
Task-Driven Discrete Representation Learning

Tung-Long Vuong
Monash University

Abstract

In recent years, deep discrete representation learning (DRL) has achieved significant success across various domains. Most DRL frameworks (e.g., the widely used VQ-VAE and its variants) have primarily focused on generative settings, where the quality of a representation is implicitly gauged by the fidelity of its generation. In fact, the goodness of a discrete representation remain ambiguously defined across the literature. In this work, we adopt a practical approach that examines DRL from a task-driven perspective. We propose a unified framework that explores the usefulness of discrete features in relation to downstream tasks, with generation naturally viewed as one possible application. In this context, the properties of discrete representations as well as the way they benefit certain tasks are also relatively understudied. We therefore provide an additional theoretical analysis of the trade-off between representational capacity and sample complexity, shedding light on how discrete representation utilization impacts task performance. Finally, we demonstrate the flexibility and effectiveness of our framework across diverse applications.

1 Introduction

Deep neural representation learning has been a key factor in the success of deep learning across various domains, including images (Pathak et al., 2016; Goodfellow et al., 2014; Kingma et al., 2016), videos, and audio (Reed et al., 2017; Oord et al., 2016; Kalchbrenner et al., 2017). The fundamental idea is to

use deep neural networks to learn compact, expressive representations from high-dimensional data, typically represented as continuous-valued vectors (Kingma & Welling, 2013; Chen et al., 2020; Chen & He, 2021). However, for certain tasks such as complex reasoning, planning, or predictive learning (Hafner et al., 2020; Allen et al., 2021; Ozair et al., 2021; Van Den Oord et al., 2017), it can be more advantageous to interpret non-identical but similar vectors as internal representations of shared discrete symbols from a fixed codebook, leading to the concept of discrete representation learning (Ozair et al., 2021; Hu et al., 2017).

In response, discrete representations have been actively explored and applied in various contexts. The pioneering, also most popular, unsupervised discrete representation learning framework is *Vector Quantization Variational Auto-Encoder* (VQ-VAE) (Van Den Oord et al., 2017), followed by its variants (Roy et al., 2018; Takida et al., 2022; Yu et al., 2021; Razavi et al., 2019; Williams et al., 2020; Vuong et al., 2023). (Van Den Oord et al., 2017) successfully extends VAE model for learning discrete latent representations. Concretely, the vector quantized models learn a compact discrete representation using a deterministic encoder-decoder architecture in the first stage, and subsequently apply this highly compressed representation for various downstream tasks, namely image generation (Esser et al., 2021), cross-modal translation (Kim et al., 2022), image recognition (Yu et al., 2021) or text-to-image generation (Ramesh et al., 2021; Rombach et al., 2022). Another line of research which can be categorized as discrete representation learning is *Prototypical Representation Learning* (Yarats et al., 2021; Ming et al., 2022; Lin et al., 2023; Ye et al., 2024), where discrete representations are used to model clusters, classes, or patterns in data in a supervised fashion. The role of prototypes varies across tasks. Ming et al. (2022) use prototypes for out-of-distribution (OOD) detection and data points are examined through their relation to these representations; prototypes are also considered as summaries, for instance, of an agent’s exploratory experiences (Yarats et al., 2021) or of continuous representations for classification tasks (Lin et al., 2023). The learning ob-

jectives also vary based on differences in usage. For example, Islam et al. (2022) encourage the alignment of augmented data representations with the original data while maximizing the information-theoretic dependency between the data and their predicted discrete representations. Alternatively, Liu et al. (2022) refines the prototype learning process by employing discrete information bottlenecks for learning structured representation spaces. From a prototypical perspective, any classification problem in fact can be viewed as a form of discrete representation learning, where continuous representations are classified based on their similarity to the class-specific vectors. This effectively transforms continuous features into discrete class labels by matching them to the most similar prototype.

Although deep discrete representation learning has been widely applied across various domains, the learning processes for discrete representations vary significantly between tasks, making it difficult to analyze their properties and advantages comprehensively. Much of the existing research on discrete representations focuses primarily on VQ-VAE and its variants. For instance, works such as (Roy et al., 2018; Takida et al., 2022; Yu et al., 2021; Zheng et al., 2022; Dhariwal et al., 2020; Williams et al., 2020) explore issues like codebook collapse and propose solutions to mitigate them. This highlights the need for a generalized framework that encapsulates discrete representation learning, along with theoretical analyses to better understand its properties.

Contribution. In this work, we introduce a generalized framework for learning task-driven deep discrete representations. We present a theoretical analysis on learned discrete representations, highlighting the trade-off between optimal task accuracy and sample complexity. Specifically, reducing the size of the codebook decreases the optimal accuracy but requires fewer samples to achieve generalization. On the other hand, increasing the codebook size improves accuracy but necessitates a larger sample size. We translate the theoretical results into practical insights about how to effectively optimize discrete latents in specific downstream tasks. We conduct various experiments to demonstrate the flexibility and generalizability of our framework where it achieves competitive performance across diverse applications.

Organization. The paper is structured as follows. First, we provide a general overview of representation learning and introduce the background on Wasserstein distance, which serves as the foundation for our proposed framework. In Section 3.1, we present the discrete representation framework and derive a trainable

form compatible with gradient backpropagation frameworks such as PyTorch. Section 3.2 contains a detailed analysis of our framework. Next, we conduct experiments in Section 4 to empirically validate our theoretical findings, followed by concluding remarks in Section 5. Additional results and detailed proofs are provided in the supplementary material.

2 Preliminaries

Notations. We first introduce the notations and basic concepts in the paper. We use calligraphic letters (i.e., \mathcal{X}) for spaces, upper case letters (i.e. X) for random variables, lower case letters (i.e. x) for their values and \mathbb{P} for (observed) probability distributions. We consider the task-driven setting where the observation distribution is $\mathbb{P}(x)$ (or \mathbb{P}_x) and labelling function $f_y(\cdot)$ is a deterministic mapping from x to the outcomes/labels the task of interest (e.g., a categorical label for a classification problem, a real value/vector for a regression problem).

General Representation Learning. The objective of representation learning is to learn the *encoder* $f_e : \mathcal{X} \rightarrow \mathcal{Z}$ first map the data examples $x \in \mathcal{X}$ to the latent $z \in \mathcal{Z}$. The latent representation $z = f_e(x)$ is expected to encapsulate task-related information from x and is then used to predict the task $f_y(x)$ using a decoder f_d . f_e, f_d usually is learned by the following optimization problem (OP):

$$\min_{f_d, f_e} \mathbb{E}_{x \sim \mathbb{P}_x} [\ell_y(f_d(f_e(x)), f_y(x))] \quad (1)$$

where $\ell_y(u, v)$ is the loss between u and v which is depended on our task of interest. For instance, ℓ_y is cross-entropy loss for classification task, L1 or L2 loss for regression task.

Optimal transport. Let $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m b_j \delta_{y_j}$ are discrete probability distributions where a_j and b_j represent discrete probability masses such that $\sum_{i=1}^n a_i = 1$ and $\sum_{j=1}^m b_j = 1$. The Kantorovich formulation of the OT distance (Kantorovich, 2006) between two discrete distributions α and β is

$$W_c(\mu, \nu) := \inf_{\mathbb{P} \in \Pi(\mu, \nu)} \langle \mathbf{C}, \mathbb{P} \rangle, \quad (2)$$

where $\langle \cdot, \cdot \rangle$ denotes the Frobenius dot-product; $\mathbf{C} \in \mathbb{R}_+^{n \times m}$ is the cost matrix of the transport; $\mathbb{P} \in \mathbb{R}_+^{n \times m}$ is the transport matrix/plan; $\Pi(\mu, \nu) := \{\mathbb{P} \in \mathbb{R}_+^{n \times m} : \mathbb{P} \mathbf{1}_n = \mu, \mathbb{P} \mathbf{1}_m = \nu\}$ is the transport polytope of μ and ν ; $\mathbf{1}_n$ is the n -dimensional column vector of ones. For arbitrary measures, Eq. (2) can be generalized as

$$W_c(\mu; \nu) := \inf_{\pi \sim \mathcal{P}(X \sim \mu, Y \sim \nu)} \mathbb{E}_{(x, y) \sim \pi} [c(x, y)], \quad (3)$$

where the infimum is taken over the set of all joint distributions $\mathcal{P}(X \sim \alpha, Y \sim \beta)$ with marginals α and β respectively. $c : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ is any measurable cost function. If $c(x, y) = D^p(x, y)$ is a distance for $p \leq 1$, W_p , the p -root of W_c , is called the p -Wasserstein distance. Finally, for a measurable map $T : \mathcal{X} \mapsto \mathcal{Y}$, $T\#\mu$ denotes the push-forward measure of μ that, for any measurable set $B \subset \mathcal{Y}$, satisfies $T\#\mu(B) = \mu(T^{-1}(B))$. For discrete measures, the push-forward operation is $T\#\mu = \sum_{j=1}^n a_j \delta_{T(x_j)}$.

3 Discrete Representation Learning

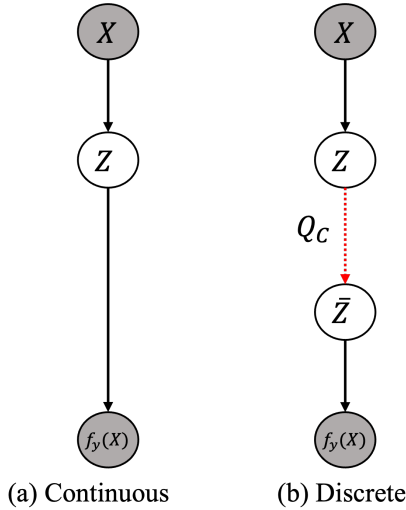


Figure 1: Graphical model: (a) continuous representation vs (b) discrete representation where discrete representation \bar{z} is constrained by set of M discrete representation i.e., $\bar{Z} \in C = \{c_1, \dots, c_M\}$.

Codebook. Unlike continuous representation learning, in the discrete representation learning (DRL) framework, the image of f_e is constrained to a finite set of discrete representations, commonly referred to as the *codebook*. The *codebook* consists of a set of codewords $C = \{c_m\}_{m=1}^M$, where each codeword belongs to the latent space \mathcal{Z} ($C \in \mathcal{Z}^M$).

Quantization. Denote the discrete representation function $\bar{f}_e = Q_C \circ f_e$ in which the *encoder* $f_e : \mathcal{X} \rightarrow \mathcal{Z}$ first map the data examples $x \in \mathcal{X}$ to the latent $z \in \mathcal{Z}$. Following, a quantization Q_C projecting z onto $C : \bar{z} = Q_C(z) = \operatorname{argmin}_{c \in C} d_z(z, c)$ is a quantization operator which d_z is a distance on \mathcal{Z} .

Objective function. Since the image of \bar{f}_e is constrained by C , from probabilistic viewpoint, we can endow discrete distribution $\mathbb{P}_{c,\pi} = \sum_{m=1}^M \pi_m \delta_{c_m}$ over the codebook C where the category probabilities $\pi \in \Delta_M = \{\alpha \geq 0 : \|\alpha\|_1 = 1\}$ lie in the M -simplex.

Then $\bar{f}_e : \mathcal{X} \mapsto \mathcal{Z}$ be a deterministic encoder such that $\bar{f}_e\#\mathbb{P}_x = \mathbb{P}_{c,\pi}$. The OP of discrete representation learning framework w.r.t training data \mathbb{P}_x can be formulated as the minimizer of following OP:

$$\min_{\mathbb{P}_{c,\pi}, f_d} \min_{\bar{f}_e : \bar{f}_e\#\mathbb{P}_x = \mathbb{P}_{c,\pi}} \mathbb{E}_{x \sim \mathbb{P}_x} [\ell_y(f_d(\bar{f}_e(x)), f_y(x))] \quad (4)$$

This OP explicitly impose the constraint on \bar{f}_e i.e., \bar{f}_e is a deterministic discrete encoder mapping a data example x directly to a codeword which prevent us from end-to-end training using gradient backpropagation. To make it trainable, we replace \bar{f}_e by a continuous encoder $f_e : \mathcal{X} \rightarrow \mathcal{Z}$ in the following theorem.

Theorem 3.1. *If we seek f_d and f_e in a family with infinite capacity (e.g., the family of all measurable functions), the the two OPs of interest in (4) is equivalent to the following OPs:*

$$\min_{\mathbb{P}_{c,\pi}} \min_{f_d, f_e} \mathbb{E}_{x \sim \mathbb{P}_x} [\ell_y(f_d(Q_C(f_e(x))), f_y(x))] + \lambda \mathcal{W}_{d_z}(f_e\#\mathbb{P}_x, \mathbb{P}_{c,\pi}), \quad (5)$$

where $\mathcal{W}_{d_z}(\cdot, \cdot)$ is wasserstein distance with d_z is the metric on representation space, the parameter $\lambda > 0$ and $Q_C(z) = \operatorname{argmin}_{c \in C} d_z(z, c)$.

This theorem transforms the constrained optimization problem in Eq. (4) into the unconstrained optimization problem in Eq. (5), which involves task performance $f_d(\cdot)$ from the codeword $Q_C(f_e(x))$ and regularization using the Wasserstein distance on the latent space. We note that although this theorem requires the function classes of f_e and f_d to have infinite capacity, practical considerations necessitate that we restrict ourselves to classes of finite capacity for learning to be feasible. Therefore, we interpret (5) as a practical realization that converges to (4) under ideal conditions.

3.1 Theoretical Analysis

In the previous section, we introduced a general framework for discrete representation learning, where the output of the representation function is constrained by the codebook $C = \{c_m\}_{m=1}^M$. The size of this codebook plays a crucial role in DRL framework, as it controls the expressiveness of the representations. To provide a more comprehensive understanding of OP in Eq. (4), we examine how the codebook size influences the performance upper-bound in Eq.(4), assuming an infinite data scenario. Furthermore, we investigate the sample complexity, specifically determining the number of samples required to achieve a low approximation error for the performance upper-bound. This analysis will be discussed in detail in the subsequent section.

3.1.1 Accuracy Trade-off

We would like to examine how the codebook size M affects the expected loss in Eq. (4), as outlined in the following Lemma:

Lemma 3.2. *Let $C^* = \{c_m^*\}_{m=1}^M, \pi^*$, and f_d^* be the optimal solution of the OP in Eq. (4), then $C^* = \{c_m^*\}_k, \pi^*$, and f_d^* are also the optimal solution of the following OP:*

$$\epsilon_M^* = \min_{f_d} \min_{\pi} \min_{\sigma \in \Sigma_{\pi}} \mathbb{E}_{x \sim \mathbb{P}_x} [\ell_y(f_y(x), f_d(c_{\sigma(x)}))]$$

and

$$\epsilon_M^* \geq \epsilon_{M+1}^* \quad \forall M$$

and

$$\epsilon_M^* \rightarrow 0 \text{ as } M \rightarrow \infty$$

where Σ_{π} is the set of assignment functions $\sigma : \mathcal{X} \rightarrow \{1, \dots, M\}$ such that $\mathbb{P}_x(\sigma^{-1}(m)), m = 1, \dots, M$ are proportional to $\pi_k, m = 1, \dots, M$ (e.g., σ is the nearest assignment: $\sigma^{-1}(m) = \{x \mid f_e(x) = c_m, m = \operatorname{argmin}_m d_z(f_e(x), c_m)\}$ is set of latent representations which are quantized to m^{th} codeword).

Lemma 3.2 states that for the optimal solution $C^* = \{c_k^*\}_{m=1}^M, \pi^*, f_e^*$, and f_d^* of the OP in (4), each $f_y(x)$ is assigned to the centroid $f_d^*(c_{\sigma^*}^*(x)) = f_d^*(Q_C(f_e^*(x)))$ which forms optimal clustering centroids of the optimal clustering solution minimizing the loss of interest. This lemma also demonstrates that constraining the image of the representation function by the codebook $C = \{c_m\}_{m=1}^M$ leads to a decrease in performance. The smaller the value of M the lower the optimal accuracy that can be achieved.

Remark.1 (Multi-level Discrete Representation) Real-world data often exhibits a natural hierarchical structure, where data can be grouped at varying levels of granularity. Lemma 3.2 ensures that we obtain the best description of the data given a predefined codebook size, and this description becomes more detailed as the codebook size increases. This enables the construction of multi-level discrete representations, allowing us to explore the data at different levels of detail. For instance, with a small codebook size, you can analyze global patterns, while a larger codebook size allows you to zoom in to examine finer groupings and more specific relationships within the data.

3.1.2 Sample Complexity

Given a specific codebook size M and data distribution \mathbb{P}_x , solving OP in Eq. (4) provides the optimal accuracy corresponding to M and \mathbb{P}_x . However, in practice, models are often trained on a finite set of samples drawn i.i.d. from an underlying distribution.

We wish that models trained on finite samples can perform well even on previously unseen samples from the underlying distribution. Therefore, we focus on analyzing the gap between the optimal solution obtained from the training data and its performance on both the training data and the underlying distribution.

Let the training set be $S = \{x_1, \dots, x_N\}$ consisting of N sample $x_i \sim \mathbb{P}_x$ and define the empirical measure \mathbb{P}_x^N by placing mass N^{-1} at each of members of S i.e., $\mathbb{P}_x^N = \sum_{i=1}^N \frac{1}{N} \delta_{x_i}$. Given $C^* = \{c_m^*\}_{m=1}^M, \pi^*, f_e^*$, and f_d^* be the optimal solution of the OP in Eq. (12) w.r.t to empirical training data \mathbb{P}_x^N , the empirical loss w.r.t to distribution \mathbb{P}_x^N is defined as:

$$\mathcal{L}(\mathbb{P}_x^N) = \mathbb{E}_{x \sim \mathbb{P}_x^N} [\ell(f_y(x), f_d^*(Q_C(f_e^*(x))))], \quad (6)$$

Notably, Lemma 3.2 establishes that solving the optimization problem (OP) in Eq. (4) is equivalent to solving a clustering problem in the output space. However, discrete representations exhibit distinct characteristics. Specifically, defining $\bar{f}_y(x) = f_d^*(\operatorname{argmin}_{c'' \in C} \ell_y(x, f_d^*(c'')))$, we obtain:

$$\begin{aligned} \mathcal{L}(\mathbb{P}_x^N) &= \mathbb{E}_{x \sim \mathbb{P}_x^N} [\ell(f_y(x), f_d^*(\bar{f}_e(x)))] \\ &= \underbrace{\mathbb{E}_{x \sim \mathbb{P}_x^N} [\ell(\bar{f}_y(x), f_y(x))]}_{\text{Clustering Loss: } \mathcal{L}_C(\mathbb{P}_x^N)} \\ &\quad + \underbrace{\mathbb{E}_{x \sim \mathbb{P}_x^N} [\ell(f_d^*(\bar{f}_e^*(x)), f_y(x)) - \ell(\bar{f}_y(x), f_y(x))]}_{\text{Assignment Discrepancy Loss: } \mathcal{L}_A(\mathbb{P}_x^N)} \\ &= \mathcal{L}_C(\mathbb{P}_x^N) + \mathcal{L}_A(\mathbb{P}_x^N) \end{aligned}$$

The overall error is decomposed into two components: **Clustering Loss**: The first term $\ell(\bar{f}_y(x), f_y(x))$ represents the distance between the label $f_y(x)$ and its nearest centroid in the output space, which is determined by mapping the codebook to the output space.

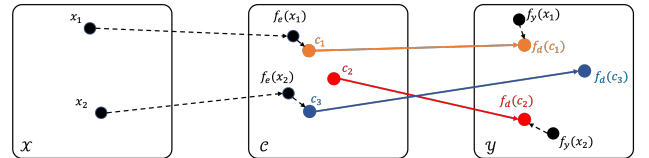


Figure 2: Mismatching between the nearest assignments in the latent and output spaces.

Assignment Discrepancy Loss: It can be seen that $\bar{f}_e^*(x)$ is the nearest assignment on latent space while $\bar{f}_y(x) = f_d^*(\operatorname{argmin}_{c'' \in C} \ell_y(x, f_d^*(c'')))$ is the nearest assignment on output space. The second term i.e., the discrepancy between $\ell(f_d^*(\bar{f}_e^*(x)), f_y(x)) - \ell(\bar{f}_y(x), f_y(x))$ arises from the mismatch between assignments in the latent and output spaces. Specifically, in Figure 2, for sample x , the decoded centroid $f_d(c_3)$

in the latent space may not be the closest centroid in the output space.

Following that, the general loss is defined similarly:

$$\mathcal{L}(\mathbb{P}_x) = \mathcal{L}_C(\mathbb{P}_x) + \mathcal{L}_A(\mathbb{P}_x)$$

The discrepancy between the empirical loss and the general loss can be bounded as follows:

$$\begin{aligned} & |\mathcal{L}(\mathbb{P}_x^N) - \mathcal{L}(\mathbb{P}_x)| \\ &= |\mathcal{L}_C(\mathbb{P}_x^N) + \mathcal{L}_A(\mathbb{P}_x^N) - \mathcal{L}_C(\mathbb{P}_x) - \mathcal{L}_A(\mathbb{P}_x)| \\ &\leq |\mathcal{L}_C(\mathbb{P}_x^N) - \mathcal{L}_C(\mathbb{P}_x)| + |\mathcal{L}_A(\mathbb{P}_x^N) - \mathcal{L}_A(\mathbb{P}_x)| \end{aligned}$$

The first term $|\mathcal{L}_C(\mathbb{P}_x^N) - \mathcal{L}_C(\mathbb{P}_x)|$ is the generalization bounds for k-mean problem, which is studied in (Telgarsky & Dasgupta, 2013; Bachem et al., 2017). In this study, our focus is on the second term, $|\mathcal{L}_A(\mathbb{P}_x^N) - \mathcal{L}_A(\mathbb{P}_x)|$, which examines the impact of discrete representations on the optimization problem.

Corollary 3.3. *Given a dataset of size N , a loss function ℓ_y that is a proper metric and upper-bounded by a positive constant L , and an encoder class \mathcal{F}_e (i.e., $\tilde{f}_e \in \mathcal{F}_e$), then it holds with probability at least $1 - \delta$ that*

$$|\mathcal{L}_A(\mathbb{P}_x^n) - \mathcal{L}_A(\mathbb{P}_x)| \leq \epsilon \quad (7)$$

where $\epsilon = LM \frac{\sqrt{\log 1/\delta + \log |\mathcal{F}_e| + 1}}{\sqrt{N}}$.

This bound shows that, given $C^* = \{c_m^*\}_{m=1}^M, \pi^*, \sigma^*$, and f_d^* as the optimal solution to the optimization problem in Eq. (4) with respect to the empirical training data \mathbb{P}_x^N , and apply this solution to ground-truth underlying distribution \mathbb{P}_x , the number of samples N required to achieve a generalization error of at most ϵ with high probability $1 - \delta$ is given by: $N(\epsilon, \delta) = \Omega\left(\frac{L^2 M^2}{\epsilon^2} \max(\log |\mathcal{E}|/\delta, 1)\right)$. This analysis indicates that the convergence rate of $|\mathcal{L}_A(\mathbb{P}_x^N) - \mathcal{L}_A(\mathbb{P}_x)|$ is $\mathcal{O}(N^{-1/2})$, which aligns with the convergence rate of the clustering generalization error $|\mathcal{L}_C(\mathbb{P}_x^N) - \mathcal{L}_C(\mathbb{P}_x)|$ as established in (Bachem et al., 2017). Consequently, the overall convergence rate of $|\mathcal{L}(\mathbb{P}_x^N) - \mathcal{L}(\mathbb{P}_x)|$ also follows $\mathcal{O}(N^{-1/2})$.

Moreover, the sample complexity $N(\epsilon, \delta)$ is influenced by the codebook size M , scaling as $\Omega(M^2)$, whereas in (Bachem et al., 2017), it scales as $\Omega(M \log M)$ in (Bachem et al., 2017). Therefore, the sample complexity for the original loss $|\mathcal{L}(\mathbb{P}_x^N) - \mathcal{L}(\mathbb{P}_x)|$ is also $\Omega(M^2)$. This implies that when M is small, fewer samples are required for the model to generalize effectively, whereas larger values of M demand a significantly higher number of samples.

Remark.2 (Sample Complexity Trade-off) From the two analyses above, we can observe a trade-off between optimal accuracy and sample complex-

ity. Specifically, reducing the size of the codebook decreases the optimal accuracy, but requires fewer samples to guarantee the generalization. Conversely, increasing the codebook size improves accuracy but demands more samples. The trade-off give us an hint choose appropriate codebook size depend on our task of interest. Specifically, for some that that not require very detail representation, we can use small codebook size to gain advantage from small sample complexity.

4 Experiments

In this section, we conduct experiments to validate two key properties of our framework. Specifically, we focus on: (i) **sample complexity trade-off and performance Trade-off** (Section 4.1): we apply state abstraction problem in reinforcement learning (RL) setting as a natural use case of discrete representation in RL domains to showcase the trade-off properties. By aggregating states into the same abstract states, similar action distributions are produced, enabling abstract states to be parameterized as discrete representations. We show that discrete-state representations can improve the sample efficiency of RL agents, requiring fewer samples to achieve comparable performance; (ii) **The multi-level discrete representation** (Section 4.2): we explore the representation-alignment problem in the domain generalization (DG) setting, where the goal is to align representation distributions across different domains to learn representations that are invariant to domain shifts. Our experiments demonstrate that employing multi-level discrete representations enhances the representation alignment process, leading to improved performance.

4.1 State Abstraction RL

Setting. We consider the RL environment characterized by a Block Markov Decision Process (MDP) defined by the tuple $\langle \mathcal{X}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma \rangle$, representing the set of states, actions, reward function, transition functions, and discount factor, respectively. Block MDP assumes that a true, hidden state space \mathcal{S} exists with a much smaller and more compact representation than the observed states in \mathcal{X} . In state abstraction RL, we aim to learn an abstraction function ϕ that maps the grounded states $x \in \mathcal{X}$ to an abstract state $z \in \mathcal{Z}$, where \mathcal{Z} is considered an abstract set of \mathcal{X} ; that is, we want to recover \mathcal{S} through \mathcal{Z} by learning an abstraction model. For our demonstration, we incorporate our discrete learning framework into the method of learning continuous Markov abstraction (Allen et al., 2021), which learns an abstraction mapping ϕ from the observation space to abstract state space while preserving the Markov property of the original environment.

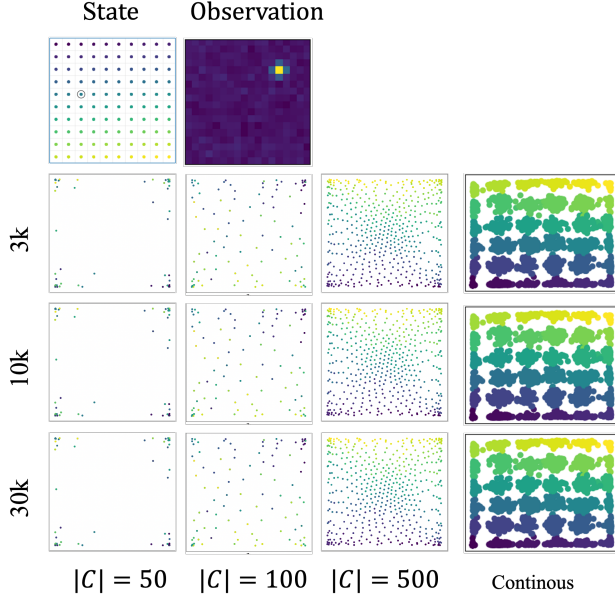


Figure 3: **(Top row)** The visualization true underlying state on the left, the observations which are noisy version of true state in the middle and the task of RL agent i.e., moving from random point to goal (red point) in the right. **(From second row)** The visualization of the learned latent codebook; from second top to bottom using $3k$, $10k$, and $30k$ samples for training; from left to right: discrete representation with the codebook sizes of 50, 100, 500 and continuous representation from the algorithm of (Allen et al., 2021).

Discrete Representation for State Abstraction.

Since there is no reconstruction in an RL problem, we need to define the task f_y . Given image observation as inputs, the encoder f_e is defined as a feature extractor network, with the output z from the encoder quantized to the nearest codeword through the quantization operator Q_C as in Eq (5), which together form the discrete abstraction model $\phi := \hat{f}_e = Q_C \circ f_e$. The task is to learn the Markov abstraction state through the weighted sum of the Inverse loss and Ratio loss defined in (Allen et al., 2021) as follows.

$$\mathcal{L}_{\text{inv}} = - \sum_i \log f(a = a_i | \phi(x), \phi(x')) \quad (8)$$

$$\mathcal{L}_{\text{ratio}} = - \sum_i \log(y = y_i | \phi(x), \phi(\tilde{x})) \quad (9)$$

The inverse loss trains a network to predict the actions taken a_i , given the current and the next abstraction $\phi(x)$, $\phi(x')$, and the ratio loss trains another network to predict whether two arbitrary abstractions $\phi(x)$, $\phi(\tilde{x})$ are from two consecutive grounded states, with y_i the binary labels denote if the two states are consecutive. Thus, unlike in the reconstruction tasks,

the decoders f_d here are defined as the action prediction network in the inverse model and the binary classifier of the next states for estimating density ratios. The labels corresponding to $f_y(x)$ in (5) are the binary labels y_i , and the actions a_i .

Task. We evaluate our framework using the offline setting in Visual Gridworld environment from (Allen et al., 2021). The input image observation of the Gridworld encapsulates the underlying states (Figure 3 (top row)). An abstraction model is initially trained on offline experiences collected through random exploration without any reward signals. Once trained, the model is frozen, and a DQN agent is trained on the learned latent space using reward signals to navigate from a random starting point to the goal (red point). We compare the agent’s performance when using an abstraction model learned through the discrete representation framework versus its continuous counterpart.

Stage-1. Figure 3 (starting from the second row) illustrates the representations learned by using discrete-state encoding across different codebook sizes and training sample counts, compared to the continuous-state representation of the abstraction model (where the continuous representation is considered as a discrete one with a very large codebook) with different number of training sample and codebook sizes. Our experiments show that when the latent codebook size exceeds the number of underlying discrete states, the model successfully reconstructs a meaningful latent structure, reducing noise compared to continuous representations. Even when the environment’s state space is larger than the available codebook, the model still learns reasonable representations by organizing similar states into neighboring regions in the latent space.

Stage-2. Figure 4 presents the performance of the DQN agent using the state representation learned from stage-1 to evaluate the quality of abstraction model. Comparing discrete-state and continuous-state representations, we observe that with a limited number of samples (e.g., $3k$ samples in 4.a and $10k$ samples in Figure 4.b), the discrete-state representation outperforms the continuous-state representation. However, with sufficient training data, the continuous-state representation ultimately achieves the best performance. This highlights the trade-off between representation capacity and sample complexity.

When comparing discrete representations, as shown in Figure 4.b with $10k$ samples, it is evident that performance improves as the codebook size increases, given a sufficient number of samples.

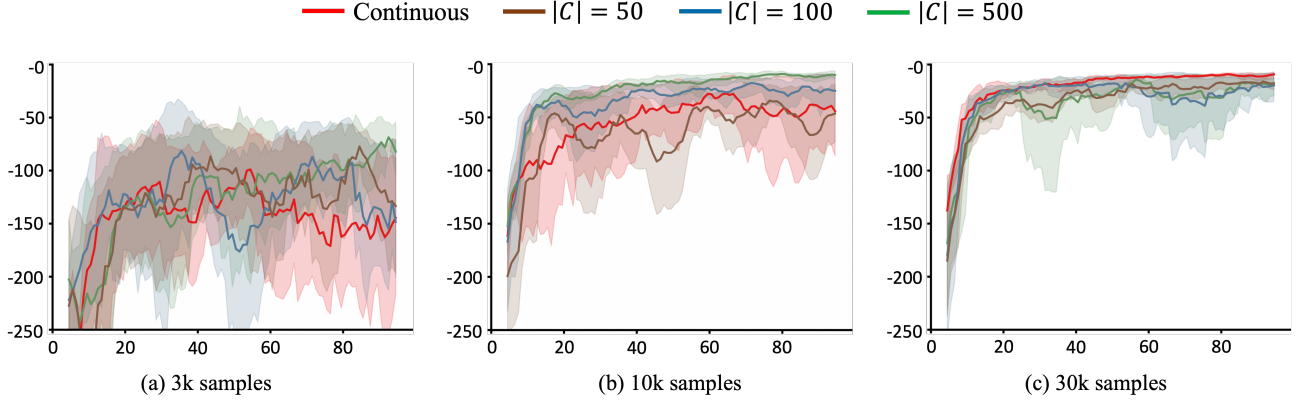


Figure 4: Results on 10×10 gridworld environment, tested with different codebook sizes and training samples.

4.2 Domain Generalization

Setting. Given E training source domains $\{\mathbb{P}^e(x, y)\}_{e=1}^E$. The objective of DG is to exploit the ‘commonalities’ present in the training domains to improve generalization to any unseen test domain.

Domain Representation Alignment. Various approaches have been proposed to address the DG problem (Long et al., 2017; Ganin et al., 2016; Li et al., 2018a; Gong et al., 2016; Li et al., 2018c; Tachet des Combes et al., 2020; Arjovsky et al., 2020; Krueger et al., 2021; Mitrovic et al., 2020; Wang et al., 2022; Yao et al., 2022), in the experiment, we focus in representation alignment-based approach which aim to learn a representation function f_e for data X such that $f_e(X)$ is invariant or consistent across different domains. Two representative work on this approach are DANN (Ganin et al., 2016) which have focused on learning such domain-invariant representations by reducing the divergence between latent marginal distributions $\mathbb{E}[f_e(X)|E]$ where E represents a domain variable and CDANN (Li et al., 2018b) which aligns the conditional distributions $\mathbb{E}[f_e(X)|Y = y, E]$ across domains.

Let encoder f_e , classifier f_d , and domain discriminator D , the objective for DANN and CDANN can be formulated as follows:

$$\min_{f_e, f_d, D} \sum_{e=1}^E \mathcal{L}_W(\mathbb{P}^e) + \mathcal{L}_D(\mathbb{P}^e) \quad (10)$$

where

- $\mathcal{L}_W(\mathbb{P}^e) = \mathbb{E}_{(x, y) \sim \mathbb{P}^e} [\ell_y(f_d(f_e(x)), y)]$
- $\mathcal{L}_D(\mathbb{P}^e) = \mathbb{E}_{(x, y) \sim \mathbb{P}^e} [\ell_y(D(\mathcal{R}(f_e(x))), \phi(x)), e)]$

with ℓ_y is cross-entropy-loss; \mathcal{R} is the Gradient Reversal Layer (GRL) as introduced by Ganin et al. (2016);

$\phi(x) = f_y(x) = y$ for CDANN and *empty* for DANN.

In the above objective, the first term \mathcal{L}_W is the standard classification loss, while the second term \mathcal{L}_D is the *representation-alignment loss*. Specifically, in the second term, the domain discriminator D aims to accurately predict the domain label “ e ” based on the combined feature $(f_e(x), m)$. However, with the incorporation of GRL, the gradient that backpropagates from the loss function to the encoder f_e is reversed in the opposite direction. Concurrently, this loss enforces that the representation function f_e transforms the input x into a latent representation $z = f_e(x)$ in such a way that D is unable to determine the domain “ e ” of z , resulting in a domain-invariant representation.

However, achieving true invariance is both challenging and potentially too restrictive. In real-world scenarios, where feature distributions are often multimodal, even feature distributions inner each class. Therefore, directly aligning entire feature distributions or class-conditional distributions across domains can be particularly difficult for adversarial networks. Recent studies (Goodfellow et al., 2014; Che et al., 2016) have shown that adversarial networks are prone to failure by aligning only a subset of the distribution’s components (modes), leaving other parts mismatched. To address this problem, our idea here is using set of discrete representation to capture modalities of training-domain distribution, then aligning representation in modality-wise manner. In the following, we demonstrate how to achieve fine-grained modalities of the training-domain distribution by leveraging the multi-level discrete representation property established in Lemma 3.2.

Fine-grained Representation Alignment. First, let’s reformulate the K -class classification problem within our discrete framework. We consider the classifier on top of the encoder as a linear layer represented by the weight matrix $f_d = W = [w_k]_{k=1}^K$, where w_k is

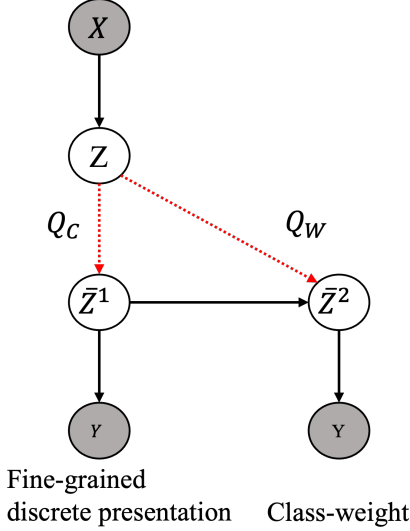


Figure 5: Fine-Grained Discrete Representation: \bar{Z}^2 corresponds to class-level weights, capturing information at the class level, while \bar{Z}^1 corresponds to fine-grained discrete representations, characterizing data at a more detailed, modality-specific level.

the vector corresponding to k^{th} class. The prediction via softmax activation is given by:

$$P(y = k | x) = \frac{\exp(f_e(x)^\top w_k)}{\sum_{j=1}^K \exp(f_e(x)^\top w_j)}$$

Therefore, given that f_e^* and W^* are the optimal solutions for the classification problem, i.e., $f_e^*, W^* = \operatorname{argmin}_{f_e, W} \sum_{e=1}^E \mathcal{L}_W(\mathbb{P}^e)$, the classification process for a given sample-label pair (x, y) is essentially performed by mapping the representation $z = f_e^*(x)$ to the closest class-vector $w_{k=y}^*$ using the metric $d_z(z, w_k) = \exp(f_e(x)^\top w_k)$ and applying the decoder $f_d(w_{k=y}^*) = y$ (mapping class-vector w_y to its index). Consequently, we can formulate the K -class classification problem as a discrete representation problem¹, where $W = \{w_1, \dots, w_K\}$ is the set of discrete representations, capturing the characteristics of the original data, i.e., a high-level clustering where all samples in a cluster share the same class label $f_y(x)$.

Our goal is to further capture the multiple modalities within the feature distribution of each class. To achieve this, we propose learning an additional set of fine-grained discrete representations $C = [c_1, \dots, c_M]$ with $M > K$, thereby increasing the codebook size to enhance representation capacity (as shown in Lemma 3.2). This process is illustrated in Figure 5,

¹We demonstrate the equivalence between k -class classification problem and the discrete representation learning presented in Eq. (4) in Appendix 6.4.

where $z = f_e(x)$ is the learned representation of input x , and z is simultaneously quantized using both the codebook C , i.e., $\bar{z}^1 = Q_C(z)$, and the codebook W , i.e., $\bar{z}^2 = Q_W(z)$. Note that, by OP in Eq (4), both $f_d(\bar{z}^1)$ and $f_d(\bar{z}^2)$ should correctly predict label $f_y(x)$ (the first term of Eq (4)). Consequently, the continuous representation z , which are quantized to \bar{z}^1 , should belong to the same class as \bar{z}^2 , forming a sub-class representing by \bar{z}^2 , effectively partitioning each class into several sub-classes, where each sub-class represents a distinct modality of the class’s feature distribution.

Since second level discrete representation is the classification, hence we have the final objective as follows:

$$\min_{f_e, f_d, d} \sum_{e=1}^E \mathcal{L}_W(\mathbb{P}^e) + \mathcal{L}_C(\mathbb{P}^e) + \mathcal{L}_D(\mathbb{P}^e) \quad (11)$$

where

- $\mathcal{L}_W(\mathbb{P}^e) = \mathbb{E}_{(x,y) \sim \mathbb{P}^e} [\ell_y(f_d(f_e(x)), y)]$
- $\mathcal{L}_C(\mathbb{P}^e) = \left[\mathbb{E}_{(x,y) \sim \mathbb{P}^e} [\ell_y(f_d(Q_C(f_e(x))), y)] + \lambda \mathcal{W}_{d_z}(f_e \# \mathbb{P}^e(x), \mathbb{P}_{c,\pi}) \right]$
- $\mathcal{L}_D(\mathbb{P}^e) = \mathbb{E}_{(x,y) \sim \mathbb{P}^e} [\ell_y(D(\mathcal{R}(f_e(x)), \phi(x)), e)]$

Compared to the objective in Eq. (10), an additional term, \mathcal{L}_C , is introduced to learn a set of discrete representations $C = \{c_1, \dots, c_M\}$, where $\phi(x) = \operatorname{argmin}_m d_z(f_e(x), c_m)$ is the index of the discrete representation c_m to which sample x is assigned. This leads to a key distinction in the alignment strategy: : DANN aligns the entire domain representation, CDANN aligns class-conditional representations, while this approach utilizes fine-grained discrete representations $C = \{c_1, \dots, c_M\}$, rather than class-level representations $W = \{w_1, \dots, w_K\}$ for conditional alignment, we refer to this method as FDANN.

Table 1: Classification Accuracy on **VLCS** using ResNet50. Results are reported over 5 random seeds.

Algorithm	C	L	S	V	Avg
ERM	98.8 ± 0.1	63.3 ± 0.3	75.3 ± 0.5	79.2 ± 0.6	79.1
DANN	99.2 ± 0.1	63.0 ± 0.8	75.3 ± 1.8	79.3 ± 0.5	79.2
CDANN	99.1 ± 0.1	63.3 ± 0.7	75.1 ± 0.7	80.1 ± 0.2	79.3
FDANN	98.9 ± 0.2	63.7 ± 0.3	75.6 ± 0.4	79.4 ± 0.8	79.4

Table 2: Classification Accuracy on **PACS** using ResNet50. Results are reported over 5 random seeds.

Algorithm	A	C	P	S	Avg
ERM	89.3 ± 0.2	83.4 ± 0.6	97.3 ± 0.3	82.5 ± 0.5	88.1
DANN	90.7 ± 1.2	82.2 ± 0.4	97.3 ± 0.1	81.6 ± 0.4	87.9
CDANN	90.5 ± 0.3	82.4 ± 1.0	97.6 ± 0.1	80.4 ± 0.3	87.7
FDANN	90.5 ± 0.5	83.4 ± 0.2	97.8 ± 0.1	83.2 ± 0.2	88.7

Table 3: Classification Accuracy on **OfficeHome** using ResNet50. Results are reported over 5 random seeds.

Algorithm	A	C	P	R	Avg
ERM	66.1 \pm 0.4	57.7 \pm 0.4	78.4 \pm 0.1	80.2 \pm 0.2	70.6
DANN	67.2 \pm 0.1	56.2 \pm 0.1	78.6 \pm 0.2	80.0 \pm 0.5	70.5
CDANN	66.8 \pm 0.4	56.4 \pm 0.8	78.4 \pm 0.5	80.1 \pm 0.2	70.4
FDANN	69.1 \pm 0.6	58.4 \pm 0.8	79.5 \pm 0.2	81.4 \pm 0.3	72.1

Benchmark. Table 1, Table 2 and Table 3 present the results of our experiments, comparing FDANN with baseline ERM (Gulrajani & Lopez-Paz, 2021), DANN and CDANN across three datasets VLCS, PACS and OfficeHome from the DomainBed benchmark (Gulrajani & Lopez-Paz, 2021) (the detailed settings are provided in Appendix 7). Our method consistently outperforms the baseline approaches on all datasets, demonstrating the effectiveness of aligning representations based on the modalities captured by the set of fine-grained discrete representations.

Table 4: Classification Accuracy on **PACS** with different codebook size using ResNet50. Results are reported over 5 random seeds.

M	A	C	P	S	Avg
CDANN	90.5 \pm 0.3	82.4 \pm 1.0	97.6 \pm 0.1	80.4 \pm 0.3	87.7
$K \times 4$	90.2 \pm 0.3	83.2 \pm 0.7	97.9 \pm 0.2	82.3 \pm 1.5	88.2
$K \times 8$	90.5 \pm 0.8	83.8 \pm 0.6	97.6 \pm 0.3	82.1 \pm 1.8	88.7
$K \times 16$	90.5 \pm 0.5	83.4 \pm 0.2	97.8 \pm 0.1	83.2 \pm 0.2	88.7
$K \times 32$	90.2 \pm 0.5	83.8 \pm 0.8	97.3 \pm 0.4	82.0 \pm 1.2	88.4

Effect of Codebook Size. We conducted an additional ablation study to further investigate the effect of codebook size on performance. It is important to note that CDANN can be considered a special case of our FDANN when $M = K \times 1$. As shown in Table 4, performance generally improves as the codebook size increases. However, when K becomes excessively large (i.e., $M = K \times 32$), a decline is observed, likely due to trade-offs in the optimization process.

5 Limitations and Conclusion

In this work, we introduce a generalized framework for learning deep discrete representations from a task-driven perspective. We provide a theoretical analysis of the learned discrete representations, emphasizing the trade-off between optimal task accuracy and sample complexity. Specifically, reducing the codebook size lowers the achievable accuracy but requires fewer samples to generalize effectively. Conversely, increasing the codebook size enhances accuracy but demands a larger sample size. We translate these theoretical findings into practical insights for optimizing discrete

latents in specific downstream tasks. Finally, we perform extensive experiments to validate our analyses. However, our theoretical framework relies on the use of Wasserstein distances to align latent spaces. The computational complexity of calculating the Wasserstein distance is well-documented and can introduce significant overhead to the learning process. In future work, we aim to explore alternative methods to develop a more efficient learning framework.

Acknowledgements

We sincerely thank our colleagues Bang Giang Le, Viet Cuong Ta, Trung Le, and Dinh Phung for their valuable discussions and support throughout the paper writing and experiment implementation.

References

- Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep.*, 32:96, 2019.
- Cameron Allen, Neev Parikh, Omer Gottesman, and George Konidaris. Learning markov state abstractions for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 34:8229–8241, 2021.
- Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020. URL <https://arxiv.org/abs/1907.02893>.
- Olivier Bachem, Mario Lucic, S Hamed Hassani, and Andreas Krause. Uniform deviation bounds for k-means clustering. In *International conference on machine learning*, pp. 283–291. PMLR, 2017.
- Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473, 2018.
- Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.
- Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li. Mode regularized generative adversarial networks. *arXiv preprint arXiv:1612.02136*, 2016.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

- Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.
- Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.
- Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, 2021.
- Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.
- Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pp. 2839–2848. PMLR, 2016.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.
- Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.
- Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.
- Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *International conference on machine learning*, pp. 1558–1567. PMLR, 2017.
- Riashat Islam, Hongyu Zang, Manan Tomar, Aniket Didolkar, Md Mofijul Islam, Samin Yeasar Arnob, Tariq Iqbal, Xin Li, Anirudh Goyal, Nicolas Heess, et al. Representation learning in deep rl via discrete information bottleneck. *arXiv preprint arXiv:2212.13835*, 2022.
- Nal Kalchbrenner, Aaron Oord, Karen Simonyan, Ivo Danihelka, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. Video pixel networks. In *International Conference on Machine Learning*, pp. 1771–1779. PMLR, 2017.
- LK Kantorovich. On a problem of monge. *Journal of Mathematical Sciences*, 133(4), 2006.
- Taehoon Kim, Gwangmo Song, Sihaeng Lee, Sangyun Kim, Yewon Seo, Soonyoung Lee, Seung Hwan Kim, Honglak Lee, and Kyunghoon Bae. L-verse: Bidirectional generation between image and text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16526–16536, 2022.
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29, 2016.
- David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghui Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.
- Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.
- Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5400–5409, 2018a.
- Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018b.
- Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 624–639, 2018c.
- Ci-Siang Lin, Min-Hung Chen, and Yu-Chiang Frank Wang. Frequency-aware self-supervised long-tailed learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 963–972, 2023.

- Dianbo Liu, Alex Lamb, Xu Ji, Pascal Notsawo, Mike Mozer, Yoshua Bengio, and Kenji Kawaguchi. Adaptive discrete communication bottlenecks with dynamic vector quantization. *arXiv preprint arXiv:2202.01334*, 2022.
- Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *arXiv preprint arXiv:1705.10667*, 2017.
- Yifei Ming, Yiyao Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? *arXiv preprint arXiv:2203.04450*, 2022.
- Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922*, 2020.
- Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.
- Sherjil Ozair, Yazhe Li, Ali Razavi, Ioannis Antonoglou, Aaron Van Den Oord, and Oriol Vinyals. Vector quantized models for planning. In *international conference on machine learning*, pp. 8302–8313. PMLR, 2021.
- Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.
- Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415, 2019.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.
- Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.
- Scott Reed, Aäron Oord, Nal Kalchbrenner, Sergio Gómez Colmenarejo, Ziyu Wang, Yutian Chen, Dan Belov, and Nando Freitas. Parallel multi-scale autoregressive density estimation. In *International conference on machine learning*, pp. 2912–2921. PMLR, 2017.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.
- Aurko Roy, Ashish Vaswani, Arvind Neelakantan, and Niki Parmar. Theory and experiments on vector quantized autoencoders. *arXiv preprint arXiv:1805.11063*, 2018.
- Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.
- Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J Gordon. Domain adaptation with conditional distribution matching and generalized label shift. *Advances in Neural Information Processing Systems*, 33:19276–19289, 2020.
- Yuhta Takida, Takashi Shibuya, Weihsiang Liao, Chieh-Hsin Lai, Junki Ohmura, Toshimitsu Uesaka, Naoki Murata, Shusuke Takahashi, Toshiyuki Kumakura, and Yuki Mitsufuji. SQ-VAE: Variational Bayes on discrete representation with self-annealed stochastic quantization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20987–21012. PMLR, 17–23 Jul 2022.
- Matus J Telgarsky and Sanjoy Dasgupta. Moment-based uniform deviation bounds for k -means and friends. *Advances in Neural Information Processing Systems*, 26, 2013.
- Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528. IEEE, 2011.
- Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
- Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027, 2017.
- Tung-Long Vuong, Trung Le, He Zhao, Chuanxia Zheng, Mehrtash Harandi, Jianfei Cai, and Dinh Phung. Vector quantized wasserstein auto-encoder. *arXiv preprint arXiv:2302.05917*, 2023.
- Ruoyu Wang, Mingyang Yi, Zhitang Chen, and Shengyu Zhu. Out-of-distribution generalization with causal invariant transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 375–385, 2022.

- Will Williams, Sam Ringer, Tom Ash, David MacLeod, Jamie Dougherty, and John Hughes. Hierarchical quantized autoencoders. *Advances in Neural Information Processing Systems*, 33:4524–4535, 2020.
- Xufeng Yao, Yang Bai, Xinyun Zhang, Yuechen Zhang, Qi Sun, Ran Chen, Ruiyu Li, and Bei Yu. Pcl: Proxy-based contrastive learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7097–7107, 2022.
- Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Reinforcement learning with prototypical representations. In *International Conference on Machine Learning*, pp. 11920–11931. PMLR, 2021.
- Hangting Ye, Wei Fan, Xiaozhuang Song, Shun Zheng, He Zhao, Dandan Guo, and Yi Chang. Ptarl: Prototype-based tabular representation learning via space calibration. *arXiv preprint arXiv:2407.05364*, 2024.
- Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. In *International Conference on Learning Representations*, 2021.
- Chuanxia Zheng, Long Tung Vuong, Jianfei Cai, and Dinh Phung. Movq: Modulating quantized vectors for high-fidelity image generation. *Advances in Neural Information Processing Systems*, 35, 2022.

Appendix

The appendix provides supplementary material and additional details to support the main findings and methods proposed in this paper. It is organized into several sections:

Section 6 provides the proofs for the theoretical results discussed in the main paper. Specifically:

- Section 6.1 presents the proof of Theorem 3.1, which transforms the constrained discrete representation learning framework into a trainable form.
- Section 6.2 provides the proof for the Lemma 3.2 on the accuracy trade-off.
- Section 6.3 includes the Corollary 3.3 on sample complexity.
- Finally, Section 6.4 demonstrates the equivalence between the common classification problem and the discrete representation framework.

Section 7 outlines the details of the experimental setup, including datasets and hyperparameters. The source code is provided in the supplementary material.

6 Theoretical development

In this Section, we present all proofs relevant to theory developed in our paper.

Notations. We use calligraphic letters (i.e., \mathcal{X}) for spaces, upper case letters (i.e. X) for random variables, lower case letters (i.e. x) for their values and \mathbb{P} for (observed) probability distributions. We consider the task-driven setting where the observation distribution is $\mathbb{P}(x)$ (or \mathbb{P}_x) and labelling function $f_y(\cdot)$ is a deterministic mapping from x to the outcomes/labels the task of interest (e.g., a categorical label for a classification problem, a real value/vector for a regression problem).

Objective function. The OP of discrete representation learning framework w.r.t training data \mathbb{P}_x is formulated as the minimizer of following OP:

$$\min_{\mathbb{P}_{c,\pi}, f_d} \min_{\bar{f}_e: \bar{f}_e \# \mathbb{P}_x = \mathbb{P}_{c,\pi}} \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d(\bar{f}_e(x)), f_y(x))] \quad (12)$$

Where

- $\mathbb{P}_{c,\pi} = \sum_{m=1}^M \pi_m \delta_{c_m}$ is the discrete distribution over the codebook $C = \{c_m\}_{m=1}^M$ where the category probabilities $\pi \in \Delta_M = \{\alpha \geq 0 : \|\alpha\|_1 = 1\}$ lie in the M -simplex.
- The discrete representation function $\bar{f}_e = Q_C \circ f_e$ in which the *encoder* $f_e : \mathcal{X} \rightarrow \mathcal{Z}$ first map the data examples $x \in \mathcal{X}$ to the latent $z \in \mathcal{Z}$. Following, a quantization Q_C projecting z onto $C : \bar{z} = Q_C(z) = \operatorname{argmin}_{c \in C} d_z(z, c)$ is a quantization operator which d_z is a distance on \mathcal{Z} .
- The decoder f_d using discrete representation \bar{z} to make prediction and ℓ is the loss of interest.

6.1 Proof theorem 3.1 in the main paper

Lemma 6.1. Consider C, π, f_d , and f_e as a feasible solution of the OP in Eq. (12). Let us denote $\bar{f}_e(x) = \operatorname{argmin}_{c \in C} d_z(f_e(x), c) = Q_C(x)$, then $\bar{f}_e(x)$ is a Borel measurable function.

Proof. We denote the set A_k on the latent space as

$$A_k = \{z : d_z(z, c_k) < d(z, c_j), \forall j \neq k\} = \{z : Q_C(z) = c_k\}.$$

A_k is known as a Voronoi cell w.r.t. the metric d_z . If we consider a continuous metric d_z , A_k is a measurable set. Given a Borel measurable function B , we prove that $\bar{f}_e^{-1}(B)$ is a Borel measurable set on the data space.

Let $B \cap \{c_1, \dots, c_K\} = \{c_{i_1}, \dots, c_{i_m}\}$, we prove that $\bar{f}_e^{-1}(B) = \cup_{j=1}^m f_e^{-1}(A_{i_j})$. Indeed, take $x \in \bar{f}_e^{-1}(B)$, then $\bar{f}_e(x) \in B$, implying that $\bar{f}_e(x) = Q_C(x) = c_{i_j}$ for some $j = 1, \dots, m$. This means that $f_e(x) \in A_{i_j}$ for some $j = 1, \dots, m$. Therefore, we reach $\bar{f}_e^{-1}(B) \subset \cup_{j=1}^m f_e^{-1}(A_{i_j})$.

We now take $x \in \cup_{j=1}^m f_e^{-1}(A_{i_j})$. Then $f_e(x) \in A_{i_j}$ for $j = 1, \dots, m$, hence $\bar{f}_e(x) = Q_C(x) = c_{i_j}$ for some $j = 1, \dots, m$. Thus, $\bar{f}_e(x) \in B$ or equivalently $x \in \bar{f}_e^{-1}(B)$, implying $\bar{f}_e^{-1}(B) \supset \cup_{j=1}^m f_e^{-1}(A_{i_j})$.

Finally, we reach $\bar{f}_e^{-1}(B) = \cup_{j=1}^m f_e^{-1}(A_{i_j})$, which concludes our proof because f_e is a measurable function and A_{i_j} are measurable sets. \square

Theorem 6.2. (Theorem 3.1 in the main paper) *If we seek f_d and f_e in a family with infinite capacity (e.g., the family of all measurable functions), the the two OPs of interest in (12) is equivalent to the following OP*

$$\min_{\mathbb{P}_{c,\pi}} \min_{f_d, f_e} \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d(Q_C(f_e(x))), f_y(x))] + \lambda \mathcal{W}_{d_z}(f_e \# \mathbb{P}_x, \mathbb{P}_{c,\pi}), \quad (13)$$

where and the parameter $\lambda > 0$ and $Q_C(z) = \operatorname{argmin}_{c \in C} d_z(z, c)$

Proof. Given the optimal solution $C^{*1}, \pi^{*1}, f_d^{*1}$, and f_e^{*1} of the OP in (12), we conduct the optimal solution for the OP in (18). Let us conduct $C^{*2} = C^{*1}, f_d^{*2} = f_d^{*1}$. We next define $\bar{f}_e^{*2}(x) = \operatorname{argmin}_c d_z(f_e^{*1}(x), c) = Q_{C^{*1}}(f_e^{*1}(x)) = Q_{C^{*2}}(f_e^{*1}(x))$. We prove that $C^{*2}, \pi^{*2}, f_d^{*2}$, and \bar{f}_e^{*2} are optimal solution of the OP in (18). By this definition, we yield $\bar{f}_e^{*2} \# \mathbb{P}_x = \mathbb{P}_{C^{*2}, \pi^{*2}}$ and hence $\mathcal{W}_{d_z}(\bar{f}_e^{*2} \# \mathbb{P}_x, \mathbb{P}_{C^{*2}, \pi^{*2}}) = 0$. Therefore, we need to verify the following:

- (i) \bar{f}_e^{*2} is a Borel-measurable function.
- (ii) Given a feasible solution C, π, f_d , and \bar{f}_e of (18), we have

$$\mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d^{*2}(\bar{f}_e^{*2}(x)), f_y(x))] \leq \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d(\bar{f}_e(x)), f_y(x))]. \quad (14)$$

We derive as

$$\begin{aligned} & \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d^{*2}(\bar{f}_e^{*2}(x)), f_y(x))] + \lambda \mathcal{W}_{d_z}(\bar{f}_e^{*2} \# \mathbb{P}_x, \mathbb{P}_{C^{*2}, \pi^{*2}}) \\ &= \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d^{*2}(\bar{f}_e^{*2}(x)), x)] \\ &= \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d^{*1}(Q_{C^{*2}}(f_e^{*1}(x))), f_y(x))] \\ &= \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d^{*1}(Q_{C^{*1}}(f_e^{*1}(x))), f_y(x))] \\ &\leq \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d^{*1}(Q_{C^{*1}}(f_e^{*1}(x))), f_y(x))] + \lambda \mathcal{W}_{d_z}(f_e^{*1} \# \mathbb{P}_x, \mathbb{P}_{C^{*1}, \pi^{*1}}). \end{aligned} \quad (15)$$

Moreover, because $\bar{f}_e \# \mathbb{P}_x = \mathbb{P}_{C,\pi}$ which is a discrete distribution over the set of codewords C , we obtain $Q_C(\bar{f}_e(x)) = \bar{f}_e(x)$. Note that C, π, f_d , and \bar{f}_e is also a feasible solution of (12) because \bar{f}_e is also a specific encoder mapping from the data space to the latent space, we achieve

$$\begin{aligned} & \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d(Q_C(\bar{f}_e(x))), f_y(x))] + \lambda \mathcal{W}_{d_z}(\bar{f}_e \# \mathbb{P}_x, \mathbb{P}_{C,\pi}) \\ &\geq \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d^{*1}(Q_{C^{*1}}(\bar{f}_e^{*1}(x))), f_y(x))] + \lambda \mathcal{W}_{d_z}(\bar{f}_e^{*1} \# \mathbb{P}_x, \mathbb{P}_{C^{*1}, \pi^{*1}}). \end{aligned}$$

Noting that $\bar{f}_e \# \mathbb{P}_x = \mathbb{P}_{C,\pi}$ and $Q_C(\bar{f}_e(x)) = \bar{f}_e(x)$, we arrive at

$$\begin{aligned} & \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d(\bar{f}_e(x)), f_y(x))] \\ &\geq \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d^{*1}(Q_{C^{*1}}(\bar{f}_e^{*1}(x))), f_y(x))] + \lambda \mathcal{W}_{d_z}(\bar{f}_e^{*1} \# \mathbb{P}_x, \mathbb{P}_{C^{*1}, \pi^{*1}}). \end{aligned} \quad (16)$$

Combining the inequalities in (15) and (16), we obtain Inequality (14) as

$$\mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d^{*2}(\bar{f}_e^{*2}(x)), f_y(x))] \leq \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d(\bar{f}_e(x)), f_y(x))]. \quad (17)$$

This concludes our proof. \square

6.2 Proof Lemma 3.2 in the main paper: Accuracy Trade-off

We first demonstrate the equivalence between the optimization problem in Eq. (12) and the Wasserstein distance between the prediction distribution $f_d \# \mathbb{P}_{c,\pi}$ and the label distribution $f_y \# \mathbb{P}_x$, as outlined in Proposition 1.3. Following this, we relate this result to a clustering perspective (Accuracy Trade-off Lemma).

Proposition 6.3. *We can equivalently reformulate the OP in Eq. (12) as minimizing the Wasserstein distance between the prediction distribution $f_d \# \mathbb{P}_{c,\pi}$ and the label distribution $f_y \# \mathbb{P}_x$:*

$$\min_{C,\pi,f_d} \min_{\bar{f}_e: \bar{f}_e \# f_y \# \mathbb{P}_x = \mathbb{P}_{c,\pi}} \mathbb{E}_{x \sim \mathbb{P}_x, c \sim \bar{f}_e(x)} [\ell(f_d(c), x)] = \min_{C,\pi,f_d} \mathcal{W}_\ell(f_d \# \mathbb{P}_{c,\pi}, f_y \# \mathbb{P}_x), \quad (18)$$

Proof. We first prove that

$$\mathcal{W}_\ell(f_d \# \mathbb{P}_{c,\pi}, f_y \# \mathbb{P}_x) = \min_{\bar{f}_e: \bar{f}_e \# f_y \# \mathbb{P}_x = \mathbb{P}_{c,\pi}} \mathbb{E}_{x \sim \mathbb{P}_x, c \sim \bar{f}_e(x)} [\ell(f_d(c), f_y(x))], \quad (19)$$

where \bar{f}_e is a **stochastic discrete** encoder mapping data example x directly to the codebooks.

Let \bar{f}_e be a **stochastic discrete** encoder such that $\bar{f}_e \# \mathbb{P}_x = \mathbb{P}_{c,\pi}$ (i.e., $x \sim \mathbb{P}_x$ and $c \sim \bar{f}_e(x)$ implies $c \sim \mathbb{P}_{c,\pi}$). We consider $\gamma_{d,c}$ as the joint distribution of (x, c) with $x \sim \mathbb{P}_x$ and $c \sim \bar{f}_e(x)$. We also consider $\gamma_{f_c,d}$ as the joint distribution including $(y, y') \sim \gamma_{f_c,d}$ where $x \sim \mathbb{P}_x, c \sim \bar{f}_e(x)$, $y = f_y(x)$ and $y' = f_d(c)$. This follows that $\gamma_{f_c,d} \in \Gamma(f_d \# \mathbb{P}_{c,\pi}, f_y \# \mathbb{P}_x)$ which admits $f_d \# \mathbb{P}_{c,\pi}$ and $f_y \# \mathbb{P}_x$ as its marginal distributions. We also have:

$$\begin{aligned} \mathbb{E}_{x \sim \mathbb{P}_x, c \sim \bar{f}_e(x)} [\ell(f_d(c), f_y(x))] &= \mathbb{E}_{(x,c) \sim \gamma_{d,c}} [\ell(f_d(c), f_y(x))] \stackrel{(1)}{=} \mathbb{E}_{(y,y') \sim \gamma_{f_c,d}} [\ell(y, y')] \\ &\geq \min_{\gamma_{f_c,d} \in \Gamma(f_d \# \mathbb{P}_{c,\pi}, f_y \# \mathbb{P}_x)} \mathbb{E}_{(y,y') \sim \gamma_{f_c,d}} [\ell(y, y')] \\ &= \mathcal{W}_\ell(f_d \# \mathbb{P}_{c,\pi}, f_y \# \mathbb{P}_x). \end{aligned}$$

Note that we have the equality in (1) due to $(f_y, f_d) \# \gamma_{d,c} = \gamma_{f_c,d}$ (both f_y and f_d are deterministic functions).

Therefore, we reach

$$\min_{\bar{f}_e: \bar{f}_e \# \mathbb{P}_x = \mathbb{P}_{c,\pi}} \mathbb{E}_{x \sim \mathbb{P}_x, c \sim \bar{f}_e(x)} [\ell(f_d(c), f_y(x))] \geq \mathcal{W}_\ell(f_d \# \mathbb{P}_{c,\pi}, f_y \# \mathbb{P}_x).$$

Let $\gamma_{f_c,d} \in \Gamma(f_y \# \mathbb{P}_x, f_d \# \mathbb{P}_{c,\pi})$. Let $\gamma_{f_c,c} \in \Gamma(f_d \# \mathbb{P}_{c,\pi}, \mathbb{P}_{c,\pi})$ be a deterministic coupling such that $c \sim \mathbb{P}_{c,\pi}$ and $y = f_d(c)$ imply $(c, y) \sim \gamma_{f_c,c}$. Using the gluing lemma (see Lemma 5.5 in (Santambrogio, 2015)), there exists a joint distribution $\alpha \in \Gamma(\mathbb{P}_{c,\pi}, f_d \# \mathbb{P}_{c,\pi}, f_y \# \mathbb{P}_x)$ which admits $\gamma_{f_c,d}$ and $\gamma_{f_c,c}$ as the corresponding joint distributions. By denoting $\gamma_{d,c} \in \Gamma(f_y \# \mathbb{P}_x, \mathbb{P}_{c,\pi})$ as the marginal distribution of α over $f_y \# \mathbb{P}_x, \mathbb{P}_{c,\pi}$, we then have

$$\begin{aligned} \mathbb{E}_{(y,y') \sim \gamma_{f_c,d}} [\ell(y, y')] &= \mathbb{E}_{(c,y',y) \sim \alpha} [\ell(y, y')] \\ &= \mathbb{E}_{(c,y) \sim \gamma_{d,c}, y' = f_d(c)} [\ell(y, y')] \\ &= \mathbb{E}_{(c,y) \sim \gamma_{d,c}} [\ell(f_d(c), y)] \\ &= \mathbb{E}_{x \sim \mathbb{P}_x, c \sim \bar{f}_e(x)} [\ell(f_d(c), f_y(x))] \\ &\geq \min_{\bar{f}_e: \bar{f}_e \# \mathbb{P}_x = \mathbb{P}_{c,\pi}} \mathbb{E}_{x \sim \mathbb{P}_x, c \sim \bar{f}_e(x)} [\ell(f_d(c), f_y(x))], \end{aligned}$$

where $\bar{f}_e(x) = \gamma_{d,c}(\cdot | x)$.

This follows that

$$\begin{aligned} \mathcal{W}_\ell(f_d \# \mathbb{P}_{c,\pi}, f_y \# \mathbb{P}_x) &= \min_{\gamma_{f_c,d} \in \Gamma(f_d \# \mathbb{P}_{c,\pi}, f_y \# \mathbb{P}_x)} \mathbb{E}_{(y,y') \sim \gamma_{f_c,d}} [\ell(y, y')] \\ &\geq \min_{\bar{f}_e: \bar{f}_e \# \mathbb{P}_x = \mathbb{P}_{c,\pi}} \mathbb{E}_{x \sim \mathbb{P}_x, c \sim \bar{f}_e(x)} [\ell(f_d(c), f_y(x))]. \end{aligned}$$

This completes the proof for the equality in Eq. (19), which means that $\min_{C,\pi,f_d} \mathcal{W}_\ell(f_d \# \mathbb{P}_{c,\pi}, f_y \# \mathbb{P}_x)$ is equivalent to

$$\min_{C,\pi,f_d} \min_{\bar{f}_e: \bar{f}_e \# \mathbb{P}_x = \mathbb{P}_{c,\pi}} \mathbb{E}_{x \sim \mathbb{P}_x, c \sim \bar{f}_e(x)} [\ell(f_d(c), f_y(x))], \quad (20)$$

We now further prove the above OP is equivalent to

$$\min_{C,\pi,f_d} \min_{\bar{f}_e: \bar{f}_e \# \mathbb{P}_x = \mathbb{P}_{c,\pi}} \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d(\bar{f}_e(x)), f_y(x))], \quad (21)$$

where \bar{f}_e is a **deterministic discrete** encoder mapping data example x directly to the codebooks.

It is obvious that the OP in (21) is special case of that in (20) when we limit to search for deterministic discrete encoders. Given the optimal solution $C^{*1}, \pi^{*1}, f_d^{*1}$, and \bar{f}_e^{*1} of the OP in (20), we show how to construct the optimal solution for the OP in (21). Let us construct $C^{*2} = C^{*1}$, $f_d^{*2} = f_d^{*1}$. Given $x \sim \mathbb{P}_x$, let us denote $\bar{f}_e^{*2}(x) = \operatorname{argmin}_c \ell(f_d^{*2}(c), f_y(x))$. Thus, \bar{f}_e^{*2} is a deterministic discrete encoder mapping data example x directly to the codebooks. We define $\pi_k^{*2} = \Pr(\bar{f}_e^{*2}(x) = c_k : x \sim \mathbb{P}_x)$, $k = 1, \dots, K$, meaning that $\bar{f}_e^{*2} \# \mathbb{P}_x = \mathbb{P}_{C^{*2}, \pi^{*2}}$. From the construction of \bar{f}_e^{*2} , we have

$$\mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d^{*2}(\bar{f}_e^{*2}(x)), f_y(x))] \leq \mathbb{E}_{x \sim \mathbb{P}_x, c \sim \bar{f}_e^{*1}(x)} [\ell(f_d^{*1}(c), f_y(x))].$$

Furthermore, because $C^{*2}, \pi^{*2}, f_d^{*2}$, and \bar{f}_e^{*2} are also a feasible solution of the OP in (21), we have

$$\mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d^{*2}(\bar{f}_e^{*2}(x)), f_y(x))] \geq \mathbb{E}_{x \sim \mathbb{P}_x, c \sim \bar{f}_e^{*1}(x)} [\ell(f_d^{*1}(c), f_y(x))].$$

This means that

$$\mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d^{*2}(\bar{f}_e^{*2}(x)), f_y(x))] = \mathbb{E}_{x \sim \mathbb{P}_x, c \sim \bar{f}_e^{*1}(x)} [\ell(f_d^{*1}(c), f_y(x))],$$

and $C^{*2}, \pi^{*2}, f_d^{*2}$, and \bar{f}_e^{*2} are also the optimal solution of the OP in (21). □

Lemma 6.4. (Lemma 3.2 in the main paper) Let $C^* = \{c_m^*\}_{m=1}^M, \pi^*$, and f_d^* be the optimal solution of the OP in Eq. (12), then $C^* = \{c_m^*\}_k, \pi^*$, and f_d^* are also the optimal solution of the following OP:

1. $\epsilon_M^* = \min_{f_d} \min_{\pi} \min_{\sigma \in \Sigma_\pi} \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_y(x), f_d(c_{\sigma(x)}))]$
2. $\epsilon_M^* \geq \epsilon_{M+1}^* \quad \forall M$
3. $\epsilon_M^* \rightarrow 0$ as $M \rightarrow \infty$

where Σ_π is the set of assignment functions $\sigma : \mathcal{X} \rightarrow \{1, \dots, M\}$ such that $\mathbb{P}_x(\sigma^{-1}(m)), m = 1, \dots, M$ are proportional to $\pi_k, m = 1, \dots, M$ (e.g., σ is the nearest assignment: $\sigma^{-1}(m) = \{x \mid \bar{f}_e(x) = c_m, m = \operatorname{argmin}_m d_z(f_e(x), c_m)\}$ is set of latent representations which are quantized to m^{th} codeword).

Proof. (1)

Based on the result of Proposition 6.3, the OP in Eq. (12) can be rewritten as:

$$\min_{C,\pi} \min_{f_d} \min_{\bar{f}_e: \bar{f}_e \# \mathbb{P}_x = \mathbb{P}_{C,\pi}} \mathbb{E}_{x \sim \mathbb{P}_x, c \sim \bar{f}_e(x)} [\ell(f_d(c), x)] = \min_{C,\pi} \min_{f_d} \mathcal{W}_\ell(f_d \# \mathbb{P}_{C,\pi}, f_y \# \mathbb{P}_x)$$

By definition, it is clear that

$$f_d \# \mathbb{P}_{C,\pi} = \sum_{k=1}^K \pi_k \delta_{f_d(c_k)}.$$

Therefore, the OP in right hand side can be rewritten as:

$$\min_{C,\pi} \min_{f_d} \mathcal{W}_\ell \left(f_y \# \mathbb{P}_x, \sum_{k=1}^K \pi_k \delta_{f_d(c_k)} \right). \quad (22)$$

By using the Monge definition, we have

$$\mathcal{W}_\ell \left(f_y \# \mathbb{P}_x, \sum_{k=1}^K \pi_k \delta_{f_d(c_k)} \right) = \min_{T: T \# f_y \# \mathbb{P}_x = f_d \# \mathbb{P}_{c, \pi}} \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_y(x), T(f_y(x)))]$$

Since $T \# f_y \# \mathbb{P}_x = f_d \# \mathbb{P}_{c, \pi}$, $T(x) = f_d(c_k)$ for some k . Additionally, $|T^{-1}(f_d(c_k))|, k = 1, \dots, K$ are proportional to $\pi_k, k = 1, \dots, K$. Denote $\sigma: \mathcal{X} \rightarrow \{1, \dots, K\}$ such that $T(f_y(x)) = f_d(c_{\sigma(x)}), \forall x \in \mathcal{X}$, we have $\sigma \in \Sigma_\pi$. It follows that

$$\mathcal{W}_\ell \left(f_y \# \mathbb{P}_x, \sum_{k=1}^K \pi_k \delta_{f_d(c_k)} \right) = \min_{\sigma \in \Sigma_\pi} \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_y(x), f_d(c_{\sigma(x)}))].$$

Finally, the optimal solution of the OP in Eq. (12) is equivalent to

$$\min_{f_d} \min_{\pi} \min_{\sigma \in \Sigma_\pi} \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_y(x), f_d(c_{\sigma(x)}))],$$

which directly implies the conclusion. \square

Proof. (2)-(3) Given $f_e^*, C_M^* = \{c_m^*\}_{m=1}^M, \pi^*$ and f_d^* be the optimal solution of the OP in Eq. (12) w.r.t to empirical training data \mathbb{P}_x^n , then the optimal loss is:

$$\begin{aligned} \epsilon_M^* &= \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d^*(Q_C(f_e^*(x))), f_y(x))] \\ &= \int_{\mathcal{X}} \mathbb{P}(x) \ell(f_d^*(Q_C(f_e^*(x))), f_y(x)) dx \end{aligned} \quad (23)$$

We construct a solution with a codebook size of $(M+1)$ as follows.

- First choose $x_{\max} = \arg \min_{x \sim \mathbb{P}_x} \ell(f_d(\bar{f}_e(x)), f_y(x))$
- We construct $C_{M+1} = C_M^* \cup \{c_{M+1}\} = \{c_m^*\}_{m=1}^M \cup \{c_{M+1}\}$ such that $f_d^*(c) = f_y(x_{\max})$
- Then we construct f_e such that $Q_C(f_e(x_{\max})) = c_{M+1}$ and $f_e(x) = f_e^*(x)$ for all $x \neq x_{\max}$.

Then given f_e, C_{M+1}, f_d^* , then loss w.r.t to \mathbb{P}_x is:

$$\begin{aligned} \epsilon_{M+1} &= \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d^*(Q_C(f_e(x))), f_y(x))] \\ &= \int \mathbb{P}(x) \ell(f_d^*(Q_C(f_e(x))), f_y(x)) dx \\ &= \int_{\mathbb{X} \setminus \{x_{\max}\}} \mathbb{P}(x) \ell(f_d^*(Q_C(f_e(x))), f_y(x)) dx + \mathbb{P}(x_{\max}) \ell(f_d^*(Q_C(f_e(x_{\max}))), f_y(x)) \\ &= \int_{\mathbb{X} \setminus \{x_{\max}\}} \mathbb{P}(x) \ell(f_d^*(Q_C(f_e^*(x))), f_y(x)) dx + \mathbb{P}(x_{\max}) \ell(f_d^*(Q_C(f_e(x_{\max}))), f_y(x)) \\ &= \int_{\mathbb{X} \setminus \{x_{\max}\}} \mathbb{P}(x) \ell(f_d^*(Q_C(f_e^*(x))), f_y(x)) dx \\ &\leq \int_{\mathbb{X} \setminus \{x_{\max}\}} \mathbb{P}(x) \ell(f_d^*(Q_C(f_e^*(x))), f_y(x)) dx + \mathbb{P}(x_{\max}) \ell(f_d^*(Q_C(f_e^*(x_{\max}))), f_y(x)) \\ &= \epsilon_M^* \end{aligned}$$

Since f_e, C_{M+1} , and f_d^* are feasible solutions of the OP in Eq. (12) with a codebook size of $(M+1)$, we have $\epsilon_{M+1} \geq \epsilon_{M+1}^*$. Hence, we reach $\epsilon_{M+1}^* \leq \epsilon_M^*$.

Additionally, the equation holds when $\mathbb{P}(x_{\max}) \ell(f_d^*(Q_C(f_e^*(x_{\max}))), f_y(x)) = 0$, which implies $\epsilon_M^* = 0$, directly implying the conclusion (2) and (3). \square

6.3 Proof Corollary 3.3 in the main paper: Sample Complexity

We outline the setup related to Corollary 3.3 before proceeding with the proof.

Let the training set be $S = \{x_1, \dots, x_N\}$ consisting of N sample $x_i \sim \mathbb{P}_x$ and define the empirical measure \mathbb{P}_x^N by placing mass N^{-1} at each of members of S i.e., $\mathbb{P}_x^N = \sum_{i=1}^N \frac{1}{N} \delta_{x_i}$. Given $C^* = \{c_m^*\}_{m=1}^M, \pi^*, f_e^*$, and f_d^* be the optimal solution of the OP in Eq. (12) w.r.t to empirical training data \mathbb{P}_x^N , the empirical loss w.r.t to distribution \mathbb{P}_x^N is defined as:

$$\mathcal{L}(\mathbb{P}_x^N) = \mathbb{E}_{x \sim \mathbb{P}_x^N} [\ell(f_y(x), f_d^*(Q_C(f_e^*(x))))], \quad (24)$$

Notably, Lemma 3.2 establishes that solving the optimization problem (OP) in Eq. (4) is equivalent to solving a clustering problem in the output space. However, discrete representations exhibit distinct characteristics. Specifically, defining $\bar{f}_y(x) = f_d^*(\operatorname{argmin}_{c'' \in C} \ell_y(x, f_d^*(c'')))$, we obtain:

$$\begin{aligned} \mathcal{L}(\mathbb{P}_x^N) &= \mathbb{E}_{x \sim \mathbb{P}_x^N} [\ell(f_y(x), f_d^*(\bar{f}_e(x)))] \\ &= \underbrace{\mathbb{E}_{x \sim \mathbb{P}_x^N} [\ell(\bar{f}_y(x), f_y(x))]}_{\text{Clustering Loss: } \mathcal{L}_C(\mathbb{P}_x^N)} + \underbrace{\mathbb{E}_{x \sim \mathbb{P}_x^N} [\ell(f_d^*(\bar{f}_e^*(x)), f_y(x)) - \ell(\bar{f}_y(x), f_y(x))]}_{\text{Assignment Discrepancy Loss: } \mathcal{L}_A(\mathbb{P}_x^N)} \\ &= \mathcal{L}_C(\mathbb{P}_x^N) + \mathcal{L}_A(\mathbb{P}_x^N) \end{aligned}$$

The overall error is decomposed into two components:

Clustering Loss: The first term $\ell(\bar{f}_y(x), f_y(x))$ represents the distance between the label $f_y(x)$ and its nearest centroid in the output space, which is determined by mapping the codebook to the output space.

Assignment Discrepancy Loss: It can be seen that $\bar{f}_e^*(x)$ is the nearest assignment on latent space while $\bar{f}_y(x) = f_d^*(\operatorname{argmin}_{c'' \in C} \ell_y(x, f_d^*(c'')))$ is the nearest assignment on output space. The second term i.e., the discrepancy between $\ell(f_d^*(\bar{f}_e^*(x)), f_y(x)) - \ell(\bar{f}_y(x), f_y(x))$ arises from the mismatch between assignments in the latent and output spaces. Specifically, in Figure 2, for sample , the decoded centroid $f_d(c_3)$ in the latent space may not be the closest centroid in the output space.

Following that, the general loss is defined similarly:

$$\mathcal{L}(\mathbb{P}_x) = \mathcal{L}_C(\mathbb{P}_x) + \mathcal{L}_A(\mathbb{P}_x)$$

The discrepancy between the empirical loss and the general loss can be bounded as follows:

$$\begin{aligned} |\mathcal{L}(\mathbb{P}_x^N) - \mathcal{L}(\mathbb{P}_x)| &= |\mathcal{L}_C(\mathbb{P}_x^N) + \mathcal{L}_A(\mathbb{P}_x^N) - \mathcal{L}_C(\mathbb{P}_x) - \mathcal{L}_A(\mathbb{P}_x)| \\ &\leq |\mathcal{L}_C(\mathbb{P}_x^N) - \mathcal{L}_C(\mathbb{P}_x)| + |\mathcal{L}_A(\mathbb{P}_x^N) - \mathcal{L}_A(\mathbb{P}_x)| \end{aligned}$$

The first term $|\mathcal{L}_C(\mathbb{P}_x^N) - \mathcal{L}_C(\mathbb{P}_x)|$ is the generalization bounds for k-mean problem, which is studied in (Telgarsky & Dasgupta, 2013; Bachem et al., 2017). In this study, our focus is on the second term, $|\mathcal{L}_A(\mathbb{P}_x^N) - \mathcal{L}_A(\mathbb{P}_x)|$, which examines the impact of discrete representations on the optimization problem.

We begin the proof by presenting Proposition 6.5, which is used to prove Corollary 3.3 in the main paper.

Proposition 6.5 (Concentration for Discrete Distributions (Hsu et al., 2012; Agarwal et al., 2019)). *Let z be a discrete random variable that takes values in $\{1, \dots, d\}$, distributed according to q . We write q as a vector where $q = [\mathbb{P}(z = j)]_{j=1}^d$. Assume we have n iid samples, and that our empirical estimate of q is $\hat{q} = \left[\frac{1}{N} \sum_{i=1}^N \mathbf{1}[z_j = j] \right]_{j=1}^d$. We have that $\forall \epsilon > 0$*

$$Pr \left(\|\hat{q} - q\|_2 \geq 1/\sqrt{N} + \epsilon \right) \leq e^{-N\epsilon^2}$$

which implies that

$$Pr \left(\|\hat{q} - q\|_1 \geq \sqrt{d}(1/\sqrt{N} + \epsilon) \right) \leq e^{-N\epsilon^2}$$

Corollary 6.6. Corollary 3.3 in the main paper *Given a dataset of size N , a loss function ℓ_y that is a proper metric and upper-bounded by a positive constant L , and an encoder class \mathcal{F}_e (i.e., $\bar{f}_e \in \mathcal{F}_e$), then it holds with probability at least $1 - \delta$ that*

$$|\mathcal{L}_A(\mathbb{P}_x^N) - \mathcal{L}_A(\mathbb{P}_x)| \leq \epsilon \quad (25)$$

where $\epsilon = LM \frac{\sqrt{\log 1/\delta + \log |\mathcal{F}_e| + 1}}{\sqrt{N}}$.

Proof.

$$\begin{aligned}
 |\mathcal{L}_A(\mathbb{P}_x^N) - \mathcal{L}_A(\mathbb{P}_x)| &= \left| \int_{\mathcal{X}} (\mathbb{P}(x)^N - \mathbb{P}(x)) (\ell(f_d(\bar{f}_e^*(x)), f_y(x)) - \ell(\bar{f}_y(x), f_y(x))) dx \right| \\
 &\stackrel{(1)}{\leq} \left| \int_{\mathcal{X}} (\mathbb{P}(x)^N - \mathbb{P}(x)) \ell(f_d(\bar{f}_e^*(x)), \bar{f}_y(x)) dx \right| \\
 &= \left| \int_{\mathcal{X}} (\mathbb{P}(x)^N \ell(f_d(\bar{f}_e^*(x)), \bar{f}_y(x)) dx - \int_{\mathcal{X}} \mathbb{P}(x) \ell(f_d(\bar{f}_e^*(x)), \bar{f}_y(x)) dx \right| \\
 &= |\mathcal{L}_A(\bar{f}_y, \mathbb{P}_x^N) - \mathcal{L}_A(\bar{f}_y, \mathbb{P}_x)|
 \end{aligned}$$

We have $\stackrel{(1)}{\leq}$ because:

- Since $\bar{f}_y(x) = f_d^*(\operatorname{argmin}_{c'' \in C} \ell_y(x, f_d^*(c'')))$, then $\ell(f_d^*(\bar{f}_e^*(x)), f_y(x)) - \ell(\bar{f}_y(x), f_y(x)) \geq 0$ for all x .
- by triangle inequality $\ell(f_d^*(\bar{f}_e^*(x)), \bar{f}_y(x)) \geq \ell(f_d^*(\bar{f}_e^*(x)), f_y(x)) - \ell(\bar{f}_y(x), f_y(x))$.
- Since $C^* = \{c_m^*\}_{m=1}^M, \pi^*, f_e^*$, and f_d^* be the optimal solution of the OP in Eq. (12) w.r.t to empirical training data \mathbb{P}_x^N , $\ell(f_d^*(\bar{f}_e^*(x)), f_y(x)) - \ell(\bar{f}_y(x), f_y(x)) = 0$ for all x such that $\mathbb{P}^N(x) > 0$.
- $(\mathbb{P}(x)^N - \mathbb{P}(x)) \leq 0$ for all x such that $\mathbb{P}^N(x) = 0$.

Therefore,

$$\begin{aligned}
 \mathcal{L}_A(\bar{f}_y, \mathbb{P}_x) &= \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(\bar{f}_y(x), f_d^*(\bar{f}_e^*(x)))] \\
 &= \int_{\mathcal{X}} \mathbb{P}(x) \ell(f_d(\bar{f}_e^*(x)), \bar{f}_y(x)) dx \\
 &= \int_{\mathcal{X}} \mathbb{P}(x) \sum_{c \in C} \mathbf{1}_{[\bar{f}_e^*(x)=c]} \ell(f_d^*(c), \bar{f}_y(x)) dx \\
 &= \sum_{c \in C} \int_{\mathcal{X}} \mathbb{P}(x) \mathbf{1}_{[\bar{f}_e^*(x)=c]} \ell(f_d^*(c), f_d^*(\min_{c'' \in C} \ell_y(x, f_d^*(c'')))) dx \\
 &= \sum_{c \in C} \sum_{c' \in C} \int_{\mathcal{X}} \mathbb{P}(x) \mathbf{1}_{[\bar{f}_e^*(x)=c, \operatorname{argmin}_{c'' \in C} \ell_y(x, f_d^*(c''))=c']} \ell(f_d^*(c), f_d^*(c')) dx
 \end{aligned} \tag{26}$$

Similarly, let $\mathbb{P}^N(x) = 0 \forall x \notin S$, we have:

$$\mathcal{L}_A(\bar{f}_y, \mathbb{P}_x^N) = \sum_{c \in C} \sum_{c' \in C} \int_{\mathcal{X}} \mathbb{P}^N(x) \mathbf{1}_{[\bar{f}_e^*(x)=c, \operatorname{argmin}_{c'' \in C} \ell_y(x, f_d^*(c''))=c']} \ell(f_d^*(c), f_d^*(c')) dx$$

The discrepancy between the empirical loss and the general loss, arising from the mismatched assignment between the output space and the latent space, can be bounded as follows:

$$\begin{aligned}
 & |\mathcal{L}(\bar{f}_y, \mathbb{P}_x^N) - \mathcal{L}(\bar{f}_y, \mathbb{P}_x)| \tag{27} \\
 &= \left| \sum_{c \in C} \sum_{c' \in C} \int_{\mathcal{X}} \left(\mathbb{P}^N(x) \mathbf{1}_{[\bar{f}_e^*(x)=c, \arg\min_{c'' \in C} \ell_y(x, f_d^*(c''))=c']} - \mathbb{P}(x) \mathbf{1}_{[\bar{f}_e^*(x)=c, \arg\min_{c'' \in C} \ell_y(x, f_d^*(c''))=c']} \right) \ell(f_d^*(c), f_d^*(c')) dx \right| \\
 &= \left| \sum_{c \in C} \sum_{c' \in C} \ell(f_d^*(c), f_d^*(c')) \int_{\mathcal{X}} \left(\mathbb{P}^N(x) \mathbf{1}_{[\bar{f}_e^*(x)=c, \arg\min_{c'' \in C} \ell_y(x, f_d^*(c''))=c']} - \mathbb{P}(x) \mathbf{1}_{[\bar{f}_e^*(x)=c, \arg\min_{c'' \in C} \ell_y(x, f_d^*(c''))=c']} \right) dx \right| \\
 &= \sum_{c \in C} \sum_{c' \in C} |\ell(f_d^*(c), f_d^*(c'))| \left| \int_{\mathcal{X}} \mathbb{P}^N(x) \mathbf{1}_{[\bar{f}_e^*(x)=c, \arg\min_{c'' \in C} \ell_y(x, f_d^*(c''))=c']} - \mathbb{P}(x) \mathbf{1}_{[\bar{f}_e^*(x)=c, \arg\min_{c'' \in C} \ell_y(x, f_d^*(c''))=c']} \right| dx \\
 &= L \sum_{c \in C} \sum_{c' \in C} \left| \int_{\mathcal{X}} \mathbb{P}^N(x) \mathbf{1}_{[\bar{f}_e^*(x)=c, \arg\min_{c'' \in C} \ell_y(x, f_d^*(c''))=c']} - \mathbb{P}(x) \mathbf{1}_{[\bar{f}_e^*(x)=c, \arg\min_{c'' \in C} \ell_y(x, f_d^*(c''))=c']} \right| dx \\
 &= L \|\hat{P}_C - P_C\|_1 \tag{28}
 \end{aligned}$$

where \hat{P}_C and P_C are the estimated and the true probability distribution vector of the product codebook $C \times C' \in \mathcal{C} \times \mathcal{C}$, i.e. $\hat{P}_C = [\frac{1}{N} \sum_{i=1}^N \mathbf{1}_{[\bar{f}_e^*(x)=c, \arg\min_{c'' \in C} \ell_y(x, f_d^*(c''))=c']}]_{c, c' \in C}$ and $P_C = [\mathbb{P}(\bar{f}_e^*(x) = c, \arg\min_{c'' \in C} \ell_y(x, f_d^*(c'')) = c')]_{c, c' \in C}$.

Applying the concentration inequality from Proposition 6.5, we have:

$$Pr \left(\|\hat{P}_C - P_C\|_1 \geq M(1/\sqrt{N} + \epsilon) \right) \leq e^{-N\epsilon^2},$$

Next, we apply the union bound to the above result across all optimal functions in the encoder class \mathcal{F}_e , which gives:

$$Pr \left(\exists \bar{f}_e \in \mathcal{F}_e, \|\hat{P}_C - P_C\|_1 \geq M(1/\sqrt{N} + \epsilon) \right) \leq |\mathcal{F}_e| e^{-N\epsilon^2},$$

Let $\delta = |\mathcal{F}_e| e^{-N\epsilon^2}$, then with probability of at least $1 - \delta$

$$Pr \left(\|\hat{P}_C - P_C\|_1 \leq M(1/\sqrt{N} + \epsilon) \right) \geq 1 - \delta, \quad \forall \bar{f}_e \in \mathcal{F}_e \tag{29}$$

Finally, combining this with (29), we have with probability at least $1 - \delta$:

$$|\mathcal{L}_A(\bar{f}_y, \mathbb{P}_x^n) - \mathcal{L}_A(\bar{f}_y, \mathbb{P}_x)| \leq LM \frac{\sqrt{\log 1/\delta + \log |\mathcal{F}_e|} + 1}{\sqrt{N}}$$

□

6.4 Equivalence between K -class classification and Discrete Representation Framework

We consider the classifier on top of the encoder as a linear layer represented by the weight matrix $f_d = W = [w_k]_{k=1}^K$, where w_k is the vector corresponding to class “k”. The prediction via softmax activation is given by:

$$P(y | x) = \frac{\exp(f_e(x)^\top W_k)}{\sum_{j=1}^K \exp(f_e(x)^\top W_j)}$$

Given that the loss of interest is the *cross-entropy loss*, which is commonly used for classification tasks, the functions f_e and f_d are learned through the following optimization problem (OP):

$$\begin{aligned}
 & \min_{W, f_e} \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d(f_e(x)), f_y(x))] \\
 &= \min_{W, f_e} \mathbb{E}_{x \sim \mathbb{P}_x} \left[f_y(x) \log \left(\frac{\exp(f_e(x)^\top W_{f_y(x)})}{\sum_{j=1}^K \exp(f_e(x)^\top W_j)} \right) \right] \tag{30}
 \end{aligned}$$

This loss reaches its optimal solution as $f_e(x) = \lambda W_{f_y(x)}$ with λ as a coefficient, and $f_e(x)^\top W_j = 0$ for all $j \neq f_y(x)$. Since the optimal loss remains unchanged for any value of λ , we can, without loss of generality, set $\lambda = 1$.

The above optimal solution for f_e may be too restrictive. In fact, given W , we only need to find f_e such that $W_{f_y(x)} = Q_W(f_e(x)) = \arg \max_{w \in W} d_z(f_e(x), w)$.

Given that f_e^* and W^* are the optimal solutions of the optimization problem in Eq. (30), then $\bar{f}_e^* \# \mathbb{P}_x = \mathbb{P}_{c,\pi}$, where $\bar{f}_e^* = Q_W \circ f_e^*$, $C = W^* = \{w_k\}_{k=1}^K$, $\pi = f_y \# \mathbb{P}_x$, $\mathbb{P}_{c,\pi} = \sum_{k=1}^K \pi_k \delta_{c_k}$ and $f_d^* = W^*$ also the optimal solution of OP in Eq. (12):

$$\min_{\mathbb{P}_{c,\pi}, f_d} \min_{\bar{f}_e: \bar{f}_e \# \mathbb{P}_x = \mathbb{P}_{c,\pi}} \mathbb{E}_{x \sim \mathbb{P}_x} [\ell(f_d(\bar{f}_e(x)), f_y(x))]$$

7 Experimental Settings

7.1 Domain Generalization

We adopt the training and evaluation protocol as described in the DomainBed benchmark (Gulrajani & Lopez-Paz, 2021), including dataset splits, model selection on the validation set, and optimizer settings. To improve computational efficiency, as recommended by (Cha et al., 2021), we restrict our HP search space. Particularly, we utilize the Adam optimizer with a learning rate of 5×10^{-5} , as detailed in (Gulrajani & Lopez-Paz, 2021), without any adjustments for dropout or weight decay. The batch size is set to 32.

Our method’s unique hyperparameter, the regularization term λ , is optimized within the range of $[0.01, 0.1, 1.0]$, and the number of discrete representations $|\mathcal{M}|$ is fixed at 16 times the number of classes.

SWAD-specific hyperparameters remain at their default settings, and the evaluation frequency is set to 300 for all datasets.

7.2 Datasets

To evaluate the effectiveness of the proposed method, we utilize five datasets: PACS (Li et al., 2017), VLCS (Torr-alba & Efros, 2011), Office-Home (Venkateswara et al., 2017), Terra Incognita (Beery et al., 2018) and Domain-Net (Peng et al., 2019) which are the common DG benchmarks with multi-source domains.

- **PACS** (Li et al., 2017): 9991 images of seven classes in total, over four domains: Art_painting (A), Cartoon (C), Sketches (S), and Photo (P).
- **VLCS** (Torr-alba & Efros, 2011): five classes over four domains with a total of 10729 samples. The domains are defined by four image origins, i.e., images were taken from the PASCAL VOC 2007 (V), LabelMe (L), Caltech (C) and Sun (S) datasets.
- **Office-Home** (Venkateswara et al., 2017): 65 categories of 15,500 daily objects from 4 domains: Art, Clipart, Product (vendor website with white-background) and Real-World (real-object collected from regular cameras).

7.3 State abstraction Reinforcement Learning

We consider the Visual Gridworld environment from (Allen et al., 2021). The underlying grid states is a 10×10 grid structure, where the agent occupies individual squares and can navigate in four possible adjacent squares. The agent receives the image observation from the environment of size 30×30 pixels, with its current location highlighted. The images are interfered with truncated Gaussian noise, making it crucial for an effective encoder to distinguish between the noise and the agent’s true position for downstream tasks. See Figure (3) for an illustration.

The learned encoder serves as a pre-trained feature extractor from image observation for training a DQN agent for downstream tasks. A random goal position is designated through out this training phase, and the agent receives a reward of (-1) for each step taken until it reaches the goal, upon which the episode stops and the agent is reset to another random non-goal position on the grid.

All agents are trained on pre-collected data samples of random-walk interactions and random resets in the environment. No reward signal is used to assist the pretraining stage. The encoder network architecture of ϕ is taken from the original paper, which is a 2-layer fully connected neural network with 32 hidden dimensions using the Tanh activation function. For quantization methods, the outputs of the encoder are quantized into a codeword in a codebook. The training Loss of Markov abstraction is the weighted sum of the following three loss function

$$\mathcal{L}_{\text{Markov}} = \alpha \mathcal{L}_{\text{Inv}} + \beta \mathcal{L}_{\text{Ratio}} + \eta \mathcal{L}_{\text{Smooth}}$$

We follow the original configuration to set the coefficients of the Inverse and Ratio loss to 1 and the smooth coefficient to 0 in this task. In our experiment, the Continuous baseline is the original version of Markov abstraction, using similar setups as in (Allen et al., 2021). For our quantization method, there is an additional Wasserstein loss from the objective in (13), we incorporate this quantization loss into the Markov loss as

$$\mathcal{L} = \mathcal{L}_{\text{Markov}} + \lambda \mathcal{L}_{\text{quantize}}$$

we set $\lambda = 100$ for all experiments with reinforcement learning through gridsearch the value of $\lambda \in \{0.1, 1, 10, 100\}$. All experiments are repeated with 5 different seeds.