Universal Rate-Distortion-Classification Representations for Lossy Compression

Nam Nguyen, *Graduate Student Member, IEEE,* Thuan Nguyen, *Member, IEEE,* Thinh Nguyen, *Senior Member, IEEE*, and Bella Bose, *Fellow, IEEE*

Abstract

In lossy compression, Blau and Michaeli [2] introduced the information rate-distortion-perception (RDP) function, extending traditional rate-distortion theory by incorporating perceptual quality. More recently, this framework was expanded by defining the rate-distortion-perception-classification (RDPC) function, integrating multi-task learning that jointly optimizes generative tasks such as perceptual quality and classification accuracy alongside reconstruction tasks [3]. To that end, motivated by the concept of a universal RDP encoder introduced in [4], we investigate universal representations that enable diverse distortion-classification tradeoffs through a single fixed encoder combined with multiple decoders. Specifically, theoretical analysis and numerical experiments demonstrate that for the Gaussian source under mean squared error (MSE) distortion, the entire distortion-classification tradeoff region can be achieved using one universal encoder. In addition, this paper characterizes achievable distortion-classification regions for fixed universal representations in general source distributions, identifying conditions that ensure minimal distortion penalty when reusing encoders across varying tradeoff points. Experimental results using MNIST and SVHN datasets validate our theoretical insights, showing that universal encoders can obtain distortion performance comparable to task-specific encoders, thus supporting the practicality and effectiveness of our proposed universal representations.

Index Terms

Lossy compression, rate-distortion-classification tradeoff, universal encoder representations, image compression, deep learning.

This work was presented in part at the IEEE Information Theory Workshop (ITW), September 2025 [1]. This work was supported by the National Science Foundation Grant CCF2417898. *(Corresponding author: Nam Nguyen.)*

Nam Nguyen, Thinh Nguyen and Bella Bose are with the School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, US (e-mail: {nguynam4, thinhq, bella.bose}@oregonstate.edu) .

Thuan Nguyen is with the Department of Engineering, Engineering Technology, and Surveying, East Tennessee State University, Johnson City, TN, USA (email: nguyent11@etsu.edu).

I. Introduction

Rate-distortion theory has long served as the foundation for lossy compression, characterizing the minimum distortion achievable at a given bit rate [5]. Conventional systems are typically evaluated using full-reference distortion metrics such as Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Multi-Scale Structural Similarity Index (MS-SSIM) [6], [7]. However, recent research has demonstrated that minimizing distortion alone is insufficient to yield perceptually convincing reconstructions. This limitation is particularly evident in deep learning (DL)-based image compression, where empirical evidence suggests that improvements in perceptual quality often come at the expense of increased distortion [8], [9].

To address this limitation, Blau and Michaeli [2] introduced the rate-distortion-perception (RDP) framework, which incorporates perceptual quality, measured via distributional divergence, as an independent optimization axis. The RDP formulation reveals a fundamental tradeoff among rate, distortion fidelity, and perceptual realism. Practical implementations, particularly those using GANs [10], have demonstrated high perceptual quality at low bit rates [11]–[14]. Common no-reference perceptual metrics include Fréchet Inception Distance (FID) [15], Naturalness Image Quality Evaluator (NIQE), Perception-based Image Quality Evaluator (PIQE), and Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [16]–[18].

In the context of signal restoration, the authors in [19] pioneered the extension of the perception-distortion tradeoff by introducing the classification-distortion-perception (CDP) framework. Specifically, they incorporated the classification error rate of the restored signal as a third dimension, complementing distortion and perceptual quality. Their work rigorously established the inherent tradeoff within the CDP framework, demonstrating that it is impossible to simultaneously minimize distortion, perceptual difference, and classification error rate.

Further extending this perspective, recent studies have integrated classification tasks into lossy compression frameworks, enabling multi-task optimization that bridges the gap between image compression and visual analysis [3], [20]. Initially proposed by Zhang et al. [20], the rate-distortion-classification (RDC) model established a unified framework for optimizing the tradeoff among rate, distortion, and classification accuracy in lossy image compression. Through extensive statistical analysis of multi-distribution sources, the RDC model was shown to exhibit favorable theoretical properties, including monotonic non-increasing behavior and convexity under specific

conditions. Building upon this foundation, Wang et al. [3] formalized the concept further by introducing the rate-distortion-perception-classification (RDPC) function. Their results rigorously demonstrated inherent tradeoffs within the RDPC framework, revealing that enhancements in classification accuracy generally incur higher distortion or reduced perceptual quality.

Given these intricate tradeoffs in lossy compression scenarios, a fundamental question emerges: are these tradeoffs inherently determined by the encoder's chosen representation, or can a single encoder representation adapt to multiple objectives via different decoding strategies? To investigate this, the concept of a universal RDP framework was introduced in [4], utilizing a single fixed encoder in conjunction with multiple decoders to achieve diverse points within the distortion-perception space without retraining the encoder.

In this paper, motivated by this universal approach, we investigate universal representations that enable diverse distortion-classification tradeoffs through a single fixed encoder combined with multiple decoders. Specifically, theoretical analysis and numerical experiment demonstrate that for the Gaussian source under MSE distortion, the entire distortion-classification tradeoff region can be achieved using one universal encoder. In addition, this paper characterizes achievable distortion-classification regions for fixed universal representations in general source distributions, identifying conditions that ensure minimal distortion penalty when reusing encoders across varying tradeoff points. Experimental results using MNIST and SVHN datasets validate our theoretical insights, showing that universal encoders can obtain distortion performance comparable to task-specific encoders, thus supporting the practicality and effectiveness of our proposed universal representations.

II. Related Work

Lossy compression has traditionally been studied through the framework of rate-distortion theory [5], which characterizes the minimum bit rate required to encode a source within a specified distortion level. Classical information theory has also explored distribution-preserving lossy compression, aiming to maintain statistical properties of the source in the reconstruction [21]–[23]. Recent advances in generative modeling [24] have reignited interest in these foundations, particularly in the context of machine learning and representation learning. Modern frameworks increasingly leverage rate-distortion principles to learn compact, information-constrained representations [25]–[27].

Within the rate-distortion-perception framework, distortion is typically evaluated using full-reference metrics that compare the reconstructed signal with the original. These include mean squared error (MSE), peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM) [6]. In contrast, perceptual quality is assessed using no-reference metrics, which rely on the statistical properties of the output alone. Examples include Fréchet Inception Distance (FID) [15], Perception-based Image Quality Evaluator (PIQE) [18], Naturalness Image Quality Evaluator (NIQE) [17], and Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [16].

The emergence of generative adversarial networks (GANs) has significantly advanced the field of perceptual compression. GAN-based models can produce visually realistic reconstructions, motivating the use of trained discriminators as perceptual quality evaluators [14]. Theoretically, this is supported by the correspondence between GAN objectives and statistical divergence measures [11], [13], [28]. Building on these insights, several works have integrated GAN-based regularization into compressive autoencoder frameworks [2], [8], enabling high perceptual quality even at extremely low bitrates [29].

Blau and Michaeli [9] first formalized the perception-distortion tradeoff, which was later extended into the RDP framework [2]. These contributions inspired a broader class of models for distribution-preserving lossy compression [12], underscoring the necessity of jointly optimizing for perceptual realism and distortion fidelity.

Further extending this perspective, the classification-distortion-perception framework [19] introduced classification accuracy as an additional optimization axis. This formulation revealed that perceptual quality, distortion, and classification performance are fundamentally at odds: improving one often degrades the others. In multi-task learning settings, recent work [3], [20] proposed the rate-distortion-classification and rate-distortion-perception-classification functions, formalizing the joint optimization of these competing objectives. Their analyses revealed structural properties such as convexity and monotonicity of the tradeoff regions and empirically confirmed that improving classification accuracy typically incurs higher distortion or reduced perceptual quality.

To alleviate the need for retraining encoders for every task-specific operating point, the concept of universal RDP representations was proposed in [4]. This framework fixes the encoder and trains multiple decoders to support diverse perceptual-distortion tradeoffs. The authors proved that such universal representations are approximately optimal under certain conditions, offering practical advantages for multi-objective compression without sacrificing performance.

III. Rate-Distortion-Classification Representations

Consider a source generating observable data $X \sim p_X$, which inherently contains multiple target labels represented by variables $S_1, \ldots, S_K \sim p_{S_1,\ldots,S_K}$. The observable data X and these intrinsic variables are correlated, following a joint probability distribution $p_{X,S_1,\ldots,S_K}$ over the space $\mathcal{X} \times \mathcal{S}_1 \times \cdots \times \mathcal{S}_K$. While the target variables are not directly observable, they can be inferred from X. For example, if X is an image, classification tasks can be object recognition or scene understanding. As illustrated in Fig. 1, the lossy compression process consists of

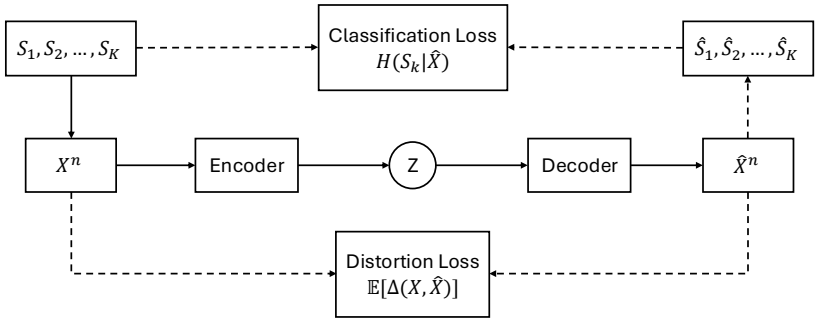


Figure 1: Illustration of task-oriented lossy compression framework.

an encoder and a decoder. Consider a source that generates an independent and identically distributed (i.i.d.) sequence $X_1, X_2, \ldots, X_n \sim p_X$. The encoder, represented by the function $f : \mathcal{X}^n \to \{1, 2, \ldots, 2^{nR}\}$, maps the input sequence X^n into a compressed message M at a rate of R bits. This message is then processed by the decoder, defined as $g : \{1, 2, \ldots, 2^{nR}\} \to \hat{\mathcal{X}}^n$, which reconstructs the sequence $\hat{X}^n$. The goal of this process is to preserve essential information while efficiently compressing data to meet the requirements of downstream tasks.

Distortion constraint. The reconstructed output $\hat{X}$ must satisfy the following distortion constraint:

$$\mathbb{E}(\Delta(X, \hat{X})) \leq D, \tag{1}$$

where $\Delta : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$ represents a distortion metric such as Hamming distortion or mean squared error. The expectation is computed over the joint distribution $p_{X,\hat{X}} = p_{\hat{X}|X} p_X$.

Classification constraint. We impose the following classification constraint:

$$H(S_k|\hat{X}) \leq C_k, \qquad k \in [K], \tag{2}$$

for some $C_k > 0$. This ensures that the uncertainty of the classification variable S_k given the reconstructed source $\hat{X}$ does not exceed C_k [3].

Information RDC function. To quantify the achievable rate under both distortion and classification constraints, the information rate-distortion-classification function for a source $X \sim p_X$ is defined as follows:

Definition 1 (Information Rate-Distortion-Classification Function)**.** [3] For a source $X \sim p_X$ and a single associated classification variable S, the *information rate-distortion-classification function* is defined as:

$$R(D, C) = \min_{p_{\hat{X}|X}} \quad I(X; \hat{X}) \tag{3a}$$

$$\text{s.t.} \quad \mathbb{E}[\Delta(X, \hat{X})] \leq D, \tag{3b}$$

$$H(S|\hat{X}) \leq C. \tag{3c}$$

IV. Gaussian Source Case

This section examines the rate-distortion-classification tradeoff for a scalar Gaussian source. Leveraging the analytical tractability of the Gaussian setting, we derive closed-form expressions for both the RDC and DCR functions, offering valuable insights into the relationship between compression, distortion, and classification accuracy.

For a scalar Gaussian source, the closed-form expression of $R(D, C)$ is given in the following theorem by Wang et al. [3].

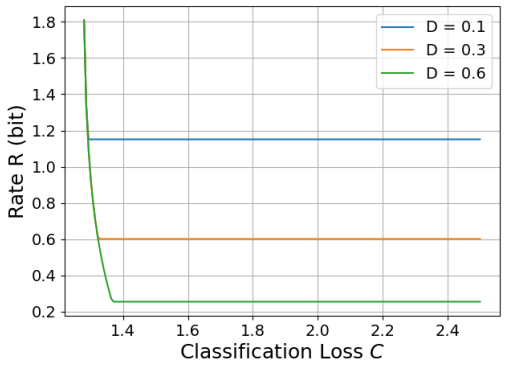


Figure 2: Illustration of the information rate-distortion-classification function of a Gaussian source.

Theorem 1 (Information Rate-Distortion-Classification Function for a Gaussian Source)**.** [3] Let $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ be a Gaussian source and $S \sim \mathcal{N}(\mu_S, \sigma_S^2)$ be an associated classification variable, with a covariance of $\text{Cov}(X, S) = \theta_1$. The problem (2) is feasible if $C \geq \frac{1}{2}\log\left(1 - \frac{\theta_1^2}{\sigma_S^2\sigma_X^2}\right) + h(S)$. For the MSE distortion (i.e., $\mathbb{E}[\Delta(X, \hat{X})] = \mathbb{E}[(X - \hat{X})^2]$), the *information rate-distortion-classification function* is achieved by a jointly Gaussian estimator $\hat{X}$ and given by

$$R(D, C) = \begin{cases} \frac{1}{2}\log\frac{\sigma_X^2}{D}, & D \leq \sigma_X^2\left(1 - \frac{1}{\rho^2}\left(1 - e^{-2h(S)+2C}\right)\right) \\ -\frac{1}{2}\log\left(1 - \frac{1}{\rho^2}\left(1 - e^{-2h(S)+2C}\right)\right), & D > \sigma_X^2\left(1 - \frac{1}{\rho^2}\left(1 - e^{-2h(S)+2C}\right)\right) \\ 0, & C > h(S) \text{ and } D > \sigma_X^2. \end{cases}$$

where $\rho = \frac{\theta_1}{\sigma_S\sigma_X}$ represents the correlation coefficient between X and S, while $h(\cdot)$ denotes the differential entropy.

The visualization of Theorem 1 is provided in Figure 2, with parameters set to $\sigma_X^2 = \sigma_S^2 = 1.0$ and $\theta_1 = 0.5$. The tradeoff between rate, distortion, and classification is clearly illustrated.

This characterization highlights the intricate tradeoff between reconstruction fidelity and classification performance. In the first case of the theorem, the RDC function reduces to the classical rate-distortion formulation, with the classification constraint inactive. As the constraint tightens (i.e., C decreases), a higher rate is required to jointly satisfy both distortion and classification objectives. This dependency is reflected in the curvature of the RDC boundary illustrated in Figure 2.

Furthermore, we characterize the minimum achievable distortion as a function of C and R by the following definition.

Definition 2. For a source $X \sim p_X$ and classification variable S, the *information distortion-classification-rate (DCR) function* is defined as:

$$D(C, R) = \min_{p_{\hat{X}|X}} \quad \mathbb{E}[(X - \hat{X})^2] \tag{4a}$$

$$\text{s.t.} \quad I(X; \hat{X}) \leq R, \tag{4b}$$

$$H(S|\hat{X}) \leq C. \tag{4c}$$

Our first contribution is the derivation of a closed-form expression for $D(C, R)$ in the Gaussian source setting, as formally stated in Theorem 2.

Theorem 2 (Information Distortion-Classification-Rate Function for a Gaussian Source)**.** Consider a Gaussian source $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and an associated classification variable $S \sim \mathcal{N}(\mu_S, \sigma_S^2)$ with covariance $\mathrm{Cov}(X, S) = \theta_1$. The problem (4) is feasible if the classification loss satisfies $C \geq \frac{1}{2}\log\left(1 - \frac{\theta_1^2}{\sigma_S^2\sigma_X^2}\right) + h(S)$. Under the MSE distortion, the *information distortion-classification-rate function* is achieved by a jointly Gaussian estimator $\hat{X}$ and is given by

$$D(C,R) = \begin{cases} \sigma_X^2 e^{-2R}, & C > \frac{1}{2}\log\left(1 - \frac{\theta_1^2(\sigma_X^2 - \sigma_X^2 e^{-2R})}{\sigma_S^2\sigma_X^4}\right) + h(S) \\ \sigma_X^2 - \frac{\sigma_S^2\sigma_X^4}{\theta_1^2}\left(1 - e^{-2h(S)+2C}\right), & \\ & \frac{1}{2}\log\left(1 - \frac{\theta_1^2}{\sigma_S^2\sigma_X^2}\right) + h(S) \leq C \leq \frac{1}{2}\log\left(1 - \frac{\theta_1^2(\sigma_X^2 - \sigma_X^2 e^{-2R})}{\sigma_S^2\sigma_X^4}\right) + h(S) \\ 0, & C > h(S) \text{ and } R > h(X). \end{cases}$$

Proof. Consider the distortion-classification-rate function $D(C, R)$ under the MSE distortion criterion as follows

$$D(C,R) = \min_{p_{\hat{X}|X}} \quad \mathbb{E}[(X - \hat{X})^2] \tag{5a}$$

$$\text{s.t.} \quad I(X;\hat{X}) \leq R, \tag{5b}$$

$$h(S|\hat{X}) \leq C. \tag{5c}$$

where (X, S) are jointly Gaussian random variables with covariance $\mathrm{Cov}(X, S) = \theta_1$. The optimal solution is attained when $\hat{X}$ is also Gaussian and jointly distributed with X [3], [4], [30]. Indeed, we can replace any random variable $\hat{X}$ with a Gaussian random variable $\hat{X}_G$, having same mean and variance as X, such that (a) $\mathbb{E}[(X - \hat{X})^2] \geq \mathbb{E}[(X - \hat{X}_G)^2]$, (b) $I(X;\hat{X}) \geq I(X;\hat{X}_G)$, and (c) $h(S|\hat{X}) \geq h(S|\hat{X}_G)$. The proof for claims (a) and (b) can be found in the proof of Theorem 1 in [4], and claim (c) is inherited from the entropy power inequality in [30], [31]. Thus, the optimization reduces to a parameter search over the mean $\mu_{\hat{X}}$, variance $\sigma_{\hat{X}}^2$, and covariance $\mathrm{Cov}(X, \hat{X}) = \theta_2$.

By applying the closed-form expressions for differential entropy and mutual information of jointly Gaussian variables [5], we obtain:

$$I(X;\hat{X}) = -\frac{1}{2}\log\left(1 - \frac{\theta_2^2}{\sigma_X^2\sigma_{\hat{X}}^2}\right), \tag{6}$$

And for the classification constraint:

$$h(S|\hat{X}) = h(S) - I(S;\hat{X}) \leq C,$$
$$I(S;\hat{X}) \geq h(S) - C,$$
$$-\frac{1}{2}\log\left(1 - \frac{\theta_1^2}{\sigma_S^2\sigma_X^4} \times \frac{\theta_2^2}{\sigma_{\hat{X}}^2}\right) \geq h(S) - C.$$

Additionally, the mean squared error between X and $\hat{X}$ can be expressed as [4]:

$$\mathbb{E}[(X-\hat{X})^2] = (\mu_X - \mu_{\hat{X}})^2 + \sigma_X^2 + \sigma_{\hat{X}}^2 - 2\theta_2. \tag{7}$$

Then, the $D(C,R)$ problem can be formulated as:

$$D(C,R) = \min_{\mu_{\hat{X}}, \sigma_{\hat{X}}, \theta_2} (\mu_X - \mu_{\hat{X}})^2 + \sigma_X^2 + \sigma_{\hat{X}}^2 - 2\theta_2 \tag{8a}$$
$$\text{s.t.} -\frac{1}{2}\log\left(1 - \frac{\theta_2^2}{\sigma_X^2\sigma_{\hat{X}}^2}\right) \leq R, \tag{8b}$$
$$-\frac{1}{2}\log\left(1 - \frac{\theta_1^2}{\sigma_S^2\sigma_X^4}\frac{\theta_2^2}{\sigma_{\hat{X}}^2}\right) \geq h(S) - C. \tag{8c}$$

Since both the rate constraint (8b) and the classification constraint (8c) are independent of the mean $\mu_{\hat{X}}$, and θ_2 only depends on the variance of X and $\hat{X}$ when X and $\hat{X}$ are Gaussian distributions, the objective function is minimized when the means match, i.e., $(\mu_X - \mu_{\hat{X}})^2 + \sigma_X^2 + \sigma_{\hat{X}}^2 - 2\theta_2 \geq \sigma_X^2 + \sigma_{\hat{X}}^2 - 2\theta_2$, we let $\mu_X = \mu_{\hat{X}}$ in the subsequent derivations.

To ensure that the mutual information expression in (8b) is well-defined, it is necessary that $1 - \frac{\theta_2^2}{\sigma_X^2\sigma_{\hat{X}}^2} > 0$, i.e., $\frac{\theta_2^2}{\sigma_{\hat{X}}^2} < \sigma_X^2$. Under this condition, the mutual information between S and $\hat{X}$ is upper bounded as

$$I(S;\hat{X}) = -\frac{1}{2}\log\left(1 - \frac{\theta_1^2}{\sigma_S^2\sigma_X^4} \times \frac{\theta_2^2}{\sigma_{\hat{X}}^2}\right) \leq -\frac{1}{2}\log\left(1 - \frac{\theta_1^2}{\sigma_S^2\sigma_X^2}\right),$$

which implies that constraint (8c) becomes infeasible if $C < \frac{1}{2}\log\left(1 - \frac{\theta_1^2}{\sigma_S^2\sigma_X^2}\right) + h(S)$. Therefore, to guarantee feasibility, we assume throughout that

$$C \geq \frac{1}{2}\log\left(1 - \frac{\theta_1^2}{\sigma_S^2\sigma_X^2}\right) + h(S).$$

The optimization problem (8) can be effectively solved using the Karush-Kuhn-Tucker (KKT) conditions. Our approach systematically explores all possible combinations of active and inactive rate and classification constraints to characterize the optimal solution.

Case 1. Constraint (8b) is active and constraint (8c) is inactive.

Recall the classical Shannon rate-distortion function for a Gaussian source $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ [5]:

$$R(D) = \begin{cases} \frac{1}{2}\log(\frac{\sigma_X^2}{D}), & 0 \leq D \leq \sigma_X^2 \\ 0, & D > \sigma_X^2. \end{cases}$$

And,

$$D(R) = \sigma_X^2 e^{-2R},$$

with the optimal solution attained by some $p_{\hat{X}|X}$ where $\hat{X} \sim \mathcal{N}(\mu_X, \sigma_X^2 - \sigma_X^2 e^{-2R})$. In this case, we have $D(C,R) = \sigma_X^2 e^{-2R}$, achieved by choosing $\sigma_{\hat{X}}^2 = \sigma_X^2 - \sigma_X^2 e^{-2R}$ when the rate constraint (8b) is active. It implies that:

$$-\frac{1}{2}\log\left(1 - \frac{\theta_2^2}{\sigma_X^2 \sigma_{\hat{X}}^2}\right) = R \Rightarrow \theta_2 = \sigma_X^2 - \sigma_X^2 e^{-2R}.$$

The constraint (8c) is not active if

$$-\frac{1}{2}\log\left(1 - \frac{\theta_1^2}{\sigma_S^2 \sigma_X^4}\frac{\theta_2^2}{\sigma_{\hat{X}}^2}\right) > h(S) - C.$$

Hence,

$$C > \frac{1}{2}\log\left(1 - \frac{\theta_1^2(\sigma_X^2 - \sigma_X^2 e^{-2R})}{\sigma_S^2 \sigma_X^4}\right) + h(S).$$

Therefore, $D(C,R) = \sigma_X^2 e^{-2R}$ if $C > \frac{1}{2}\log\left(1 - \frac{\theta_1^2(\sigma_X^2 - \sigma_X^2 e^{-2R})}{\sigma_S^2 \sigma_X^4}\right) + h(S)$.

Case 2. Constraint (8b) is inactive and constraint (8c) is active.

The classification constraint (8c) is active, infer that:

$$\begin{aligned} -\frac{1}{2}\log\left(1 - \frac{\theta_1^2}{\sigma_S^2 \sigma_X^4}\frac{\theta_2^2}{\sigma_{\hat{X}}^2}\right) &= h(S) - C, \\ \Rightarrow \frac{\theta_2^2}{\sigma_{\hat{X}}^2} &= \frac{\sigma_S^2 \sigma_X^4}{\theta_1^2}\left(1 - e^{-2h(S)+2C}\right). \end{aligned} \tag{9}$$

We choose $\sigma_{\hat{X}}^2 = \theta_2 = \frac{\sigma_S^2 \sigma_X^4}{\theta_1^2}\left(1 - e^{-2h(S)+2C}\right)$ and substitute into the distortion expression:

$$\begin{aligned} \mathbb{E}[(X - \hat{X})^2] &= \sigma_X^2 + \sigma_{\hat{X}}^2 - 2\theta_2, \\ &= \sigma_X^2 - \frac{\sigma_S^2 \sigma_X^4}{\theta_1^2}\left(1 - e^{-2h(S)+2C}\right). \end{aligned}$$

Substitute (9) into the rate expression (6), we get:

$$I(X;\hat{X}) = -\frac{1}{2}\log\left(1 - \frac{\sigma_S^2 \sigma_X^2}{\theta_1^2}\left(1 - e^{-2h(S)+2C}\right)\right).$$

So, the rate constraint (8b) is inactive when

$$-\frac{1}{2}\log\left(1-\frac{\sigma_S^2\sigma_X^2}{\theta_1^2}\left(1-e^{-2h(S)+2C}\right)\right)<R,$$

$$\Rightarrow C<\frac{1}{2}\log\left(1-\frac{\theta_1^2(\sigma_X^2-\sigma_X^2e^{-2R})}{\sigma_S^2\sigma_X^4}\right)+h(S).$$

Therefore, $D(C,R)=\sigma_X^2-\frac{\sigma_S^2\sigma_X^4}{\theta_1^2}\left(1-e^{-2h(S)+2C}\right)$ if $C<\frac{1}{2}\log\left(1-\frac{\theta_1^2(\sigma_X^2-\sigma_X^2e^{-2R})}{\sigma_S^2\sigma_X^4}\right)+h(S)$.

Case 3: Both the rate constraint (8b) and the classification constraint (8c) are active.

From case 2, we know that the classification constraint (8c) is active if

$$\sigma_{\hat{X}}^2=\theta_2=\frac{\sigma_S^2\sigma_X^4}{\theta_1^2}\left(1-e^{-2h(S)+2C}\right),$$

and,

$$\mathbb{E}[(X-\hat{X})^2]=\sigma_X^2-\frac{\sigma_S^2\sigma_X^4}{\theta_1^2}\left(1-e^{-2h(S)+2C}\right).$$

The rate constraint (8b) is active if

$$I(X;\hat{X})=-\frac{1}{2}\log\left(1-\frac{\sigma_S^2\sigma_X^2}{\theta_1^2}\left(1-e^{-2h(S)+2C}\right)\right)=R.$$

$$\Rightarrow C=\frac{1}{2}\log\left(1-\frac{\theta_1^2(\sigma_X^2-\sigma_X^2e^{-2R})}{\sigma_S^2\sigma_X^4}\right)+h(S).$$

Therefore, $D(C,R)=\sigma_X^2-\frac{\sigma_S^2\sigma_X^4}{\theta_1^2}\left(1-e^{-2h(S)+2C}\right)$ if $C=\frac{1}{2}\log\left(1-\frac{\theta_1^2(\sigma_X^2-\sigma_X^2e^{-2R})}{\sigma_S^2\sigma_X^4}\right)+h(S)$.

Case 4. Both constraint (8b) and constraint (8c) are inactive.

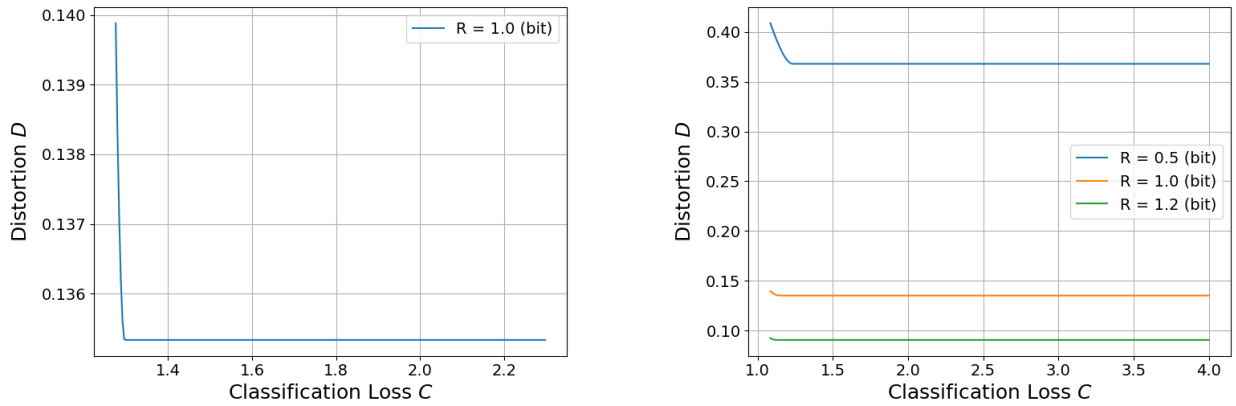
When $C>h(S)$, implying that the classification constraint (8c) is inactive, and the rate R is sufficiently large such that $R>h(X)$, meaning the rate constraint (8b) is also inactive, the minimum achievable distortion $D(C,R)$ reaches its theoretical lower bound, i.e., $D(C,R)=0$. This is achieved by setting $\hat{X}=X$, which leads to zero reconstruction error, i.e., $\mathbb{E}[(X-\hat{X})^2]=0$. Furthermore, all constraints are satisfied since $I(X;\hat{X})=h(X)<R$, and $h(S|\hat{X})=h(S|X)\leq h(S)<C$.

Therefore, $D(C,R)=0$ if $C>h(S)$ and $R>h(X)$.

In summary, combining the four cases, the closed-form expression for the information distortion-classification-rate function $D(C,R)$ under MSE distortion is given by Theorem 2. □

For any fixed R, as C increases from $\frac{1}{2}\log\left(1-\frac{\theta_1^2}{\sigma_S^2\sigma_X^2}\right)+h(S)$ to $\frac{1}{2}\log\left(1-\frac{\theta_1^2(\sigma_X^2-\sigma_X^2e^{-2R})}{\sigma_S^2\sigma_X^4}\right)+h(S)$, the distortion $D(C,R)$ decreases monotonically from

$$\sigma_X^2+\frac{\sigma_S^2\sigma_X^4(1-e^{2C-2h(S)})}{\theta_1^2}-2\frac{\sigma_S\sigma_X^3\sqrt{(1-e^{-2h(S)+2C})(1-e^{-2R})}}{\theta_1}$$



(a) $D(C, R)$ function for a Gaussian source.

(b) $D(C, R)$ functions across multiple rates.

Figure 3: Illustration of the DCR functions: (a) shows the $D(C, R)$ function for a Gaussian source, and (b) shows how the function varies across multiple rates.

to the optimal value $\sigma_X^2 e^{-2R}$. Increasing C beyond this point does not yield further improvements in distortion.

Figure 3 shows the information distortion-classification-rate function in Theorem 2, with parameters set to $\sigma_X^2 = \sigma_S^2 = 1.0$ and $\theta_1 = 0.7$. Subfigure 3(a) depicts the DCR tradeoff at a fixed rate, while subfigure 3(b) illustrates how the function varies across multiple rate values.

These plots illustrate the operational interpretation of the DCR surface. At a fixed rate, relaxing the classification constraint (i.e., increasing C) enables lower achievable distortion. Conversely, for a fixed classification constraint, increasing the rate allows the encoder to preserve more source information, thereby reducing distortion. This visualization highlights the fundamental tradeoff among rate, distortion, and classification, and identifies regimes where further relaxing classification requirements yields diminishing returns in distortion.

V. Universal Encoder Representations

Designing separate encoders for each distortion-classification constraint is often not desirable. This motivates the use of universal representations, where a single encoder supports multiple decoding constraints, each for a distinct task, as shown in Figure 4. This section introduces the universal RDC framework, quantifies the rate penalty, and presents theoretical results for both Gaussian and general sources.

A. Definitions

In the standard RDC setting, the minimum rate to satisfy a distortion-classification pair (D, C) is achieved by jointly optimizing the encoder and decoder. The proposed universal RDC framework extends this by fixing the encoder and allowing the decoder to adapt, thereby supporting all constraint pairs $(D, C) \in \Theta$, where Θ is a given set of multiple (D, C) pairs. As illustrated in Fig. 4, the encoder produces a shared latent representation Z, while each decoder i implements a conditional distribution $p_{\hat{X}_{D_i,C_i}|Z}$, generating reconstructions $\hat{X}_{D_i,C_i}$ that satisfy their respective distortion and classification constraints (D_i, C_i).

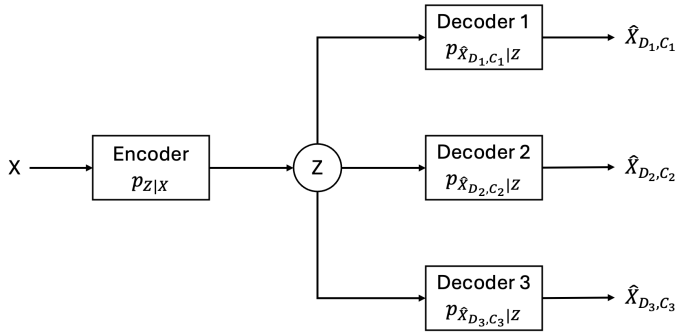


Figure 4: Illustration of the universal encoder representation framework.

A particularly interesting case arises when the set Θ includes all distortion-classification pairs (D, C) that lie along the RDC curve for a fixed rate. This setting raises a fundamental question: *How much additional rate is required to simultaneously satisfy all constraints in Θ using a single encoder, rather than designing a separate encoder for each target pair?* Ideally, the cost of universality should be minimal, implying that the excess rate required to support all tasks is close to that needed for the most demanding constraint in Θ.

To formalize this notion, we introduce *information universal rate-distortion-classification function* by adapting the definition from [4]. We assume that X is a random variable and Θ is a non-empty set of constraint pairs (D, C).

Definition 3 (Information Universal RDC Function)**.** Let Z be a representation of X, generated through a conditional distribution $p_{Z|X}$. Define $\mathcal{P}_{Z|X}(\Theta)$ as the set of such transformations for which, for every $(D, C) \in \Theta$, there exists a decoder $p_{\hat{X}_{D,C}|Z}$ such that $\mathbb{E}[\Delta(X, \hat{X}_{D,C})] \leq D$ and $H(S|\hat{X}_{D,C}) \leq C$, with the Markov chain $X \leftrightarrow Z \leftrightarrow \hat{X}_{D,C}$ holding. The *information*

universal RDC function is defined as

$$R(\Theta) = \inf_{p_{Z|X} \in \mathcal{P}_{Z|X}(\Theta)} I(X;Z). \tag{10}$$

A representation Z is said to be Θ-universal with respect to X if it allows all constraints in Θ to be satisfied using appropriate decoders. In this context, each decoder $p_{\hat{X}_{D,C}|Z}$ maps the shared representation Z to a reconstruction $\hat{X}_{D,C}$ tailored to the specific constraint pair. The quantity $R(\Theta)$ thus captures the minimal rate needed to meet all constraints in Θ with a shared encoder. In contrast, the term $\sup_{(D,C)\in\Theta} R(D,C)$ represents the rate required to satisfy only the most demanding individual constraint. The difference between these two defines the rate penalty. Similarly, we adapt the definition from [4] to quantify the rate penalty under the classification constraint as follows:

Definition 4 (Rate Penalty)**.** The *rate penalty* for a constraint set Θ is defined as

$$A(\Theta) = R(\Theta) - \sup_{(D,C)\in\Theta} R(D,C), \tag{11}$$

which quantifies the additional rate incurred when using a single encoder to satisfy all constraints in Θ.

Let $\Omega(R) = \{(D,C) : R(D,C) \leq R\}$ denote the set of distortion-classification pairs that are achievable at rate R. Ideally, we would like $A(\Omega(R)) = 0$ for all R, indicating that the entire tradeoff curve can be achieved without incurring any additional cost for universality. This would eliminate the need to design separate encoders for different distortion-classification goals at the same rate.

Given a representation Z, we define the associated achievable distortion-classification region as

$$\Omega(p_{Z|X}) = \left\{(D,C) : \exists\, p_{\hat{X}_{D,C}|Z} \text{ such that } \mathbb{E}[\Delta(X,\hat{X}_{D,C})] \leq D, \quad H(S|\hat{X}_{D,C}) \leq C\right\}.$$

Intuitively, $\Omega(p_{Z|X})$ describes the set of constraint pairs that can be satisfied using the fixed representation Z. If a representation Z satisfies $I(X;Z) = R$ and $\Omega(p_{Z|X}) = \Omega(R)$, then Z is said to achieve the maximal distortion-classification region at rate R. That is, for any other representation Z' with $I(X;Z') \leq R$, we have $\Omega(p_{Z'|X}) \subseteq \Omega(p_{Z|X})$.

B. Universal Encoder Representation Characterization

Theorem 3 (No Rate Penalty for a Gaussian Source)**.** Let $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ be a scalar Gaussian source and let $S \sim \mathcal{N}(\mu_S, \sigma_S^2)$ be an associated classification variable. Assume MSE distortion

and classification loss measured by $H(S|\hat{X})$. Let Θ denote any non-empty set of distortion-classification constraint pairs (D, C). Then,

$$A(\Theta) = 0, \tag{12}$$

which implies that satisfying the most demanding constraint in Θ is sufficient to simultaneously satisfy all others using a fixed encoder and there is no rate penalty for universality in this case.

Furthermore, consider any representation Z that is jointly Gaussian with X and satisfies

$$I(X;Z) = \sup_{(D,C)\in\Theta} R(D,C). \tag{13}$$

Then the following inclusion holds:

$$\Theta \subseteq \Omega(p_{Z|X}) = \Omega(I(X;Z)), \tag{14}$$

meaning that Z achieves the maximal distortion-classification region at rate $I(X;Z)$; i.e., all constraints in Θ are simultaneously achievable via appropriate decoders applied to a shared representation Z.

Proof. The proof method follows the approach presented in [4]. Let $R = \sup_{(D,C)\in\Theta} R(D,C)$. By definition, $\Theta \subseteq \Omega(R)$, where $\Omega(R)$ denotes the set of all achievable distortion-classification pairs at rate R. The lower boundary of this region, the optimal tradeoff curve, is characterized by

$$D = \sigma_X^2 - \frac{\sigma_S^2\sigma_X^4}{\theta_1^2}\left(1 - e^{-2h(S)+2C}\right),$$
$$C \in \left[\frac{1}{2}\log\left(1 - \frac{\theta_1^2}{\sigma_S^2\sigma_X^2}\right) + h(S), \frac{1}{2}\log(1 - \frac{\theta_1^2(\sigma_X^2 - \sigma_X^2 e^{-2R})}{\sigma_S^2\sigma_X^4}) + h(S)\right).$$

Every point in $\Omega(R)$ is component-wise dominated by a point on this boundary. Consider a representation Z that is jointly Gaussian with X, such that $I(X;Z) = R$. This implies the squared correlation coefficient between X and Z satisfies $\rho_{XZ}^2 = 1 - 2^{-2R}$ [4], where

$$\rho_{XZ} = \frac{\text{Cov}(X,Z)}{\sigma_X\sigma_Z} = \frac{\mathbb{E}[(X-\mu_X)(Z-\mu_Z)]}{\sigma_X\sigma_Z}.$$

For any point (D, C) on the boundary, define the corresponding reconstruction as:

$$\hat{X}_{D,C} = \text{sign}(\rho_{XZ})\gamma(Z - \mu_Z) + \mu_X,$$

where γ denotes a scaling coefficient and

$$\text{sign}(\rho_{XZ}) = \begin{cases} 1, & \text{for } \rho_{XZ} \geq 0, \\ -1, & \text{for } \rho_{XZ} < 0. \end{cases}$$

With this construction, we have:

$$\begin{aligned}\mu_{\hat{X}_{D,C}} &= \mathbb{E}[\text{sign}(\rho_{XZ})\gamma(Z-\mu_Z)+\mu_X],\\ &= \text{sign}(\rho_{XZ})\gamma(\mathbb{E}[Z]-\mathbb{E}[Z])+\mu_X = \mu_X.\end{aligned}$$

And,

$$\sigma_{\hat{X}_{D,C}} = \gamma^2\sigma_Z,$$

$$\text{Cov}(X,\hat{X}_{D,C}) = \gamma\text{Cov}(X,Z).$$

We now choose:

$$\sigma_{\hat{X}_{D,C}} = \theta_2 = \frac{\sigma_S^2\sigma_X^4(1-e^{-2h(S)+2C})}{\theta_1^2}.$$

Solving for γ, we obtain:

$$\gamma = \frac{\sigma_S\sigma_X^2\sqrt{1-e^{-2h(S)+2C}}}{\theta_1\sigma_Z}.$$

Hence,

$$\hat{X}_{D,C} = \text{sign}(\rho_{XZ})\frac{\sigma_S\sigma_X^2\sqrt{1-e^{-2h(S)+2C}}}{\theta_1\sigma_Z}(Z-\mu_Z)+\mu_X.$$

We now verify that this reconstruction satisfies the distortion and classification constraints.

Distortion constraint.

$$\begin{aligned}\mathbb{E}[\|X-\hat{X}_{D,C}\|^2] &= \sigma_X^2+\sigma_{\hat{X}_{D,C}}^2-2\theta_2,\\ &= \sigma_X^2-\frac{\sigma_S^2\sigma_X^4}{\theta_1^2}\left(1-e^{-2h(S)+2C}\right),\\ &= D.\end{aligned}$$

Classification constraint.

$$\begin{aligned}h(S|\hat{X}_{D,C}) &= h(S)-I(S|\hat{X}_{D,C}),\\ &= h(S)+\frac{1}{2}\log\left(1-\frac{\theta_1^2}{\sigma_S^2\sigma_X^4}\frac{\theta_2^2}{\sigma_{\hat{X}_{D,C}}^2}\right),\end{aligned}$$

where

$$\frac{\theta_2^2}{\sigma_{\hat{X}_{D,C}}^2} = \frac{\sigma_S^2\sigma_X^4(1-e^{-2h(S)+2C})}{\theta_1^2}.$$

Substituting in, we get:

$$\begin{aligned} h(S|\hat{X}_{D,C}) &= h(S) \\ &\quad + \frac{1}{2}\log\left(1 - \frac{\theta_1^2}{\sigma_S^2\sigma_X^4}\frac{\sigma_S^2\sigma_X^4(1-e^{-2h(S)+2C})}{\theta_1^2}\right), \\ &= h(S) + \frac{1}{2}\log\left(e^{-2h(S)+2C}\right), \\ &= C. \end{aligned}$$

This confirms that for any given $\mathrm{Cov}(X,Z)$, one can always choose a scalar γ to generate $\hat{X}_{D,C}$ satisfying both distortion and classification constraints. That is, every point $(D,C) \in \Theta$ can be realized by applying an appropriate decoder to a fixed Gaussian representation Z of X such that $I(X;Z) = \sup_{(D,C)\in\Theta} R(D,C)$. Therefore, $\Omega(p_{Z|X}) = \Omega(R)$, which implies the rate penalty is zero: $A(\Theta) = I(X;Z) - R = 0$. □

In addition, we consider a general source $X \sim p_X$ and characterize the distortion-classification region induced by an arbitrary representation Z under MSE distortion.

Theorem 4 (Achievable Universality Region for a General Source)**.** Let $X \sim p_X$ be a general source and S an associated classification variable. Assume distortion is measured by MSE and classification loss by $H(S|\hat{X})$. Let Z be any representation of X, and define $\tilde{X} = \mathbb{E}[X|Z]$ as the minimum mean square estimator. Then the closure of the achievable region, $\mathrm{cl}(\Omega(p_{Z|X}))$, satisfies

$$\Omega(p_{Z|X}) \subseteq \left\{ (D,C) : D \geq \mathbb{E}\|X - \tilde{X}\|^2 + \begin{array}{l} \inf\limits_{p_{\hat{X}}} W_2^2(p_{\tilde{X}}, p_{\hat{X}}) \\ \text{s.t. } H(S|\hat{X}) \leq C \end{array} \right\} \subseteq \mathrm{cl}(\Omega(p_{Z|X})),$$

where the squared Wasserstein-2 distance is $W_2^2(p_X, p_{\hat{X}}) = \inf_{p_{X,\hat{X}}} \mathbb{E}[\|X - \hat{X}\|^2]$ with the infimum taken over all joint distributions with marginals p_X and $p_{\hat{X}}$.

Moreover, $\mathrm{cl}(\Omega(p_{Z|X}))$ contains the extreme points:

$$\begin{aligned} (D^{(a)}, C^{(a)}) &= \left(\mathbb{E}[\|X - \tilde{X}\|^2], \sum_s \sum_{\tilde{x}} p_{\tilde{X}} p_{S|\tilde{X}} \log \frac{1}{p_{S|\tilde{X}}}\right), \\ (D^{(b)}, C^{(b)}) &= \left(\mathbb{E}[\|X - \tilde{X}\|^2] + W_2^2(p_{\tilde{X}}, p_{\hat{X}^{C_{\min}}}), C_{\min}\right), \end{aligned}$$

where

$$p_{\hat{X}^{C_{\min}}} = \arg\min_{p_{\hat{X}}} \quad H(S|\hat{X}) \tag{15a}$$

$$\text{s.t.} \quad \mathbb{E}[\|X - \hat{X}\|^2] \leq D. \tag{15b}$$

The minimum classification loss is: $C_{\min} = \sum_s \sum_{\hat{x}^{C_{\min}}} p_{\hat{X}^{C_{\min}}} p_{S|\hat{X}^{C_{\min}}} \log \frac{1}{p_{S|\hat{X}^{C_{\min}}}}$.

Proof. The proof approach is inspired by the main techniques introduced in [4]. For any $(D, C) \in \Omega(p_{Z|X})$, there exists a reconstruction variable $\hat{X}_{D,C}$ jointly distributed with (X, Z), such that the Markov chain $X \leftrightarrow Z \leftrightarrow \hat{X}_{D,C}$ holds, the distortion satisfies $\mathbb{E}[\Delta(X, \hat{X}_{D,C})] \leq D$, and the classification uncertainty is bounded by $H(S|\hat{X}_{D,C}) \leq C$.

Since $\tilde{X} = \mathbb{E}[X|Z]$, we have:

$$D \geq \mathbb{E}[\|X - \hat{X}_{D,C}\|^2] = \mathbb{E}[\|X - \tilde{X}\|^2] + \mathbb{E}[\|\tilde{X} - \hat{X}_{D,C}\|^2].$$

Now consider the Wasserstein-2 distance between the marginals $p_{\tilde{X}}$ and $p_{\hat{X}_{D,C}}$, which is defined as:

$$W_2^2(p_{\tilde{X}}, p_{\hat{X}_{D,C}}) = \inf_{p_{\tilde{X}', \hat{X}'}} \mathbb{E}[\|\tilde{X}' - \hat{X}'\|^2],$$

where $\tilde{X}' \sim p_{\tilde{X}}$ and $\hat{X}' \sim p_{\hat{X}_{D,C}}$. Since $(\tilde{X}, \hat{X}_{D,C})$ is one feasible coupling of these marginals,

$$\mathbb{E}[\|\tilde{X} - \hat{X}_{D,C}\|^2] \geq W_2^2(p_{\tilde{X}}, p_{\hat{X}_{D,C}}).$$

Therefore,

$$D \geq \mathbb{E}[\|X - \tilde{X}\|^2] + W_2^2(p_{\tilde{X}}, p_{\hat{X}_{D,C}}).$$

Since $H(S|\hat{X}_{D,C}) \leq C$, the marginal distribution $p_{\hat{X}_{D,C}}$ belongs to the constraint set $\{p_{\hat{X}} : H(S|\hat{X}) \leq C\}$. We have, $p_{\hat{X}_{D,C}}$ is one feasible distribution of the set $\left\{ p_{\hat{X}} : \begin{array}{c} \inf\limits_{p_{\hat{X}}} W_2^2(p_{\tilde{X}}, p_{\hat{X}}) \\ \text{s.t. } H(S|\hat{X}) \leq C \end{array} \right\}$, then

$$W_2^2(p_{\tilde{X}}, p_{\hat{X}_{D,C}}) \geq \begin{cases} \inf\limits_{p_{\hat{X}}} W_2^2(p_{\tilde{X}}, p_{\hat{X}}) \\ \text{s.t. } H(S|\hat{X}) \leq C. \end{cases}$$

This leads to the outer bound:

$$\Omega(p_{Z|X}) \subseteq \left\{ (D, C) : D \geq \mathbb{E}[\|X - \tilde{X}\|^2] + \begin{array}{c} \inf\limits_{p_{\hat{X}}} W_2^2(p_{\tilde{X}}, p_{\hat{X}}) \\ \text{s.t. } H(S|\hat{X}) \leq C. \end{array} \right\}$$

Now, to show the approximate tightness of this bound, let (D', C') be any point in the above region. For any $\epsilon > 0$, there exists a distribution $p_{\hat{X}'}$ such that:

$$H(S|\hat{X}') \leq C', \quad D' + \epsilon \geq \mathbb{E}[\|X - \tilde{X}\|^2] + W_2^2(p_{\tilde{X}}, p_{\hat{X}'}).$$

By the Markov condition, we can construct a random variable $\hat{X}'$ such that $X \leftrightarrow Z \leftrightarrow \hat{X}'$, and

$$\mathbb{E}[\|\tilde{X} - \hat{X}'\|^2] \leq W_2^2(p_{\tilde{X}}, p_{\hat{X}'}) + \epsilon.$$

Thus,

$$\mathbb{E}[\|X - \hat{X}'\|^2] = \mathbb{E}[\|X - \tilde{X}\|^2] + \mathbb{E}[\|\tilde{X} - \hat{X}'\|^2] \leq D' + 2\epsilon.$$

It follows that:

$$\begin{aligned}\Omega(p_{Z|X}) \subseteq & \left\{ (D, C) : D \geq \mathbb{E}\|X - \tilde{X}\|^2 + \begin{array}{r} \inf\limits_{p_{\hat{X}}} W_2^2(p_{\tilde{X}}, p_{\hat{X}}) \\ \text{s.t. } H(S|\hat{X}) \leq C \end{array} \right\} \\ & \subseteq \text{cl}(\Omega(p_{Z|X})).\end{aligned}$$

Now consider the characterization of conditional entropy:

$$\begin{aligned}H(S|\hat{X}) &= \sum_s \sum_{\hat{x}} p_{S,\hat{X}} \log \frac{1}{p_{S|\hat{X}}}, \\ &= \sum_s \sum_{\hat{x}} p_{\hat{X}} p_{S|\hat{X}} \log \frac{1}{p_{S|\hat{X}}}.\end{aligned}$$

By choosing $\hat{X} = \tilde{X}$, it follows that $p_{\hat{X}} = p_{\tilde{X}}$, which yields:

$$H(S|\hat{X}) = H(S|\tilde{X}) = \sum_s \sum_{\tilde{x}} p_{\tilde{X}} p_{S|\tilde{X}} \log \frac{1}{p_{S|\tilde{X}}},$$

and define:

$$(D^{(a)}, C^{(a)}) = \left(\mathbb{E}[\|X - \tilde{X}\|^2], \sum_s \sum_{\tilde{x}} p_{\tilde{X}} p_{S|\tilde{X}} \log \frac{1}{p_{S|\tilde{X}}} \right).$$

Next, define $\hat{X}^{C_{\min}} \sim p_{\hat{X}^{C_{\min}}}$:

$$p_{\hat{X}^{C_{\min}}} = \arg \min_{p_{\hat{X}}} \quad H(S|\hat{X}) \tag{16a}$$

$$\text{s.t.} \quad \mathbb{E}[\|X - \hat{X}\|^2] \leq D. \tag{16b}$$

And,

$$C_{\min} = \sum_s \sum_{\hat{x}^{C_{\min}}} p_{\hat{X}^{C_{\min}}} p_{S|\hat{X}^{C_{\min}}} \log \frac{1}{p_{S|\hat{X}^{C_{\min}}}}.$$

Let $\hat{X} = \hat{X}^{C_{\min}}$ implies $p_{\hat{X}} = p_{\hat{X}^{C_{\min}}}$, and define:

$$(D^{(b)}, C^{(b)}) = \left(\mathbb{E}[\|X - \tilde{X}\|^2] + W_2^2(p_{\tilde{X}}, p_{\hat{X}^{C_{\min}}}), C_{\min}\right).$$

Thus, by selecting $p_{\hat{X}} = p_{\tilde{X}}$ and $p_{\hat{X}} = p_{\hat{X}^{C_{\min}}}$, we confirm that both $(D^{(a)}, C^{(a)})$ and $(D^{(b)}, C^{(b)})$ lie in this region:

$$\left\{(D, C) : D \geq \mathbb{E}\|X - \tilde{X}\|^2 + \begin{array}{l}\inf\limits_{p_{\hat{X}}} W_2^2(p_{\tilde{X}}, p_{\hat{X}}) \\ \text{s.t. } H(S|\hat{X}) \leq C\end{array}\right\}.$$

□

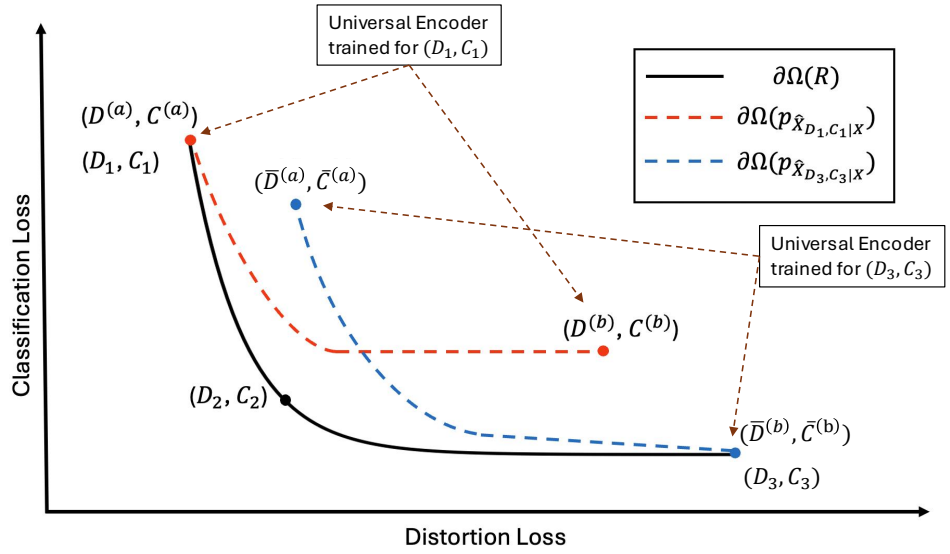


Figure 5: Universality for a general source. Shown are the boundaries of achievable distortion-classification regions corresponding to three different representations: the minimal distortion point (D_1, C_1), where $R(D_1, C_1) = R(D_1, \infty)$; the midpoint (D_2, C_2); and the minimal classification loss point (D_3, C_3), where $C_3 = C_{\min}$. Also illustrated are the two extreme points $(D^{(a)}, C^{(a)})$ and $(D^{(b)}, C^{(b)})$ associated with the representation $\hat{X}_{D_1,C_1}$, where $(D^{(a)}, C^{(a)})$ coincides with (D_1, C_1).

In addition, to gain further insight into the structure of achievable regions, let Z be the optimal reconstruction $\hat{X}_{D,C}$ associated with a point (D, C) on the distortion-classification tradeoff curve for a given rate R; that is, $I(X; \hat{X}_{D,C}) = R(D, C) = R$, and the distortion and classification loss satisfy $\mathbb{E}[\|X - \hat{X}_{D,C}\|^2] = D, \quad H(S|\hat{X}_{D,C}) = C$. We assume, for simplicity, that such an optimal reconstruction $\hat{X}_{D,C}$ exists for every (D, C) on the tradeoff boundary, and that decreasing

either D or C would violate the constraint $R(D, C) = R$. Under this assumption, the point (D, C) lies on the boundary of the closure $\mathrm{cl}(\Omega(p_{\hat{X}_{D,C}|X}))$.

According to Theorem 4, the set $\mathrm{cl}(\Omega(p_{\hat{X}_{D,C}|X}))$ includes two extreme points: the upper-left corner $(D^{(a)}, C^{(a)})$, corresponding to minimal distortion, and the lower-right corner $(D^{(b)}, C^{(b)})$, corresponding to minimal classification loss. Moreover, this closure defines a convex region that includes all these points. Figure 5 illustrates both $\Omega(R)$ and the achievable region $\Omega(p_{\hat{X}_{D,C}|X})$ for various representative points (D, C) on the tradeoff curve. The following theorem provides a quantitative characterization of this structure.

Theorem 5 (Quantitative Characterization for a General Source)**.** Let $\hat{X}_{D_1,C_1}$ denote the optimal reconstruction at point (D_1, C_1) on the conventional RDC trade-off curve, satisfying $I(X; \hat{X}_{D_1,C_1}) = R(D_1, C_1)$. Then the upper-left extreme point of $\Omega(p_{\hat{X}_{D_1,C_1}|X})$ satisfies $(D^{(a)}, C^{(a)}) = (D_1, C_1)$. Now consider the lower-right extreme points: $(D^{(b)}, C^{(b)}) \in \Omega(p_{\hat{X}_{D_1,C_1}|X})$ and $(D_3, C_3) \in \Omega(R)$, where $C_3 = C_{\min}$ and $R(D_3, C_3) = R(D_1, \infty)$. The distortion gap between these points is bounded below by:

$$D_3 - D^{(b)} \geq \sigma_X^2 + \sigma^2_{\hat{X}_{D_3,C_3}} - 2\sigma_{\hat{X}_{D_3,C_3}} \sqrt{\sigma_X^2 - D_1} - 2D_1, \tag{17}$$

and the corresponding distortion ratio satisfies:

$$\frac{D_3}{D^{(b)}} \geq \frac{\sigma_X^2 + \sigma^2_{\hat{X}_{D_3,C_3}} - 2\sigma_{\hat{X}_{D_3,C_3}} \sqrt{\sigma_X^2 - D_1}}{2D_1}. \tag{18}$$

In the case where $\sigma_X^2 = \sigma^2_{\hat{X}_{D_3,C_3}}$, the distortion gap becomes small under:

$$D_3 - D^{(b)} \approx 0 \ \text{ if } \ D_1 \approx 0 \text{ or } D_1 \approx \sigma_X^2, \tag{19}$$

$$\frac{D_3}{D^{(b)}} \approx 1 \ \text{ if } \ D_1 \approx \sigma_X^2. \tag{20}$$

Proof. The proof idea follows the result in [4]. We begin by noting that $C_3 = C_{\min}$, and from the hypothesis $R(D_3, C_3) = R(D_1, \infty)$. Next, observe that:

$$\begin{aligned} D_3 &= \mathbb{E}[\|X - \hat{X}_{D_3,C_3}\|^2], \\ &= \sigma_X^2 + \sigma^2_{\hat{X}_{D_3,C_3}} - 2\mathrm{Cov}(X, \hat{X}_{D_3,C_3}), \\ &= \sigma_X^2 + \sigma^2_{\hat{X}_{D_3,C_3}} - 2\mathbb{E}[(X - \mu_X)^T(\hat{X}_{D_3,C_3} - \mu_{\hat{X}_{D_3,C_3}})]. \end{aligned}$$

Thus,

$$\mathbb{E}[(X - \mu_X)^T(\hat{X}_{D_3,C_3} - \mu_{\hat{X}_{D_3,C_3}}) = \frac{\sigma_X^2 + \sigma^2_{\hat{X}_{D_3,C_3}} - D_3}{2}.$$

We now utilize the inequality $I(X;\mathbb{E}[X|\hat{X}_{D_3,C_3}]) \leq I(X;\hat{X}_{D_3,C_3}) = R(D_1,\infty)$, which implies: $\mathbb{E}[\|X - \mathbb{E}[X|\hat{X}_{D_3,C_3}]\|^2] \geq D_1$.

Observe that $\hat{X}_{D_3,C_3} - \mu_{\hat{X}_{D_3,C_3}}$ has a certain correlation with $X - \mu_X$, so we can use a linear predictor idea to get a simpler upper bound. By the orthogonality principle:

$$\begin{aligned}&\mathbb{E}[\|X - \mathbb{E}[X|\hat{X}_{D_3,C_3}]\|^2]\\ &\leq \mathbb{E}[\|X - \mu_X - c(\hat{X}_{D_3,C_3} - \mu_{\hat{X}_{D_3,C_3}})\|^2].\end{aligned}$$

The inequality says the best predictor $\mathbb{E}[X|\hat{X}_{D_3,C_3}]$ is no worse (in MSE) than any fixed linear predictor of the form $\mu_X + c(\hat{X}_{D_3,C_3} - \mu_{\hat{X}_{D_3,C_3}})$, where

$$\begin{aligned}c=&\frac{\text{Cov}(X - \mu_X, \hat{X}_{D_3,C_3} - \mu_{\hat{X}_{D_3,C_3}})}{\text{Var}(\hat{X}_{D_3,C_3} - \mu_{\hat{X}_{D_3,C_3}})},\\ =&\frac{\mathbb{E}[(X - \mu_X)^T(\hat{X}_{D_3,C_3} - \mu_{\hat{X}_{D_3,C_3}})]}{\sigma^2_{\hat{X}_{D_3,C_3}}}\\ =&\frac{\sigma_X^2 + \sigma^2_{\hat{X}_{D_3,C_3}} - D_3}{2\sigma^2_{\hat{X}_{D_3,C_3}}}.\end{aligned}$$

Therefore, the MSE of the optimal conditional expectation is at most the MSE of this linear estimator.

$$\begin{aligned}D_1 &\leq \mathbb{E}[\|X - \mathbb{E}[X|\hat{X}_{D_3,C_3}]\|^2],\\ &\leq \mathbb{E}[\|X - \mu_X - c(\hat{X}_{D_3,C_3} - \mu_{\hat{X}_{D_3,C_3}})\|^2],\\ &= \mathbb{E}[\|X - \mu_X\|^2] + c^2\mathbb{E}[\|\hat{X}_{D_3,P_3} - \mu_{\hat{X}_{D_3,C_3}}\|^2]\\ &- 2c\mathbb{E}[(X - \mu_X)^T(\hat{X}_{D_3,P_3} - \mu_{\hat{X}_{D_3,C_3}})],\\ &= \sigma_X^2 + c^2\sigma^2_{\hat{X}_{D_3,C_3}} - 2c\mathbb{E}[(X - \mu_X)^T(\hat{X}_{D_3,C_3} - \mu_{\hat{X}_{D_3,C_3}})],\\ &= \sigma_X^2 + \left(\frac{\sigma_X^2 + \sigma^2_{\hat{X}_{D_3,C_3}} - D_3}{2\sigma^2_{\hat{X}_{D_3,C_3}}}\right)^2 \sigma^2_{\hat{X}_{D_3,C_3}}\\ &- 2\left(\frac{\sigma_X^2 + \sigma^2_{\hat{X}_{D_3,C_3}} - D_3}{2\sigma^2_{\hat{X}_{D_3,C_3}}}\right)\left(\frac{\sigma_X^2 + \sigma^2_{\hat{X}_{D_3,C_3}} - D_3}{2}\right),\\ &= \sigma_X^2 - \frac{(\sigma_X^2 + \sigma^2_{\hat{X}_{D_3,C_3}} - D_3)^2}{4\sigma^2_{\hat{X}_{D_3,C_3}}}.\end{aligned}$$

Rearranging terms yields:

$$(\sigma_X^2 + \sigma^2_{\hat{X}_{D_3,C_3}} - D_3)^2 \leq 4\sigma^2_{\hat{X}_{D_3,C_3}}(\sigma_X^2 - D_1).$$

Under the assumptions $\sigma_X^2 + \sigma_{\hat{X}_{D_3,C_3}}^2 - D_3 \geq 0$ and $\sigma_X^2 - D_1 \geq 0$, it follows that:

$$\sigma_X^2 + \sigma_{\hat{X}_{D_3,C_3}}^2 - D_3 \leq 2\sigma_{\hat{X}_{D_3,C_3}} \sqrt{\sigma_X^2 - D_1},$$
$$D_3 \geq \sigma_X^2 + \sigma_{\hat{X}_{D_3,C_3}}^2 - 2\sigma_{\hat{X}_{D_3,C_3}} \sqrt{\sigma_X^2 - D_1}.$$

Based on [9], we can show that:

$$D^{(b)} \leq D_3 \leq 2D_1.$$

Hence,

$$D_3 - D^{(b)} \geq \sigma_X^2 + \sigma_{\hat{X}_{D_3,C_3}}^2 - 2\sigma_{\hat{X}_{D_3,C_3}} \sqrt{\sigma_X^2 - D_1} - 2D_1,$$
$$\frac{D_3}{D^{(b)}} \geq \frac{\sigma_X^2 + \sigma_{\hat{X}_{D_3,C_3}}^2 - 2\sigma_{\hat{X}_{D_3,C_3}} \sqrt{\sigma_X^2 - D_1}}{2D_1}.$$

For $D_1 = 0$:

$$D_3 - D^{(b)} \geq \sigma_X^2 + \sigma_{\hat{X}_{D_3,C_3}}^2 - 2\sigma_{\hat{X}_{D_3,C_3}} \sigma_X,$$

Moreover, in the case of $\sigma_X^2 = \sigma_{\hat{X}_{D_3,C_3}}^2$, we have:

$$D_3 - D^{(b)} \overset{D_1 \approx 0 \text{ and } \sigma_X^2 = \sigma_{\hat{X}_{D_3,C_3}}^2}{\approx} 0,$$

For $D_1 = \sigma_X^2$:

$$D_3 - D^{(b)} \geq \sigma_X^2 + \sigma_{\hat{X}_{D_3,C_3}}^2 - 2\sigma_X^2,$$
$$\frac{D_3}{D^{(b)}} \geq \frac{\sigma_X^2 + \sigma_{\hat{X}_{D_3,C_3}}^2}{2\sigma_X^2}.$$

Again, in the case of $\sigma_X^2 = \sigma_{\hat{X}_{D_3,C_3}}^2$, then:

$$D_3 - D^{(b)} \overset{D_1 \approx \sigma_X^2 \text{ and } \sigma_X^2 = \sigma_{\hat{X}_{D_3,C_3}}^2}{\approx} 0,$$
$$\frac{D_3}{D^{(b)}} \overset{D_1 \approx \sigma_X^2 \text{ and } \sigma_X^2 = \sigma_{\hat{X}_{D_3,C_3}}^2}{\approx} 1.$$

A similar method can be utilized to bound the upper-left corner of the curve, i.e., the gap between (D_1, C_1) and the upper-left extreme point $(\bar{D}^{(a)}, \bar{C}^{(a)})$ of the blue curve in Fig. 5. Let $\bar{D}^{(a)} = \mathbb{E}[\|X - \mathbb{E}[X \mid \hat{X}_{D_3,C_3}]\|^2]$. Then:

$$\bar{D}^{(a)} \leq \sigma_X^2 - \frac{(\sigma_X^2 + \sigma_{\hat{X}_{D_3,C_3}}^2 - D_3)^2}{4\sigma_{\hat{X}_{D_3,C_3}}^2},$$

which, together with $D_1 \geq \frac{1}{2}D_3$, yields:

$$\bar{D}^{(a)} - D_1 \leq \sigma_X^2 - \frac{(\sigma_X^2 + \sigma^2_{\hat{X}_{D_3,C_3}} - D_3)^2}{4\sigma^2_{\hat{X}_{D_3,C_3}}} - \frac{D_3}{2},$$

$$\frac{\bar{D}^{(a)}}{D_1} \leq \frac{\sigma_X^2 - \frac{(\sigma_X^2 + \sigma^2_{\hat{X}_{D_3,C_3}} - D_3)^2}{4\sigma^2_{\hat{X}_{D_3,C_3}}}}{D_3/2}.$$

For $D_3 = 0$:

$$\bar{D}^{(a)} - D_1 \leq \sigma_X^2 - \frac{(\sigma_X^2 + \sigma^2_{\hat{X}_{D_3,C_3}})^2}{4\sigma^2_{\hat{X}_{D_3,C_3}}},$$

If $\sigma_X^2 = \sigma^2_{\hat{X}_{D_3,C_3}}$, then:

$$\bar{D}^{(a)} - D_1 \overset{D_3 \approx 0 \text{ and } \sigma_X^2 = \sigma^2_{\hat{X}_{D_3,C_3}}}{\approx} 0,$$

For $D_3 = 2\sigma_X^2$:

$$\bar{D}^{(a)} - D_1 \leq \sigma_X^2 - \frac{(\sigma^2_{\hat{X}_{D_3,C_3}} - \sigma_X^2)^2}{4\sigma^2_{\hat{X}_{D_3,C_3}}} - \sigma_X^2,$$

$$\frac{\bar{D}^{(a)}}{D_1} \leq \frac{\sigma_X^2 - \frac{(\sigma^2_{\hat{X}_{D_3,C_3}} - \sigma_X^2)^2}{4\sigma^2_{\hat{X}_{D_3,C_3}}}}{\sigma_X^2}.$$

Similarly, in the case of $\sigma_X^2 = \sigma^2_{\hat{X}_{D_3,C_3}}$, we have:

$$\bar{D}^{(a)} - D_1 \overset{D_3 \approx 2\sigma_X^2 \text{ and } \sigma_X^2 = \sigma^2_{\hat{X}_{D_3,C_3}}}{\approx} 0,$$

$$\frac{\bar{D}^{(a)}}{D_1} \overset{D_3 \approx 2\sigma_X^2 \text{ and } \sigma_X^2 = \sigma^2_{\hat{X}_{D_3,C_3}}}{\approx} 1.$$

□

VI. Experimental Results

A. Illustration of No Rate Penalty for a Gaussian Source

We numerically validate the absence of a rate penalty when transitioning from multiple rate-specific Gaussian encoders (conventional model) to a single universal encoder. The experiment considers a scalar Gaussian source $X \sim \mathcal{N}(0, \sigma_X^2)$ and a classification variable $S \sim \mathcal{N}(0, \sigma_S^2)$, correlated via coefficient ρ. The goal is to evaluate the classification-distortion-rate function under mean squared error distortion and classification constraint $C = H(S \mid \hat{X})$.

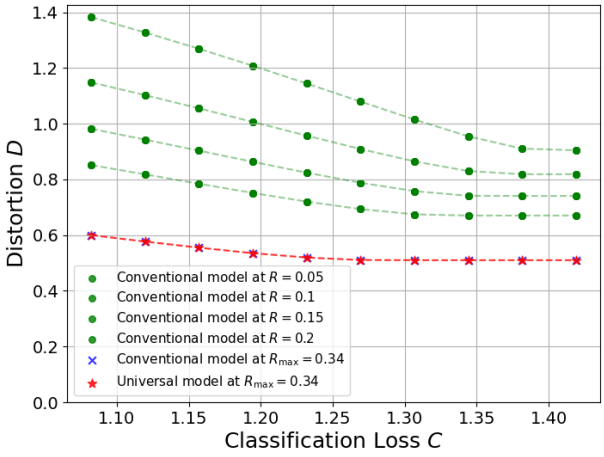


Figure 6: DRC function at various rates for a Gaussian source. The universal model achieves the same region as the conventional model, verifying Theorem 3.

We fix the variances $\sigma_X^2 = \sigma_S^2 = 1.0$ and set $\rho = 0.7$, resulting in a maximum achievable rate of $R_{\max} = \frac{1}{2}\log\left(\frac{1}{1-\rho^2}\right) = 0.34$. The conventional model is evaluated at five rate values: $[0.05, 0.1, 0.15, 0.2, 0.34]$, with the corresponding (C, D) tradeoffs derived using Theorem 2. For each rate, we compute the achievable region by varying the decoder and measuring the resulting classification and distortion performance. The universal model, guided by Theorem 3, constructs a single high-rate Gaussian representation at $R_{\max}$ and varies the decoder to explore the entire (C, D) space.

Figure 6 displays the achievable (C, D) regions under three settings. The green points correspond to the union of all conventional models, each operating at a different rate. The blue points denote a conventional encoder at $R = R_{\max}$. The red points represent the universal encoder, also fixed at $R_{\max}$, with varying decoders. Notably, the red points align precisely with the blue boundary, demonstrating that the universal model achieves the full CDR region without additional rate overhead. This confirms Theorem 3 and affirms the feasibility of universal representations for Gaussian sources under classification constraints.

B. Universal Encoder Representation for DL-based Lossy Compression

The rate-distortion-classification tradeoff was observed in deep learning-based image compression when classifier regularization was integrated into the training pipeline [3], [20]. In such

settings, achieving a specific point in the tradeoff space typically requires training a dedicated end-to-end model that jointly optimizes the encoder and decoder for that objective. However, this approach is often computationally expensive and inflexible for practical deployment.

To address this limitation, an alternative strategy is to reuse a pre-trained encoder while adapting only the decoder or classifier to meet varying task requirements. We refer to models that jointly train both encoder and decoder for each objective as *conventional models*, and those that retain a fixed encoder and retrain only the decoder as *universal models*. In our framework, universal models leverage encoders originally trained under the conventional paradigm. Under identical datasets and hyperparameters, the only difference between the two lies in whether the encoder parameters are updated during training.

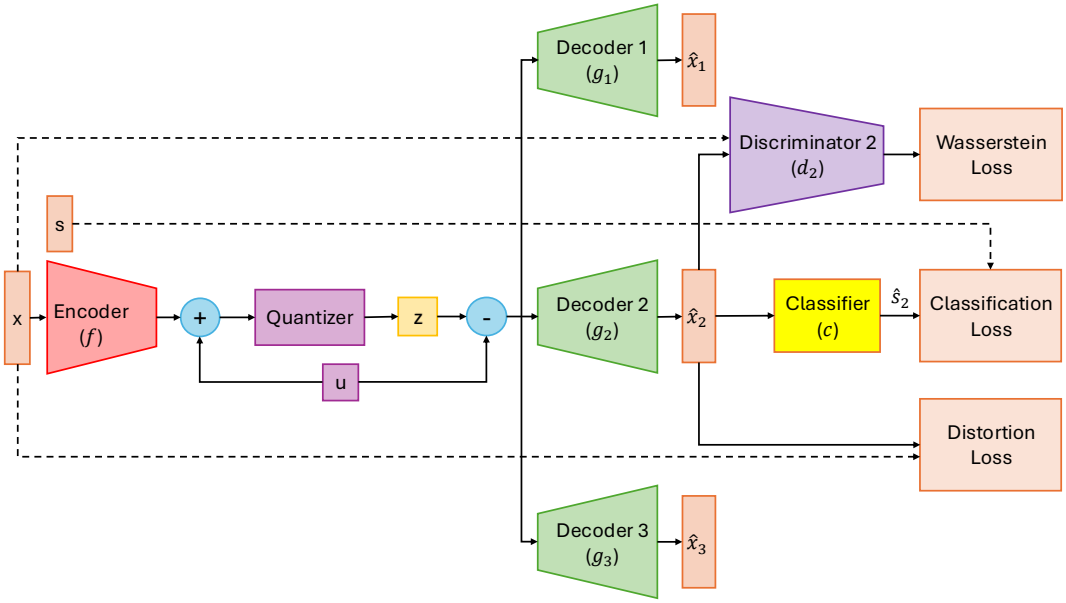


Figure 7: Figure illustrates the universal model setup. A single encoder f, trained for a specific classification-distortion tradeoff, is frozen and reused. Multiple decoders $\{g_i\}$ and discriminators $\{d_i\}$ are then trained independently using the fixed latent representation z. A shared randomness source u enables universal quantization, and a classifier C evaluates classification performance.

Training setup. We employ a stochastic autoencoder with a classifier and a GAN-based discriminator, comprising an encoder f, a decoder g, a classifier c, and a discriminator d. In the conventional setup, f, g, and d are trainable. The encoder maps input X to a latent representation $f(X) \in [-1, 1]^{\text{dim}}$ via a final `tanh` activation. This output is uniformly quantized into L levels per dimension, yielding an upper-bound compression rate of $R = \text{dim} \times \log_2 L$, as established in [8].

To perform quantization, we use dithered quantization [32], [33], assuming shared randomness $U \sim \mathcal{U}[-1/(L-1), 1/(L-1)]^{\text{dim}}$. The encoder outputs: $Z = \text{Quantize}(f(X)+U)$ and the decoder reconstructs $\hat{X} = g(Z-U)$. This approach centers quantization noise and enables gradient flow using the soft estimator from [34].

The distortion loss is measured by MSE. The output $\hat{X}$ is passed through the classifier c to produce the predicted label distribution $\hat{S} = c(\hat{X})$, with classification loss computed via cross-entropy $\text{CE}(S, \hat{S})$, an upper bound on conditional entropy, i.e. $H(S|\hat{X}) \leq \text{CE}(S, \hat{S})$ [3], [35]. To ensure that the condition $\sigma_X^2 = \sigma^2_{\hat{X}_{D_3,C_3}}$ in Theorem 5 is satisfied, we augment the training loss with a Wasserstein-1 distance regularization term, rather than relying solely on MSE and cross-entropy losses. Following the approach in [4], we use a discriminator d that takes both X and $\hat{X}$ as inputs and employs the Wasserstein-1 loss in a GAN-based framework.

The compression rate is upper bounded by $\text{dim} \times \log_2(L)$, where dim is the encoder output size and L the quantization level. The total loss is:

$$\mathcal{L} = \lambda_d \, \mathbb{E}[\|X - \hat{X}\|^2] + \lambda_c \, \text{CE}(S, \hat{S}) + \lambda_p \, W_1(p_X, p_{\hat{X}}), \tag{21}$$

where λ_d, λ_c, and λ_p controlling the tradeoffs.

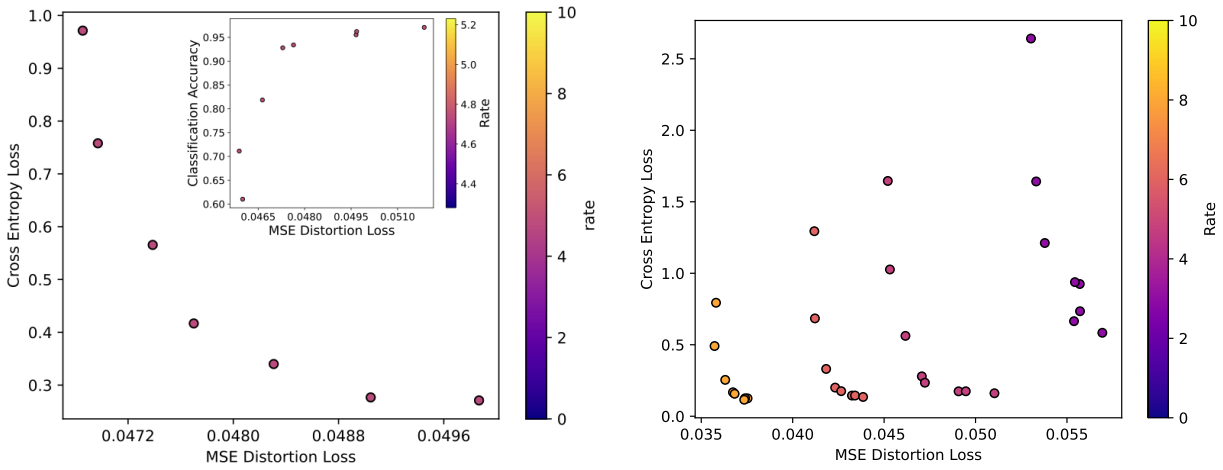
To construct the universal model, the trained encoder f is frozen, and a new decoder g_1 and discriminator d_1 are trained using:

$$\mathcal{L}_1 = \lambda_d^1 \, \mathbb{E}[\|X - \hat{X}_1\|^2] + \lambda_c^1 \, \text{CE}(S, \hat{S}) + \lambda_p^1 \, W_1(p_X, p_{\hat{X}}), \tag{22}$$

where λ_d^1, λ_c^1, and λ_p^1 adjust the task-specific tradeoffs. A schematic of the full system is shown in Figure 7.

C. Results

Figure 8 illustrates the RDC functions of conventional models on the MNIST dataset. The tradeoff between distortion and classification performance, quantified via cross-entropy loss and classification accuracy, is clearly depicted in Figure 8a at a fixed rate of $R = 4.75$, corresponding to an encoder output dimension of $\text{dim} = 3$ and quantization level $L = 3$. Figure 8b presents RDC curves across multiple rate configurations. The evaluated (dim, L) combinations include (3,2), (3,3), (3,4), and (4,4), providing a range of bit budgets for comparison. Each point in these figures corresponds to an encoder-decoder pair trained with a specific rate and loss configuration $(\lambda_d, \lambda_c, \lambda_p)$. Points sharing the same color indicate models trained under the same rate. The



(a) The RDC curve under rate $R = 4.75$.

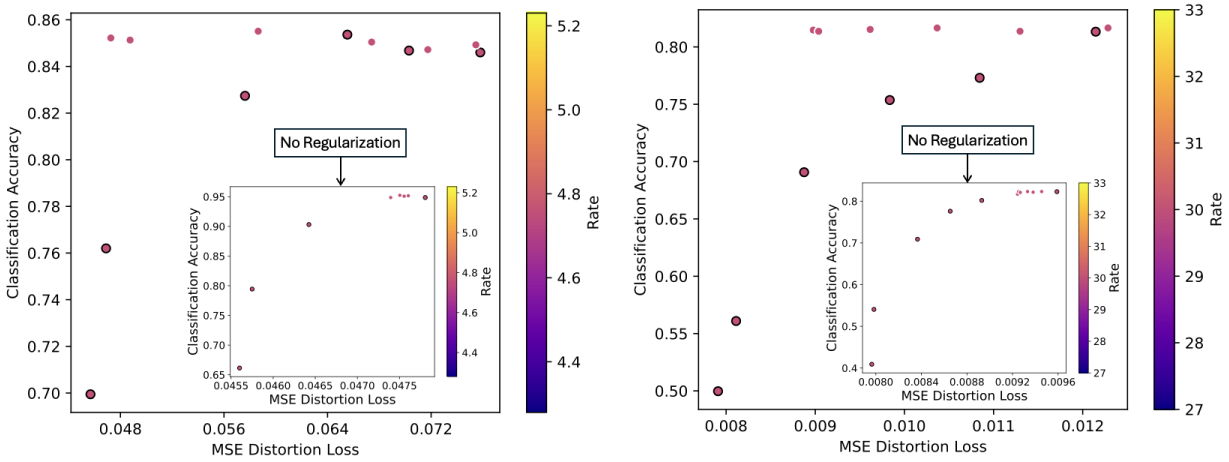
(b) The RDC curves at multiple rates.

Figure 8: RDC tradeoffs on the MNIST dataset using conventional models. (a) shows the tradeoff between distortion and classification, evaluated via cross-entropy loss or classification accuracy. (b) illustrates the RDC curves across multiple encoding rates.

results consistently reveal a tradeoff: under a fixed rate constraint, achieving lower classification loss (equivalently, improving classification accuracy) comes at the cost of increased distortion. Moreover, as the rate R increases, the entire distortion-classification curve shifts downward and to the left, signifying that higher-rate encoders can achieve both lower distortion and better classification accuracy. This trend highlights the value of allocating more bits to support multi-task compression objectives.

Figure 9 shows the RDC tradeoff on the MNIST dataset ($R = 4.75$) and SVHN dataset ($R = 30$), obtained by varying loss coefficients. Black-outlined points represent the conventional model trained jointly for specific classification-distortion objectives. Other points correspond to the universal model with decoders trained on a fixed encoder optimized for low classification loss C. The result corresponding to the fixed encoder trained at high C can be analyzed in a similar manner. Despite using a fixed encoder, the universal model achieves distortion levels comparable to the conventional model, confirming that an encoder trained for low C can still support diverse tradeoffs through decoder retraining. The subfigure illustrates the benefits of incorporating the regularization term. These observations support the validity of Theorem 5.

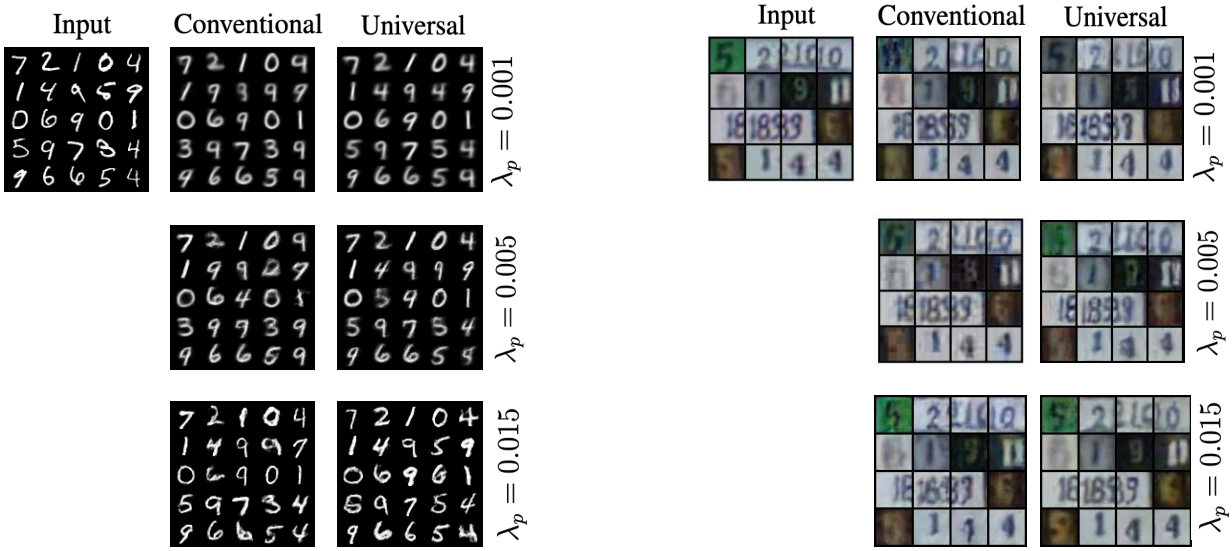
However, a noticeable classification gap remains: universal decoders cannot recover low classi-



(a) RDC curve on the MNIST dataset.

(b) RDC curve on the SVHN dataset.

Figure 9: RDC curves on the MNIST ($R = 4.75$) and SVHN ($R = 30$) datasets. The subfigures correspond to the case without the Wasserstein-1 regularization term.



(a) Decompressed outputs on the MNIST dataset.

(b) Decompressed outputs on the SVHN dataset.

Figure 10: Visualizations of decompressed outputs from selected models on the MNIST ($R = 4.75$) and SVHN ($R = 30$) datasets.

fication performance if the encoder is trained only for high-distortion objectives. This highlights the decoder's limited generative capacity when the encoder fails to preserve classification-task information.

Figure 10 presents qualitative decompression results from selected models on the MNIST and SVHN datasets at fixed compression rates of $R = 4.75$ and $R = 30$, respectively. During training, the weighting parameters were set to $\lambda_c = \lambda_p$. These examples illustrate the visual impact of increasing the Wasserstein-1 distance regularization weight λ_p. As λ_p increases, the encoder-decoder pair increasingly prioritizes minimizing $W_1(p_X, p_{\hat{X}})$, leading to a perceptual enhancement in reconstructed image quality. Specifically, the reconstructed digits and digit-background compositions appear sharper and less blurred, exhibiting improved perceptual realism. However, the reconstructed styles differ more noticeably from the original inputs, reflecting a higher distortion loss. This observation highlights the inherent tradeoff between fidelity and perceptual quality under a fixed rate constraint and underscores the importance of balancing multiple objectives in learned lossy compression frameworks.

VII. Conclusion

We proposed a universal rate-distortion-classification framework that enables a single encoder to support multiple task objectives through specialized decoders, removing the need for separate encoders per distortion-classification tradeoff. For the Gaussian source with MSE distortion, we proved that the full rate-distortion classification region is achievable with zero rate penalty using a fixed encoder. For the general source, we characterized the achievable region using MMSE estimation and the Wasserstein-2 distance, identifying conditions under which encoder reuse incurs negligible distortion penalty. Empirical results on the MNIST and SVHN datasets support our theory, showing that universal encoders, trained with Wasserstein loss regularization, achieve distortion performance comparable to task-specific models. These findings highlight the practicality and effectiveness of universal representations for multi-task lossy compression.

References

[1] N. Nguyen, T. Nguyen, T. Nguyen, and B. Bose, "Universal rate-distortion-classification representations for lossy compression," in *2025 IEEE Information Theory Workshop (ITW)*, 2025, pp. 1–6.

[2] Y. Blau and T. Michaeli, "Rethinking lossy compression: The rate-distortion-perception tradeoff," in *International Conference on Machine Learning*, 2019, pp. 675–685.

[3] Y. Wang, Y. Wu, S. Ma, and Y.-J. Angela Zhang, "Task-oriented lossy compression with data, perception, and classification constraints," *IEEE Journal on Selected Areas in Communications*, vol. 43, no. 7, pp. 2635–2650, 2025.

[4] G. Zhang, J. Qian, J. Chen, and A. Khisti, "Universal rate-distortion-perception representations for lossy compression," *IEEE Transactions on Information Theory*, pp. 1–1, 2025.

[5] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 1999.

[6] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[7] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. IEEE, 2003, pp. 1398–1402.

[8] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, "Generative adversarial networks for extreme learned image compression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 221–231.

[9] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6228–6237.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," vol. 27, 2014, pp. 2672–2680.

[11] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*, 2017, pp. 214–223.

[12] M. Tschannen, E. Agustsson, and M. Lucic, "Deep generative models for distribution-preserving lossy compression," in *Advances in Neural Information Processing Systems*, 2018, pp. 5929–5940.

[13] S. Nowozin, B. Cseke, and R. Tomioka, "f-gan: training generative neural samplers using variational divergence minimization," in *Advances in Neural Information Processing Systems*, 2016, pp. 271–279.

[14] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International Conference on Machine Learning*, 2016, pp. 1558–1566.

[15] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.

[16] A. Mittal, A. K. Moorthy, and A. C. Bovik, "Blind/referenceless image spatial quality evaluator," in *2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR)*. IEEE, 2011, pp. 723–727.

[17] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

[18] N. Venkatanath, D. Praneeth, M. C. Bh, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *2015 Twenty First National Conference on Communications (NCC)*. IEEE, 2015, pp. 1–6.

[19] D. Liu, H. Zhang, and Z. Xiong, "On the classification-distortion-perception tradeoff," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[20] Y. Zhang, "A rate-distortion-classification approach for lossy image compression," *Digital Signal Processing*, vol. 141, p. 104163, Sep. 2023.

[21] N. Saldi, T. Linder, and S. Yüksel, "Output constrained lossy source coding with limited common randomness," *IEEE Transactions on Information Theory*, vol. 61, no. 9, pp. 4984–4998, 2015.

[22] ——, "Randomized quantization and optimal design with a marginal constraint," in *2013 IEEE International Symposium on Information Theory*. IEEE, 2013, pp. 2349–2353.

[23] R. Zamir and K. Rose, "Natural type selection in adaptive lossy compression," *IEEE Transactions on information theory*, vol. 47, no. 1, pp. 99–111, 2001.

[24] S. Huang, A. Makhzani, Y. Cao, and R. Grosse, "Evaluating lossy compression rates of deep generative models," in *International Conference on Machine Learning*, 2020, pp. 4444–4454.

[25] R. Brekelmans, D. Moyer, A. Galstyan, and G. Ver Steeg, "Exact rate-distortion in autoencoders via echo noise," in *Advances in Neural Information Processing Systems*, 2019, pp. 3889–3900.

[26] A. Alemi, B. Poole, I. Fischer, J. Dillon, R. A. Saurous, and K. Murphy, "Fixing a broken elbo," in *International Conference on Machine Learning*, 2018, pp. 159–168.

[27] M. Tschannen, O. Bachem, and M. Lucic, "Recent advances in autoencoder-based representation learning," *arXiv preprint arXiv:1812.05069*, 2018.

[28] Y. Mroueh, T. Sercu, and V. Goel, "Mcgan: Mean and covariance feature matching gan," in *International Conference on Machine Learning*, 2017, pp. 2527–2535.

[29] F. Mentzer, G. D. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[30] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, "Information bottleneck for gaussian variables," in *Advances in Neural Information Processing Systems*, S. Thrun, L. Saul, and B. Schölkopf, Eds., vol. 16. MIT Press, 2003.

[31] T. Berger and R. Zamir, "A semi-continuous version of the berger-yeung problem," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1520–1526, 1999.

[32] R. M. Gray and T. G. Stockham, "Dithered quantizers," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 805–812, 1993.

[33] J. Ziv, "On universal quantization," *IEEE Transactions on Information Theory*, vol. 31, no. 3, pp. 344–347, 1985.

[34] F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. Van Gool, "Conditional probability models for deep image compression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4394–4402.

[35] M. Boudiaf, J. Rony, I. M. Ziko, E. Granger, M. Pedersoli, P. Piantanida, and I. B. Ayed, "A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses," *arXiv preprint*, 2021, arXiv 2003.08983.