathbf{Y}'$: The target $\mathbf{Y}$ is different from the predicted $\mathbf{Y}'$. $\mathcal{H}(\mathbf{Y}'|\mathbf{Z})$ always equals to 0 if we exploit neural networks as deterministic functions. In an extreme case, we can treat the predicted $\mathbf{Y}'$ as the representation $\mathbf{Z}$, which shows minimizing $\mathcal{H}(\mathbf{Y}'|\mathbf{Y}) = 0$ is the learning target. From the invariance representation learning point of view, lowering $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ is learning invariance representations with respect to $\mathbf{Y}$.

More discussions about the assumptions: For discrete entropy, $\mathbf{Z}$ is determined given $\mathbf{Y}$ implies $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ is a constant for $\mathbf{Z}$, as $\mathcal{H}(\mathbf{Z}|\mathbf{Y}) = 0$ here. However, this does not hold for differential entropy. For differential entropy, $\mathbf{Z}$ is determined given $\mathbf{Y}$ means $\mathcal{H}(\mathbf{Z}|\mathbf{Y}) = -\infty$, since given Y, Z is distributed as a delta function in this case.

C. Connections with the bound in Kawaguchi et al. (2023)

Kawaguchi et al. (2023) provide several bounds that are all applicable for both classification and regression for various cases. In the case where the encoder model ϕ (whose output is z) and the training dataset of the downstream task $\hat{S} = \{x_i, y_i\}_{i=1}^N$ are dependent (e.g., this is when $z = \phi(x)$ for $x \notin S$ is dependent of all N training data points $(x_i, y_i)_{i=1}^N$ through the training of ϕ by using $\hat{S}$), they show that any valid and general generalization bound of the information bottleneck must include two terms, $I(\mathbf{X}; \mathbf{Z}|\mathbf{Y})$ and $I(\phi; \hat{S})$, where the second term measures the effect of overfitting the encoder ϕ. This is because the encoder ϕ can compress all information to minimize $I(\mathbf{X}; \mathbf{Z}|\mathbf{Y})$ arbitrarily well while overfitting to the training data: e.g., we can simply set $\phi(x_i) = y_i$ for all $(x_i, y_i) \in \hat{S}$ and $\phi(x) = c \neq y$ for all $(x, y) \notin \hat{S}$ for some constant c to achieve the best training loss while minimizing $I(\mathbf{X}; \mathbf{Z}|\mathbf{Y})$ and performing arbitrarily poorly for test loss. Given this observation, they prove the first rigorous generalization bounds for two separate cases based on the dependence of ϕ and $\hat{S}$. Their generalization bounds scales with $I(\mathbf{X}; \mathbf{Z}|\mathbf{Y})$ without the second term $I(\phi; \hat{S})$ in case of ϕ and $\hat{S}$ being independent, and with $I(\mathbf{X}; \mathbf{Z}|\mathbf{Y})$ and $I(\phi; \hat{S})$ in case of ϕ and $\hat{S}$ being dependent.

In Theorem 2, we consider the case where ϕ and $\hat{S}$ are independent, since z in $(x, z, y) \sim P$ is drawn without dependence on the entire N data points $\{x_i, y_i\}_{i=1}^N$ in equation (18). Thus, our results are consistent and do not contradict previous findings. Unlike the previous bounds, our bound is determined by the function Q, which characterizes the dispersion or the standard deviation of a distribution by its entropy. The function Q exists for general cases, as the dispersion or the standard deviation and the entropy commonly can be estimated for a specific distribution. We thus can find a function Q to upper bound the relationship on the entropy and its dispersion or the standard deviation.

It is worth mentioning that we are not targeting a tight or an advanced bound here. Our bound is introduced to support the analysis that follows after Theorem 2, which is challenging with the previous bounds.

D. From Proposition 2, whether the optimal representation $\mathbf{Z}$ is the one that equals the ground-truth label?

Such $\mathbf{Z}$ can be one of the optimal/best representations under the negligible aleatoric uncertainty setting. However, $\mathbf{Z}$ is not unique and Proposition 2 is broader than this statement as it says that the optimal representation $\mathbf{Z}$ should be homeomorphic to the ground truth - i.e. they only need to be topologically equivalent. In this regard, the feature space may resemble a square, while the target space is a circle.

A practical benefit of Proposition 2 is its guidance on the desirable topological properties of $\mathbf{Z}$. For example, if the target space is a single connected component (i.e. $\beta_0 = 1$), then the feature space should also be similar; this does not hold in general (the task-specific loss alone cannot guarantee a single connected feature space and the topology of the feature space is also influenced by the input space. In addition, we observe empirically on depth estimation that the feature space sometimes consists of multiple disconnected components).

Table 6. Depth estimation on NYU-Depth-V2. Here, Random represents the encoder is fixed in a random state, while PH-Reg means we first train the encoder purely with PH-Reg for 1 epoch, then fix it and train the regressor.

Encoder	Regressor	$\delta_1 \uparrow$	REL $\downarrow$	RMS $\downarrow$	$\log_{10} \downarrow$
Random	Linear	0.398	0.390	1.144	0.153
PH-Reg	Linear	0.428	0.391	1.043	0.153
Random	NonLinear	0.412	0.381	1.121	0.149
PH-Rege	NonLinear	0.440	0.374	1.052	0.141

E. Discussions about the regressor

1. Do the appropriate properties (i.e., lower ID and homeomorphism) depend on the regressor?

Although the appropriate feature space may vary from regressor to regressor, the appropriate **properties**, as supported by our theorems, do not depend on the regressor. Specifically, our Theorem 1 shows that the IB tradeoff is fully determined by the values of $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ and $\mathcal{H}(\mathbf{Y}|\mathbf{Z})$. These values of the entropy terms do not depend on the regressor that maps $\mathbf{Z}$ to the predicted $\hat{\mathbf{Y}}$. In addition, for depth estimation, representations learned **purely** by PH-Reg without any other loss terms are also highly competitive (see Table 6).

2. Topology regularization with "simple" vs highly expressive regressors

For both regressors, our regularizer leads to representations with a higher signal-to-noise ratio (since $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ and $\mathcal{H}(\mathbf{Z}|\mathbf{Y})$ are minimized). This should make it easier for the regressor to estimate the true underlying signal. However, more expressive regressors have a higher capacity to estimate the underlying signal directly, so the room for improvement from the regularization is reduced. In the extreme case of too much expressiveness, our regularizer may again lead to improvements as it may limit overfitting, by minimizing noise in the learned representation.

The topology properties supported by our theorems align with invariant representation learning (i.e. invariance to noise), where invariance serves as a general prior for desirable representation properties (Bengio et al., 2013).

F. Preliminaries on Topology

The simplicial complex is a central object in topological data analysis, and it can be exploited as a tool to model the 'shape' of data. Given a set of finite samples $\mathbf{S} = \{s_i\}$, the simplicial complex K can be seen as a collection of simplices $\sigma = \{s_0, \cdots, s_k\}$ of varying dimensions: vertices ($|\sigma| = 1$), edges($|\sigma| = 2$), and the higher-dimensional counterparts($|\sigma| > 2$). The faces of a simplex $\sigma = \{s_0, \cdots, s_k\}$ is the simplex spanned by the subset of $\{s_0, \cdots, s_k\}$. The dimension of the simplicial complex K is the largest dimension of its simplices. A simplicial complex can be regarded as a high-dimensional generalization of a graph, and a graph can be seen as a 1-dimensional simplicial complex. For each $\mathbf{S}$, there exist many ways to build simplicial complexes and the Vietoris-Rips Complexes are widely used:

Definition 3 *(Vietoris-Rips Complexes). Given a set of finite samples $\mathbf{S}$ sampled from the feature space or target space and a threshold $\alpha \geq 0$, the Vietoris-Rips Complexes VR_α is defined as:*

$$VR_\alpha(\mathbf{S}) = \{\{s_0, \cdots, s_k\}, s \in \mathbf{S}|d(s_i, s_j) \leq \alpha\}, \tag{64}$$

where $d(s_i, s_j)$ is the Euclidean distance between samples s_i and s_j.

$VR_\alpha(\mathbf{S})$ is the set of all simplicial complexes $\{s_0, \cdots, s_k\}$ where the pairwise distance $d(s_i, s_j)$ is within the threshold α. Let $C_k(VR_\alpha(\mathbf{S}))$ denote the vector space generated by its k-dimensional simplices over $\mathbb{Z}_2$[2]. The boundary operator $\partial_k : C_k(VR_\alpha(\mathbf{S})) \to C_{k-1}(VR_\alpha(\mathbf{S}))$ maps each simplex to its boundary , which consists of the sum of all its faces, is a homomorphism from $C_k(VR_\alpha(\mathbf{S}))$ to $C_{k-1}(VR_\alpha(\mathbf{S}))$. It can be shown that $\partial_k \circ \partial_{k-1} = 0$, which leads to the chain complex: $\cdots \to$, and the kth homology group $H_k(VR_\alpha(\mathbf{S}))$ is defined as the quotient group $H_k(VR_\alpha(\mathbf{S})) := \ker\partial_k/\text{im}\partial_{k+1}$. ker represents kernel, which is the set of all elements that are mapped to the zero element. im represents image, which is the set of all the outputs. Rank $H_k(VR_\alpha(\mathbf{S}))$ is known as the kth Betti number β_k, which counts the number of k-dimensional holes and can be used to represent the topological features of the manifold that the set of points $\mathbf{S}$ sampled from.

---

[2]It is not specific to $\mathbb{Z}_2$, but $\mathbb{Z}_2$ is a typical choice.

(a) Illustration of the framework

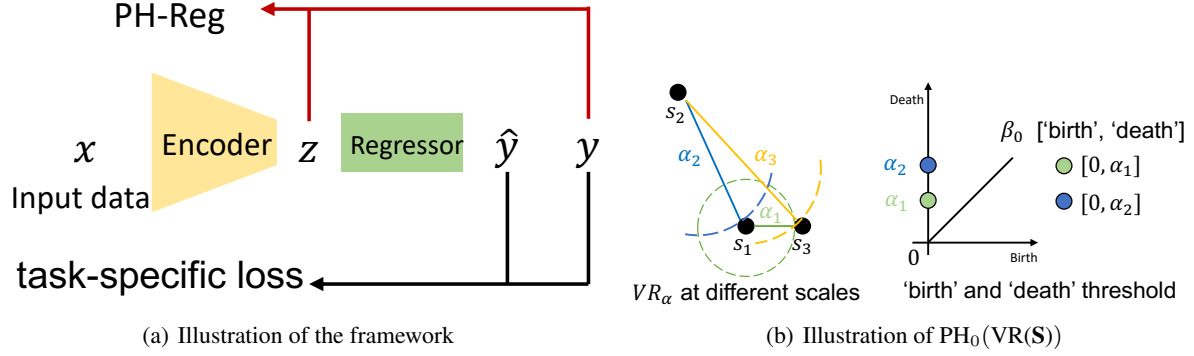(b) Illustration of $\mathrm{PH}_0(\mathrm{VR}(\mathbf{S}))$

Figure 5. Illustration of the (a) the use of PH-Reg for regression, and (b) calculating of $\mathrm{PH}_0(\mathrm{VR}(\mathbf{S}))$. Here $\mathbf{S} = \{s_1, s_2, s_3\}$. We say three connected components, i.e. β_0, $(\{\{s_1\}, \{s_2\}, \{s_3\}\})$ 'birth' when $\alpha = 0$, one 'death' (two left $(\{\{s_1, s_3\}, s_2\})$) when $\alpha = \alpha_1$, and another one 'death' (one left $(\{\{s_1, s_3, s_2\})$) when $\alpha = \alpha_2$. Thus $\mathrm{PH}_0(\mathrm{VR}(\mathbf{S})) = \{[0, \alpha_1], [0, \alpha_2]\}$.

However, the $H_k(\mathrm{VR}_\alpha(\mathbf{S}))$ is obtained based on a single α, which is easily affected by small changes in $\mathbf{S}$. Thus, it is not robust and is of limited use for real-world datasets. The persistent homology considers all the possible α instead of a single one, which results in a sequence of β_k. This is achieved through a nested sequence of simplicial complexes, called *filtration*: $\mathrm{VR}_0(\mathbf{S}) \subseteq \mathrm{VR}_{\alpha_1}(\mathbf{S}) \subseteq \cdots \subseteq \mathrm{VR}_{\alpha_m}(\mathbf{S})$ for $0 \leq \alpha_1 \leq \alpha_m$. Let $\gamma_i = [\alpha_i, \alpha_j]$ be the interval corresponding to a k-dimensional hole 'birth' at the threshold α_i and 'death' at the threshold α_j, we denote $\mathrm{PH}_k(\mathrm{VR}(\mathbf{S})) = \{\gamma_i\}$ the set of 'birth' and 'death' intervals of the k-dimensional holes. We only exploit $\mathrm{PH}_0(\mathrm{VR}(\mathbf{S}))$ in our PH-Reg, since we exploit the topological autoencoder as the topology part and higher topological features merely increase its runtime. An illustration of the calculation $\mathrm{PH}_0(\mathrm{VR}(\mathbf{S}))$ is given in Figure 5(b). We define $E(\mathbf{S}) = \sum_{\gamma \in \mathrm{PH}_0(\mathrm{VR}(\mathbf{S}))} |I(\gamma)|$, where $|I(\gamma)|$ is the length of the interval γ.

G. More details about the Coordinate Prediction task

Details about the synthetic dataset: We encode coordinates $\mathbf{y} \in \mathbb{R}^3$ into 100 dimensional vectors $\mathbf{x}_i = [f_1(\mathbf{y}_i), f_2(\mathbf{y}_i), f_3(\mathbf{y}_i), f_4(\mathbf{y}_i), \text{noise}]$, where the dimensions $1-4$ are signal and the rest 96 dimensions are noise. The encoder functions f_i are defined as:

- $f_1(\mathbf{y}_i) = y_{i_1} + y_{i_2} + y_{i_3}$

- $f_2(\mathbf{y}_i) = y_{i_1} + y_{i_2} - y_{i_3}$

- $f_3(\mathbf{y}_i) = y_{i_1} - y_{i_2} + y_{i_3}$

- $f_4(\mathbf{y}_i) = -y_{i_1} + y_{i_2} + y_{i_3}$

As shown above, the accurate coordinates $\mathbf{y}_i$ can be obtained correctly when $f_1(\mathbf{y}_i), f_2(\mathbf{y}_i), f_3(\mathbf{y}_i), f_4(\mathbf{y}_i)$ are given. We introduce noise to the remaining 96 dimensions by using f_1, f_2, f_3, f_4 on other randomly selected samples $\mathbf{y}_j$. The proximity of $\mathbf{y}_j$ to $\mathbf{y}_i$ can be intuitively seen as an indicator of the noise's relationship to the signal.

More training details: We train the models for 10000 epochs using AdamW as the optimizer with a learning rate of 0.001. We report results as mean $\pm$ standard variance over 10 runs. For the regression baseline Oridnal Entropy and the IB baseline Information Dropout, we tried various weights $\{0.01, 0.1, 1, 10\}$ and reported the best results. The trade-off parameters λ_d and λ_t are default set to 10 and 100, respectively, while λ_t is set to 10000 and λ_d is set to 1 for Mammoth, and λ_d is set to 1 for torus and circle.

Table 7. Results on AgeDB. We report results as mean ± standard variance over 3 runs. **Bold** numbers indicate the best performance.

Method	MAE ↓				GM ↓			
	ALL	Many	Med.	Few	ALL	Many	Med.	Few
Baseline (Yang et al., 2021)	7.80 ± 0.12	6.80 ± 0.06	9.11 ± 0.31	13.63 ± 0.43	4.98 ± 0.05	4.32 ± 0.06	6.19 ± 0.07	10.29 ± 0.57
+ Information dropout (Achille & Soatto, 2018b)	8.04 ± 0.14	7.14 ± 0.20	9.10 ± 0.71	13.61 ± 0.32	5.11 ± 0.06	4.49 ± 0.17	6.14 ± 0.49	10.54 ± 0.65
+ Oridnal Entropy (Zhang et al., 2023)	7.65 ± 0.13	6.72 ± 0.09	8.77 ± 0.49	13.28 ± 0.73	4.91 ± 0.14	4.29 ± 0.06	6.04 ± 0.51	10.09 ± 0.62
$+\mathcal{L}_d'$	7.75 ± 0.05	6.80 ± 0.11	8.87 ± 0.05	13.61 ± 0.50	4.96 ± 0.04	4.33 ± 0.09	6.05 ± 0.36	10.43 ± 0.40
$+\mathcal{L}_d$	7.64 ± 0.07	6.82 ± 0.07	8.62 ± 0.20	12.79 ± 0.65	4.85 ± 0.05	4.27 ± 0.06	5.91 ± 0.13	9.75 ± 0.53
$+\mathcal{L}_t$	7.50 ± 0.04	6.59 ± 0.03	8.75 ± 0.03	12.67 ± 0.24	4.77 ± 0.07	4.27 ± 0.06	6.09 ± 0.03	9.34 ± 0.70
$+\mathcal{L}_d + \mathcal{L}_t$	$\mathbf{7.32 \pm 0.09}$	$\mathbf{6.50 \pm 0.15}$	$\mathbf{8.38 \pm 0.11}$	$\mathbf{12.18 \pm 0.38}$	$\mathbf{4.69 \pm 0.07}$	$\mathbf{4.15 \pm 0.08}$	$\mathbf{5.64 \pm 0.09}$	$\mathbf{8.99 \pm 0.38}$

Table 8. Quantitative comparison of super-resolution results with public benchmark and DIV2K validation set. We report results as PSNR(dB)/SSIM. **Bold** numbers indicate the best performance.

Method	Set5	Set14	B100	Urban100	DIV2K
Baseline (Lim et al., 2017)	32.241/ 0.8656	28.614/ 0.7445	27.598/ 0.7120	26.083/ 0.7645	28.997/ 0.8189
+ Information dropout (Achille & Soatto, 2018b)	32.219/ 0.8649	28.626/ 0.7441	27.594/ 0.7113	26.059/ 0.7624	28.980/ 0.8182
+ Oridnal Entropy (Zhang et al., 2023)	32.280/ 0.8653	28.659/ 0.7445	27.614/ 0.7119	26.117/ 0.7641	29.005/ 0.8188
$+\mathcal{L}_d'$	32.252/ 0.8653	28.625/ 0.7443	27.599/ 0.7118	26.078/ 0.7638	28.989/ 0.8186
$+\mathcal{L}_d$	32.293/ 0.8660	28.644/ 0.7453	27.619/ 0.7127	26.151/ 0.7662	29.022/0.8197
$+\mathcal{L}_t$	**32.322/ 0.8663**	28.673/ 0.7455	27.624/ 0.7127	26.169/ 0.7665	29.031/ 0.8196
$+\mathcal{L}_d + \mathcal{L}_t$	32.288/ **0.8663**	**28.686/ 0.7462**	**27.627/ 0.7132**	**26.179/ 0.7670**	**29.038/ 0.8201**

H. Details about the real-world tasks

H.1. Evaluation metrics

Depth Estimation. We denote the predicted depth at position p as y_p and the corresponding ground truth depth as y_p', the total number of pixels is n. The metrics are: 1) threshold accuracy $\delta_1 \triangleq \%$ of y_p, s.t. $\max(\frac{y_p}{y_p'}, \frac{y_p'}{y_p}) < t_1$, where $t_1 = 1.25$; 2) average relative error (REL): $\frac{1}{n} \sum_p \frac{|y_p - y_p'|}{y_p}$; 3) root mean squared error (RMS): $\sqrt{\frac{1}{n} \sum_p (y_p - y_p')^2}$; 4) average ($\log_{10}$ error): $\frac{1}{n} \sum_p |\log_{10}(y_p) - \log_{10}(y_p')|$.

Age Estimation. Given N images for testing, y_i and y_i' are the i-th prediction and ground-truth, respectively. The evaluation metrics include 1)MAE: $\frac{1}{N} \sum_{i=1}^{N} |y_i - y_i'|$, and 2)Geometric Mean (GM): $(\prod_{i=1}^{N} |y_i - y_i'|)^{\frac{1}{N}}$.

H.2. Age estimation on AgeDB-DIR dataset

We exploit the AgeDB-DIR (Yang et al., 2021) for age estimation task. We follow the setting of Yang et al. (2021) and exploit their regression baseline model, which uses ResNet-50 (He et al., 2016) as the backbone. The evaluation metrics are MAE and geometric mean(GM), and the results are reported on the whole set and the three disjoint subsets, i.e. Many, Med. and Few. The trade-off parameters λ_d and λ_t are set to 0.1 and 1, respectively. Table 7 shows that both $\mathcal{L}_t$ and $\mathcal{L}_d$ can improve the performance, and combining both achieves 0.48 overall improvements (i.e. ALL) on MAE and 0.25 overall improvements on GM.

H.3. Super-resolution on DIV2K dataset

We use the DIV2K dataset (Timofte et al., 2017) for 4x super-resolution training (without the 2x pretrained model) and we evaluate on the validation set of DIV2K and the standard benchmarks: Set5 (Bevilacqua et al., 2012), Set14 (Zeyde et al., 2012), BSD100 (Martin et al., 2001), Urban100 (Huang et al., 2015). We follow the setting of Lim et al. (2017) and exploit their small-size EDSR model as our baseline architecture. We adopt the standard metrics PNSR and SSIM. The trade-off parameters λ_d and λ_t are set to 0.1 and 1, respectively. Table 3 shows that both $\mathcal{L}_d$ and $\mathcal{L}_t$ contribute to improving the baseline and adding both terms has the largest impact.

H.4. Depth estimation on NYU-Depth-v2 dataset

We exploit the NYU-Depth-v2 (Silberman et al., 2012) for the depth estimation task. We follow the setting of Lee et al. (2019) and use ResNet50 as our baseline architecture. We exploit the standard metrics of threshold accuracy $\delta_1, \delta_2, \delta_3$, average relative error (REL), root mean squared error (RMS) and average $\log_{10}$ error. The trade-off parameters λ_d and λ_t

Table 9. Quantitative comparison (MAE) on AgeDB. We report results as mean $\pm$ standard variance over 3 runs.

Method	ALL	Many	Med.	Few
Baseline	7.80 ± 0.12	6.80 ± 0.06	9.11 ± 0.31	13.63 ± 0.43
+ LDS (Yang et al., 2021)	7.67	6.98	8.86	10.89
+ FDS (Yang et al., 2021)	7.55	6.50	8.97	13.01
+ LDS + FDS (Yang et al., 2021)	7.55	7.01	8.24	10.79
+ PH-Reg (ours)	7.32 ± 0.09	6.50 ± 0.15	8.38 ± 0.11	12.18 ± 0.38
+ LDS + FDS vs. Baseline	+ 0.25	-0.19	+0.97	+2.94
+ ous vs. Baseline	+ 0.48	+0.30	+0.73	+1.45

Table 10. Quantitative comparison (PSNR(dB)) of super-resolution results with public benchmark and DIV2K validation set. **Bold** numbers indicate the best performance.

Method	Set5	Set14	B100	Urban100	DIV2K
Baseline (small-size EDSR (Lim et al., 2017))	32.24	28.61	27.60	26.08	29.00
EDSR (Lim et al., 2017)	32.46	28.80	27.71	26.64	29.25
RDN (Zhang et al., 2018)	32.47	28.81	27.72	26.61	-
Baseline + PH-Reg (ours)	32.29	28.69	27.63	26.18	29.04
RDN vs. EDSR	+0.01	+0.01	+0.01	-0.03	-
Ours vs. Baseline	+0.05	+0.08	+0.03	+0.01	+0.04

are both set to 0.1. Table 4 shows that exploiting $\mathcal{L}_t$ and $\mathcal{L}_d$ results in reduction of 6.7% and 8.9% in the δ_1 and δ_2 errors, respectively.

H.5. Different improvement gap between synthetic and real-world datasets

Our synthetic dataset is relatively simple and clean; the corresponding task (coordinate prediction) is directly related to the topology of the target space, hence the larger improvement on the synthetic data. The improvements on real-world datasets are significant (verified by Welch's t-test), and are also comparable to or better than competing works in the literature (Zhang et al., 2023; Yang et al., 2021; Lim et al., 2017; Zhang et al., 2018). Under the same settings, our improvements are competitive with recently published works. For age estimation on Age-DB, we improve the MAE (all) by almost 2-fold (0.48 vs. 0.25, see Table 9). For super-resolution on DIV2K, we improve the PSNR (Set5, see Table 10) by 0.05, while typical state-of-the-art papers in super-resolution show increments of 0.01 on PSNR (Set5) (Lim et al., 2017; Zhang et al., 2018). For depth estimation, our improvements are on par with the competing method (Zhang et al., 2023) (REL: 0.144 vs 0.143, see Table 4).

However, on NYU-Depth-V2, the representation of the head part (with many samples in the target space) is relatively well-learned, leading to a smaller impact on our regularizers. This might be a reason for the lower improvement. From the regression baseline (Figure 2(a)), the representation in the blue region (head part) already shows a lower intrinsic dimension and collapses like a line so there is little opportunity for our regularizers to have a strong impact (and hence the smaller improvement on the MAE (Many)). In contrast, the impact on the synthetic dataset and the green region (corresponding to the tail part, with limited samples in the target space) is more significant, resulting in a higher improvement.

Similar effects are observed on the other two real-world datasets, while the feature space on Age-DB tends to be a line (although the target space is discrete, it is too dense) and the feature space on DIV2K tends to be an object with a high density in the middle region (the target space is 3d).