RETHINKING THE UNIFORMITY METRIC IN SELF-SUPERVISED LEARNING

Xianghong Fang
The Chinese University of Hong Kong, Shenzhen
fangxianghong2@gmail.com

Jian Li
Tencent AI Lab
lijianjack@gmail.com

Qiang Sun *
University of Toronto & MBZUAI
qsunstats@gmail.com

Benyou Wang *
The Chinese University of Hong Kong, Shenzhen & SRIBD
wangbenyou@cuhk.edu.cn

ABSTRACT

Uniformity plays an important role in evaluating learned representations, providing insights into self-supervised learning. In our quest for effective uniformity metrics, we pinpoint four fundamental properties that such metrics should possess. Specifically, an effective uniformity metric should remain invariant to instance permutations and sample replications while being able to capture feature redundancy and dimensional collapse. Surprisingly, we find that the uniformity metric proposed by Wang & Isola (2020) fails to satisfy three of these properties, thus revealing its limitations. Particularly, their metric proves insensitive to dimensional collapse. To address these shortcomings, we introduce a new uniformity metric based on the Wasserstein distance, which fulfills all the aforementioned properties. By directly optimizing this new metric alongside alignment, we effectively mitigate dimensional collapse and consistently improve the performance of various self-supervised learning methods on downstream tasks involving CIFAR-10 and CIFAR-100 datasets. Code is available at https://github.com/statsle/Wasserstein-SSL.

1 INTRODUCTION

Self-supervised learning excels in acquiring invariant representations to various augmentations (Chen et al., 2020; He et al., 2020; Caron et al., 2020; Grill et al., 2020; Zbontar et al., 2021). It has been outstandingly successful across a wide range of domains, such as multimodality learning, object detection, and segmentation (Radford et al., 2021; Li et al., 2022; Xie et al., 2021; Wang et al., 2021; Yang et al., 2021; Zhao et al., 2021). To gain a deeper understanding of self-supervised learning, thoroughly evaluating the learned representations is a pragmatic approach (Wang & Isola, 2020; Gao et al., 2021; Tian et al., 2021; Jing et al., 2022).

Alignment, a metric quantifying the similarities between positive pairs, holds significant importance in the evaluation of learned representations (Wang & Isola, 2020). It ensures that samples forming positive pairs are mapped to nearby features, thereby rendering them invariant to irrelevant noise factors (Hadsell et al., 2006; Chen et al., 2020). However, relying solely on alignment proves inadequate for effectively assessing representation quality in self-supervised learning. This limitation becomes evident in the presence of extremely small alignment values in collapsing solutions, as observed in Siamese networks (Hadsell et al., 2006), where all outputs collapse to a single point (Chen & He, 2021), as illustrated in Figure 1. In such cases, the learned representations exhibit optimal alignment but fail

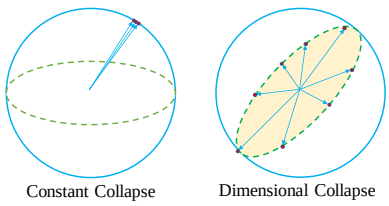


Figure 1: The left figure presents constant collapse, and the right figure visualizes dimensional collapse.

*Qiang Sun and Benyou Wang are joint corresponding authors.

to provide meaningful information for any downstream tasks. This underscores the necessity of incorporating additional factors when evaluating learned representations.

To better evaluate the learned representations, Wang & Isola (2020) formally introduced a *uniformity* metric by utilizing the logarithm of the average pairwise Gaussian potential (Cohn & Kumar, 2007). Uniformity assesses how feature embeddings are distributed uniformly across the unit hypersphere, and higher uniformity indicates that more information is preserved by the learned representations. Since its invention, uniformity has played a pivotal role in comprehending self-supervised learning and mitigating the issue of constant collapse (Arora et al., 2019; Wang & Isola, 2020; Gao et al., 2021). Nevertheless, the validity of this particular uniformity metric warrants further examination and scrutiny.

To examine the existing uniformity metric (Wang & Isola, 2020), we introduce four principled properties, also known as desiderata, that a desired uniformity metric should fulfill. Guided by these properties, we conduct a theoretical analysis that reveals certain shortcomings of the existing metric, particularly its insensitivity to dimensional collapse (Hua et al., 2021)[1]. We complement our theoretical findings with empirical evidence that underscores the metric's limitations in addressing dimensional collapse. We then introduce a new uniformity metric that satisfies all desiderata. In particular, the proposed metric is sensitive to dimensional collapse and thus is superior to the existing one. Finally, using the proposed uniformity metric as an auxiliary loss within existing self-supervised learning methods consistently improves their performance in downstream tasks.

Our main contributions are summarized as follows. (i) We introduce four desiderata that provide a novel perspective on the design of ideal uniformity metrics. Notably, the existing uniformity metric (Wang & Isola, 2020) does not satisfy all of these desiderata. Specifically, we demonstrate, both theoretically and empirically, its insensitivity to dimensional collapse. (ii) We propose a novel uniformity metric that not only fulfills all of the desiderata but also exhibits sensitivity to dimensional collapse, addressing a crucial limitation of the existing metric. (iii) Our newly proposed uniformity metric can be seamlessly incorporated as an auxiliary loss in a variety of self-supervised methods, consistently improving their performance in downstream tasks.

2 Background

2.1 Self-Supervised Representation Learning

Self-supervised learning leverages the idea that similar samples should have similar representations, while remaining invariant to unnecessary details (Wang & Isola, 2020). For instance, the Siamese network (Hadsell et al., 2006) takes as input positive pairs $(\mathbf{x}^a, \mathbf{x}^b)$, often obtained by taking two augmented views of the same sample $\mathbf{x}$. These positive pairs are then processed by an encoder network f consisting of a backbone (e.g., ResNet (He et al., 2016)) and a projection MLP head (Chen et al., 2020), leading to positive pairs of representations $(\mathbf{z}^a = f(\mathbf{x}^a), \mathbf{z}^b = f(\mathbf{x}^b)$[2]. To enforce invariance, a natural solution is to minimize the following alignment loss, defined as the expected distance between positive pairs:

$$\mathcal{L}_{\mathcal{A}} := \mathbb{E}_{(\mathbf{z}^a, \mathbf{z}^b) \sim p_{\text{pos}}} \left\| \mathbf{z}_i^a - \mathbf{z}_i^b \right\|_2^2, \tag{1}$$

where $p_{\text{pos}}(\cdot, \cdot)$ is the distribution of positive pairs.

However, optimizing the above alignment loss alone may lead to an undesired collapsing solution, where all representations collapse into a single point, as shown in Figure 1.

2.2 Existing Solutions to Constant Collapse

To prevent constant collapse, existing solutions include contrastive learning, asymmetric model architecture, and redundancy reduction.

[1]When dimensional collapse occurs, representations occupy a lower-dimensional subspace instead of the entire embedding space (Jing et al., 2022) and thus some dimensions are not fully utilized; see Figure 1.

[2]We also refer to $(\mathbf{z}^a, \mathbf{z}^b)$ as positive pairs.

Contrastive Learning Contrastive learning serves as a valuable technique for mitigating constant collapse. The key idea is to utilize negative pairs. For example, SimCLR (Chen et al., 2020) introduced an in-batch negative sampling strategy that utilizes samples within a batch as negative samples. However, its effectiveness is contingent on the use of a large batch size. To address this limitation, MoCo (He et al., 2020) used a memory bank, which stores additional representations as negative samples. Recent research endeavors have also explored clustering-based contrastive learning, which combines a clustering objective with contrastive learning techniques (Li et al., 2021; Caron et al., 2020).

Asymmetric Model Architecture The use of asymmetric model architecture represents another approach to combat constant collapse. One plausible explanation for its effectiveness is that such an asymmetric design encourages encoding more information (Grill et al., 2020). To maintain this asymmetry, BYOL (Grill et al., 2020) introduces the concept of using an additional predictor in one branch of the Siamese network while employing momentum updates and stop-gradient operators in the other branch. DINO (Caron et al., 2021), takes this asymmetry a step further by applying it to two encoders, distilling knowledge from the momentum encoder into the other one (Hinton et al., 2015). SimSiam (Chen & He, 2021) removes the momentum update from BYOL, and illustrates that the momentum update may not be essential in preventing collapse. In contrast, Mirror-SimSiam (Zhang et al., 2022a) swaps the stop-gradient operator to the other branch. Its failure challenges the assertion made in SimSiam (Chen & He, 2021) that the stop-gradient operator is the key component for preventing constant collapse. Theoretically, Tian et al. (2021) provides an examination to elucidate why an asymmetric model architecture can effectively avoid constant collapse.

Redundancy Reduction The fundamental principle behind redundancy reduction to mitigate constant collapse is to maximize the information preserved by the representations. The key idea is to decorrelate the learned representations. Barlow Twins (Zbontar et al., 2021) aims to achieve decorrelation by focusing on the cross-correlation matrix, while VICReg (Bardes et al., 2022) focuses on the covariance matrix. Zero-CL (Zhang et al., 2022b) takes a hybrid approach, combining instance-wise and feature-wise whitening techniques to achieve decorrelation.

2.3 Collapse Analysis

While the aforementioned solutions effectively prevent constant collapse, they are not as effective in preventing dimensional collapse, wherein representations occupy a lower-dimensional subspace instead of the entire space. This phenomenon has been observed in contrastive learning by visualizing the singular value spectra of representations (Jing et al., 2022; Tian et al., 2021).

To quantitatively measure the degree of collapse, Wang & Isola (2020) introduced a uniformity loss based on the logarithm of the average pairwise Gaussian potential. Given (normalized) feature representations $\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n\}$, their proposed empirical uniformity loss is:

$$\mathcal{L}_{\mathcal{U}} := \log \frac{1}{n(n-1)/2} \sum_{i=2}^{n} \sum_{j=1}^{i-1} e^{-t\|\mathbf{z}_i - \mathbf{z}_j\|_2^2}, \tag{2}$$

where $t > 0$ is a fixed parameter, often set to 2. Then $-\mathcal{LU}$ serves as the corresponding uniformity metric, with a higher value indicating greater uniformity.

We demonstrate in this work that this metric is insensitive to dimensional collapse, both theoretically in Section 3.2 and empirically in Section 5.2.

3 What makes a good uniformity metric?

In this section, we begin by presenting four fundamental properties that an effective uniformity metric should posses. Leveraging these properties as a lens, we then scrutinize the existing uniformity metric $-\mathcal{L}_{\mathcal{U}}$ proposed by Wang & Isola (2020), shedding light on its limitations.

3.1 FOUR PROPERTIES FOR UNIFORMITY

A uniformity metric $\mathcal{U} : \mathbb{R}^{mn} \to \mathbb{R}$ is a function that maps a set of learned representations to a scalar indicator of uniformity. In the following section, we introduce four properties necessary for an effective uniformity metric. Let $\mathcal{D} = \mathbf{z}_1, \ldots, \mathbf{z}_n \in \mathbb{R}^{mn}$ represent the learned representations. To avoid the trivial case, we assume that $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are not all equal, meaning that not all points collapse to a single constant point.

First, the uniformity metric should be invariant to the permutation of instances, as the distribution of representations should not be affected by permutations.

Property 1 (Instance Permutation Constraint (IPC))**.** *An effective uniformity metric $\mathcal{U}$ should satisfy*

$$\mathcal{U}(\pi(\mathcal{D})) = \mathcal{U}(\mathcal{D}), \tag{3}$$

where π is a permutation over the the instances.

Second, an effective uniformity metric should be invariant to instance clones, as cloning does not vary the distribution of embeddings.

Property 2 (Instance Cloning Constraint (ICC))**.** *An effective uniformity metric $\mathcal{U}$ should satisfy*

$$\mathcal{U}(\mathcal{D} \uplus \mathcal{D}) = \mathcal{U}(\mathcal{D}), \tag{4}$$

where $\mathcal{D} \uplus \mathcal{D} := \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n, \mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n\}$.

Third, an effective uniformity metric should decrease as feature-level cloning for each instance occurs, as this duplication introduces redundancy, which leads to dimensional collapse (Zbontar et al., 2021; Bardes et al., 2022).

Property 3 (Feature Cloning Constraint (FCC))**.** *An effective uniformity metric $\mathcal{U}$ should satisfy*

$$\mathcal{U}(\mathcal{D} \oplus \mathcal{D}) < \mathcal{U}(\mathcal{D}), \tag{5}$$

where $\mathcal{D} \oplus \mathcal{D} := \{\mathbf{z}_1 \oplus \mathbf{z}_1, \mathbf{z}_2 \oplus \mathbf{z}_2, ..., \mathbf{z}_n \oplus \mathbf{z}_n\}$ and $\mathbf{z}_i \oplus \mathbf{z}_i := (z_{i1}, \cdots, z_{im}, z_{i1}, \cdots, z_{im})^{\mathrm{T}} \in \mathbb{R}^{2m}$.

An effective uniformity metric should decrease with the addition of constant features for each instance, as this introduces uninformative features and leads to collapsed dimensions.

Property 4 (Feature Baby Constraint (FBC))**.** *An effective uniformity metric $\mathcal{U}$ should satisfy*

$$\mathcal{U}(\mathcal{D} \oplus \mathbf{0}^k) < \mathcal{U}(\mathcal{D}), \quad k \in \mathbb{N}^+, \tag{6}$$

where $\oplus$ is defined in Property 3, that is, $\mathcal{D} \oplus \mathbf{0}^k = \{\mathbf{z}_1 \oplus \mathbf{0}^k, \mathbf{z}_2 \oplus \mathbf{0}^k, ..., \mathbf{z}_n \oplus \mathbf{0}^k\}$ and $\mathbf{z}_i \oplus \mathbf{0}^k = (z_{i1}, z_{i2}, ..., z_{im}, 0, 0, ..., 0)^{\mathrm{T}} \in \mathbb{R}^{m+k}$.

Intuitively, Property 2 ensures that the uniformity metric remains insensitive to sample replication, while Properties 3 and 4 make the metric sensitive to feature redundancy and dimensional collapse, respectively. These four properties constitute intuitive yet fundamental characteristics of an effective uniformity metric.

3.2 EXAMINING THE UNIFORMITY METRIC $-\mathcal{L}_{\mathcal{U}}$

We employ the four properties introduced earlier to analyze the uniformity metric $-\mathcal{L}_{\mathcal{U}}$ defined in Eqn. (2). The following theorem summarizes our findings.

Theorem 1. *The uniformity metric $-\mathcal{L}_{\mathcal{U}}$ satisfies Property 1, but violates Properties 2, 3, and 4.*

The proof of the above theorem is collected in Appendix C.1. The violation of Property 2 indicates that the uniformity metric $-\mathcal{L}_{\mathcal{U}}$ is sensitive to sample replication, while the violations of Properties 2 and 3 suggest that $-\mathcal{L}_{\mathcal{U}}$ is insensitive to feature redundancy and dimensional collapse. Therefore, there is a pressing need to develop a new uniformity metric.

4 A NEW UNIFORMITY METRIC

In this section, we introduce a new uniformity metric to address the limitations of $-\mathcal{L}_{\mathcal{U}}$.

4.1 THE UNIFORM SPHERICAL DISTRIBUTION AND AN APPROXIMATION

As posited by (Wang & Isola, 2020), feature vectors should be approximately uniformly distributed on the unit hypersphere $\mathcal{S}^{m-1}$, thereby retaining as much information from the data as possible. Therefore, we adopt the uniform spherical distribution as our target distribution with maximal uniformity.

Our idea is to use some statistical distance between the feature distribution and the target distribution as the new uniformity loss. However, computing statistical distances involving the uniform spherical distribution can be challenging. To overcome this, we rely on the following fact, aka Fact 1, which establishes an equivalence between the uniform spherical distribution and the normalized isotropic Gaussian distribution.

Fact 1. *If $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_m)$, then $\mathbf{Y} := \mathbf{Z}/\|\mathbf{Z}\|_2$ is uniformly distributed on the unit hypersphere $\mathcal{S}^{m-1}$.*

Because the average length of $\|\mathbf{Z}\|_2$ is roughly $\sigma\sqrt{m}$ (Chandrasekaran et al., 2012), that is,

$$\frac{m}{\sqrt{m+1}} \leq \|\mathbf{Z}\|_2/\sigma \leq \sqrt{m},$$

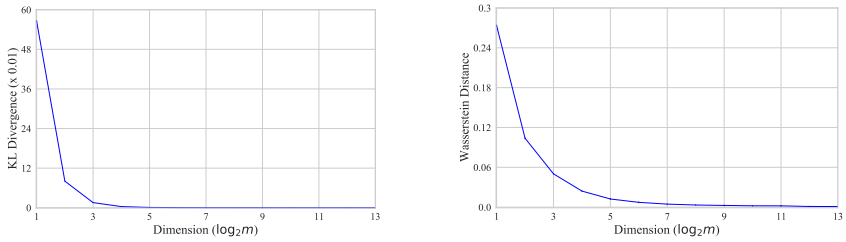
we expect $\mathbf{Z}/(\sigma\sqrt{m}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m/m)$ provides a reasonable approximation to $\mathbf{Z}/\|\mathbf{Z}\|_2$, and thus to the uniform spherical distribution. This is rigorously justified in the following theorem.

Theorem 2. *Let Y_i be the i-th coordinate of $\mathbf{Y} = \mathbf{Z}/\|\mathbf{Z}\|_2 \in \mathbb{R}^m$, where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_m)$. Then the quadratic Wasserstein distance between Y_i and $\widehat{Y}_i \sim \mathcal{N}(0, 1/m)$ converges to zero as $m \to \infty$, that is,*

$$\lim_{m\to\infty} \mathcal{W}_2(Y_i, \widehat{Y}_i) = 0.$$

The theorem above can be proved by directly utilizing the probability density functions of Y_i and $\widehat{Y}_i$. The detailed proof is deferred to Appendix B. Theorem 2 shows $\mathcal{N}(\mathbf{0}, m^{-1}\mathbf{I}_m)$ approximates the distribution of each coordinate of the uniform spherical distribution as $m \to \infty$.

We empirically compare the distributions of Y_i and $\widehat{Y}_i$ with various dimensions $m \in \{2, 4, 8, 16, 32, 64, 128, 256\}$. For each m, we sample 200,000 data points from both Y_i and $\widehat{Y}_i$, bin them into 51 groups, and calculate the empirical KL divergence and Wasserstein distance. Figure 2 plots both distances versus increasing dimensions. We observe that both distances converge to zero when m increases. Specifically, these results demonstrate that the distribution of $\widehat{Y}_i$ is a reasonable approximation to that of Y_i when $m \geq 2^4 = 16$. More comparisons between $\mathbf{Y}$ and $\widehat{\mathbf{Y}}$ can be found in Appendix D.



(a) KL Divergence (b) Wasserstein Distance

Figure 2: The KL divergence and Wasserstein distance between Y_i and $\widehat{Y}_i$ w.r.t. various dimensions.

4.2 A NEW METRIC FOR UNIFORMITY

In this section, we propose to use a statistical distance between the distribution of learned representations and $\mathcal{N}(\mathbf{0}, \mathbf{I}_m/m)$, in place of the uniform spherical distribution Unif($\mathcal{S}^{m-1}$), as our new uniformity loss.

To facilitate the computation, we adopt a Gaussian hypothesis for the learned representations and assume they follow $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. With this assumption, we use the quadratic Wasserstein distance[3]

[3]We discuss using other statistical distances as uniformity losses, such as the Kullback-Leibler divergence and Bhattacharyya distance, in Appendix E.

to calculate the distance between two distributions; see the definition in Appendix E. We need the following well-known lemma (Olkin & Pukelsheim, 1982).

Lemma 1. *Then the quadratic Wasserstein distance between $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $N(\mathbf{0}, \mathbf{I}/m)$ is*

$$\sqrt{\|\boldsymbol{\mu}\|_2^2 + 1 + \operatorname{tr}(\boldsymbol{\Sigma}) - \frac{2}{\sqrt{m}} \operatorname{tr}(\boldsymbol{\Sigma}^{\frac{1}{2}})}. \tag{7}$$

The lemma above indicates that the quadratic Wasserstein distance can be easily computed using the mean and covariance of the the representations. In practice, we estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ by using the sample mean $\widehat{\boldsymbol{\mu}}$ and the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$, respectively. Specifically, we use the following empirical quadratic Wasserstein distance as our new uniformity loss:

$$\mathcal{W}_2 := \sqrt{\|\widehat{\boldsymbol{\mu}}\|_2^2 + 1 + \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}) - \frac{2}{\sqrt{m}} \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}^{\frac{1}{2}})}. \tag{8}$$

Then $-\mathcal{W}_2$ can be used as the new uniformity metric: Smaller $\mathcal{W}_2$ indicates larger uniformity of learned representations. Additionally, our new uniformity loss can be used for improving various existing self-supervised methods.

5 Comparing Two Metrics

5.1 Theoretical Comparison

We examine the proposed metric $-\mathcal{W}_2$ in terms of the four properties introduced earlier. The following theorem summarize our findings.

Theorem 3. *The uniformity metric $-\mathcal{W}_2$ satisfies all four properties, that is, Properties 1–4.*

The proof of the above theorem resembles that of Theorem 1, and is collected in Appendix C.2. Table 1 compares $-\mathcal{L}_{\mathcal{U}}$ and $\mathcal{W}2$. We emphasize again that our new uniformity metric is invariant to instance permutations and sample replications, while effectively capturing feature redundancy and dimensional collapse.

Taking dimensional collapse as an example, consider $\mathcal{D} \oplus \mathbf{0}^k$ versus $\mathcal{D}$. A larger k indicates a more severe dimensional collapse. However, $-\mathcal{L}_{\mathcal{U}}$ fails to identify this issue, as $-\mathcal{L}_{\mathcal{U}}(\mathcal{D} \oplus \mathbf{0}^k) = -\mathcal{L}_{\mathcal{U}}(\mathcal{D})$. In sharp contrast, our proposed metric can accurately detect this dimensional collapse, as $-\mathcal{W}_2(\mathcal{D} \oplus \mathbf{0}^k) < -\mathcal{W}_2(\mathcal{D})$.

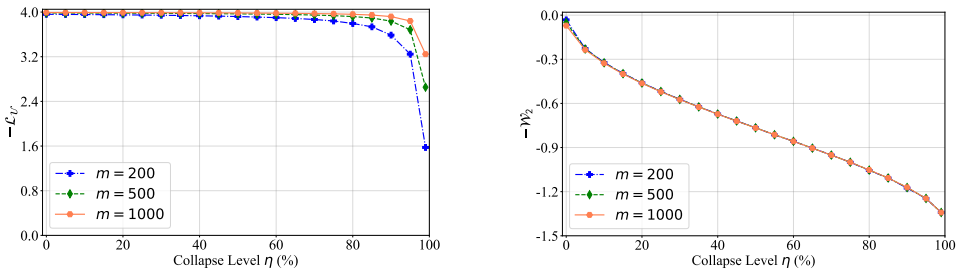
Table 1: Verifying the four properties for $-\mathcal{L}_{\mathcal{U}}$ and $-\mathcal{W}_2$.

Properties	IPC	ICC	FCC	FBC
$-\mathcal{L}_{\mathcal{U}}$	✔	✗	✗	✗
$-\mathcal{W}_2$	✔	✔	✔	✔

5.2 Empirical Comparison via Synthetic Data

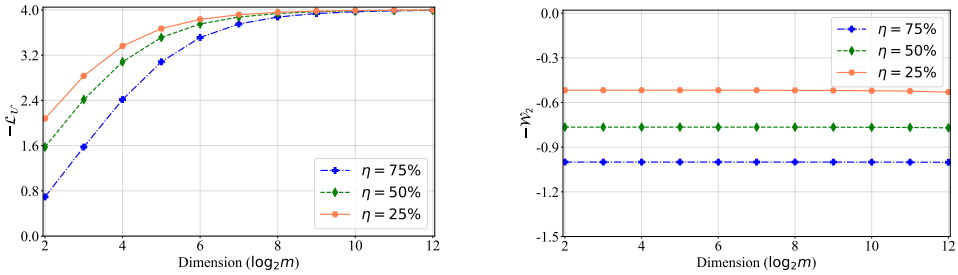
We further conduct synthetic experiments to investigate two uniformity metrics. An empirical study on the correlation between these metrics reveals that data points following a standard Gaussian distribution achieve better uniformity compared to those from other distributions; see Appendix F.1 for details. Furthermore, we generate data vectors from this distribution to facilitate a comprehensive comparison between the two metrics.

On Dimensional Collapse Degrees To synthesize data exhibiting varying degrees of dimensional collapse, we sample data vectors from the standard Gaussian distribution and zero out a proportion of the coordinates. This approach illustrates that as the proportion of zero-value coordinates increases, dimensional collapse becomes more pronounced. The proportion of zero-value coordinates is η while that of non-zero coordinates is $1 - \eta$. As shown in Figure 3(a) and Figure 3(b), $-\mathcal{W}_2$ is capable of capturing different collapse degrees, while $-\mathcal{L}_{\mathcal{U}}$ stays the same even with 80% collapse ($\eta = 80\%$), indicating that $-\mathcal{L}_{\mathcal{U}}$ is insensitive to the degrees of dimensional collapse.



(a) Sensitivity to collapse degrees for $-\mathcal{L}_{\mathcal{U}}$ (b) Sensitivity to collapse degrees for $-\mathcal{W}_2$

Figure 3: Sensitivity to dimensional collapse degrees: $-\mathcal{W}_2$ is more sensitive than $-\mathcal{L}_{\mathcal{U}}$.



(a) Effectiveness when increasing m for $-\mathcal{L}_{\mathcal{U}}$ (b) Effectiveness when increasing m for $-\mathcal{W}_2$

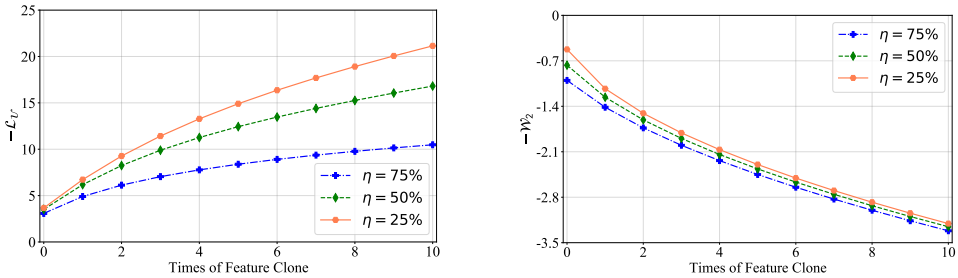
Figure 4: Effectiveness when increasing dimensions m: $-\mathcal{L}_{\mathcal{U}}$ fails to distinguish different dimensional collapse degrees for large dimensions, while $-\mathcal{W}_2$ is always able to.

On Sensitiveness of Dimensions Moreover, Figure 4 shows that $-\mathcal{L}_{\mathcal{U}}$ can not distinguish different degrees of dimensional collapse ($\eta = 25\%, 50\%$, and 75%) when the dimension m becomes large (e.g., $m \geq 2^8 = 256$). In sharp contrast, $-\mathcal{W}_2$ only depends on the degree of dimensional collapse and is independent of the dimensions m.

To complement the theoretical comparisons between the two metrics discussed in Section 5.1, we also conduct empirical comparisons in terms of FCC and FBC. ICC comparisons are collected in Appendix F.2.

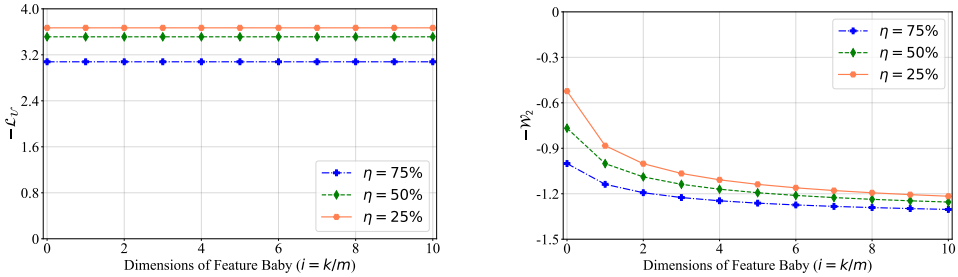
On Feature Cloning Constraint (FCC) We further investigate the impact of feature cloning by creating multiple feature clones of the dataset, e.g., $\mathcal{D} \oplus \mathcal{D}$ and $\mathcal{D} \oplus \mathcal{D} \oplus \mathcal{D}$, corresponding to one and two times cloning, respectively. Figure 5(a) demonstrates that the value of $-\mathcal{L}_{\mathcal{U}}$ remains constant as the number of clones increases, which violates the strict inequality in Eqn. (5). In contrast, in Figure 5(b), our proposed metric $-\mathcal{W}_2$ decreases, satisfying the property.

On Feature Baby Constraint (FBC) We finally analyze the effect of feature baby, where we insert k dimensional zero vectors into each instance of $\mathcal{D}$. This modified dataset is denoted as $\mathcal{D} \oplus \mathbf{0}^k$, and we examine the impact of k in both metrics. Figure 6(a) shows that the value of $-\mathcal{L}_{\mathcal{U}}$ rises as k increases, violating the strict inequality constraint in Eqn. (6). In contrast, Figure 6(b) illustrates that our proposed metric $-\mathcal{W}_2$ decreases, satisfying the constraint.



(a) $-\mathcal{L}_{\mathcal{U}}$ does NOT satisfy FCC (b) $-\mathcal{W}_2$ does satisfy FCC

Figure 5: FCC analysis.



(a) $-\mathcal{L}_{\mathcal{U}}$ does NOT satisfy FBC (b) $-\mathcal{W}_2$ does satisfy FBC

Figure 6: FBC analysis.

In summary, our empirical results align with our theoretical analysis, confirming that our proposed metric $-\mathcal{W}_2$ performs better than the existing metric $-\mathcal{L}_{\mathcal{U}}$ in capturing feature redundancy and dimensional collapse.

6 Experiments

In this section, we add the proposed uniformity loss as an auxiliary loss term for various existing self-supervised methods, and conduct experiments on CIFAR-10 and CIFAR-100 to demonstrate its effectiveness.

Models We conduct experiments on a series of self-supervised representation learning models: (i) AlignUniform (Wang & Isola, 2020), whose loss objective consists of an alignment objective and a uniform objective; (ii) three contrastive methods, i.e., SimCLR (Chen et al., 2020), MoCo (He et al., 2020), and NNCLR (Dwibedi et al., 2021); (iii) two asymmetric models, i.e., BYOL (Grill et al., 2020) and SimSiam (Chen & He, 2021); (iv) two methods via redundancy reduction, i.e., BarlowTwins (Zbontar et al., 2021) and Zero-CL (Zhang et al., 2022b). To study the behaviors of proposed Wasserstein uniformity loss in self-supervised representation learning, we impose it as an auxiliary loss term to the following models: MoCo v2, BYOL, BarlowTwins, and Zero-CL. We also propose to use a linear decay for weighting the Wasserstein uniformity loss during the training phase by taking $\alpha_t = \alpha_{\max} - t\,(\alpha_{\max} - \alpha_{\min})/T$, where t, T, $\alpha_{\max}$, $\alpha_{\min}$, α_t are current epoch, maximum epochs, maximum weight, minimum weight, and current weight, respectively. More details on experimental settings are collected in Appendix G.1.

Accuracy and representation capacity We evaluate the above methods using two different criteria: Accuracy and representation capacity. For accuracy, we use linear evaluation accuracy measured by Top-1 accuracy (Acc@1) and Top-5 accuracy (Acc@5); another is representation capacity. For representation capacity, we use the uniformity losses: $\mathcal{L}_{\mathcal{U}}$ and $\mathcal{W}_2$, and the alignment loss $\mathcal{C}_{\mathcal{A}}$.

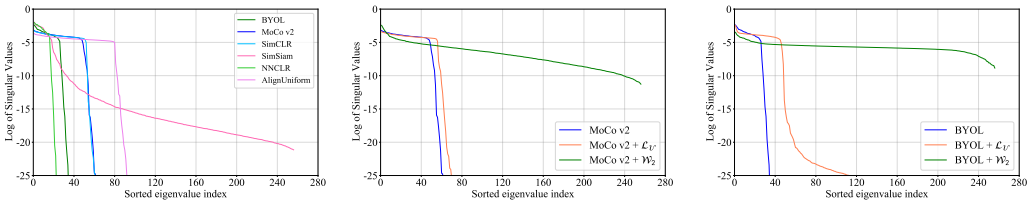
Main Results As shown in Table 2, we observe that imposing $\mathcal{W}_2$ as an additional loss term helps consistently outperform those without the loss or those with $\mathcal{L}_{\mathcal{U}}$ as the additional loss. Interestingly, although it slightly harms alignment, it results in improved uniformity and improved accuracy. This demonstrates the effectiveness of $\mathcal{W}_2$ as a uniformity loss. We emphasize that imposing an additional loss during training does not hinder the training or inference efficiency.

Convergence Analysis We test the Top-1 accuracy of these models on CIFAR-10 and CIFAR-100 via linear evaluation protocol (as described in Appendix G.1) when training them in different epochs. As shown in Figure 14. By imposing $\mathcal{W}_2$ as an additional loss for these models, it converges faster than the raw models, especially for MoCo v2 and BYOL with serious collapse problem. Our experiments show that imposing the proposed uniformity metric as an auxiliary penalty loss could largely improve uniformity but damage alignment. We further conduct uniformity and alignment analyses through all the training epochs in Figure 15 and Figure 16 respectively, see Appendix G.3.

Dimensional Collapse Analysis We visualize the singular value spectra of the learned representations (Jing et al., 2022) for various models. These spectra contain the singular values of the covariance matrix of representations from CIFAR-100 dataset, sorted in logarithmic scale order. As shown

Table 2: Main results on CIFAR-10 and CIFAR-100. Proj. and Pred. are the hidden dimensions in projector and predictor. ↑ and ↓ indicates gains and losses, respectively.

Methods	Proj.	Pred.	CIFAR-10					CIFAR-100				
			Acc@1↑	Acc@5↑	$\mathcal{W}_2 \downarrow$	$\mathcal{L}_{\mathcal{U}} \downarrow$	$\mathcal{L}_{\mathcal{A}} \downarrow$	Acc@1↑	Acc@5↑	$\mathcal{W}_2 \downarrow$	$\mathcal{L}_{\mathcal{U}} \downarrow$	$\mathcal{L}_{\mathcal{A}} \downarrow$
SimCLR	256	✗	89.85	99.78	1.04	-3.75	0.47	63.43	88.97	1.05	-3.75	0.50
NNCLR	256	256	87.46	99.63	1.23	-3.12	0.38	54.90	83.81	1.23	-3.18	0.43
SimSiam	256	256	86.71	99.67	1.19	-3.33	0.39	56.10	84.34	1.21	-3.29	0.42
AlignUniform	256	✗	90.37	99.76	0.94	-3.82	0.51	65.08	90.15	0.95	-3.82	0.53
MoCo v2	256	✗	90.65	99.81	1.06	-3.75	0.51	60.27	86.29	1.07	-3.60	0.46
MoCo v2 + $\mathcal{L}_{\mathcal{U}}$	256	✗	90.98 $\uparrow_{0.33}$	99.67	0.98 $\uparrow_{0.08}$	-3.82	0.53 $\downarrow_{0.02}$	61.21 $\uparrow_{0.94}$	87.32	0.98 $\uparrow_{0.09}$	-3.81	0.52 $\downarrow_{0.06}$
MoCo v2 + $\mathcal{W}_2$	256	✗	91.41 $\uparrow_{0.76}$	99.68	0.33 $\uparrow_{0.73}$	-3.84	0.63 $\downarrow_{0.12}$	63.68 $\uparrow_{3.41}$	88.48	0.28 $\uparrow_{0.79}$	-3.86	0.66 $\downarrow_{0.20}$
BYOL	256	256	89.53	99.71	1.21	-2.99	**0.31**	63.66	88.81	1.20	-2.87	**0.33**
BYOL + $\mathcal{L}_{\mathcal{U}}$	256	✗	90.09 $\uparrow_{0.56}$	99.75	1.09 $\uparrow_{0.12}$	-3.66	0.40 $\downarrow_{0.09}$	62.68 $\downarrow_{0.98}$	88.44	1.08 $\uparrow_{0.12}$	-3.70	0.51 $\downarrow_{0.18}$
BYOL + $\mathcal{W}_2$	256	256	90.31 $\uparrow_{0.78}$	99.77	0.38 $\uparrow_{0.83}$	-3.90	0.65 $\downarrow_{0.34}$	65.16 $\uparrow_{1.50}$	89.25	0.36 $\uparrow_{0.84}$	-3.91	0.69 $\downarrow_{0.36}$
BarlowTwins	256	✗	91.16	99.80	0.22	-3.91	0.75	68.19	90.64	0.23	-3.91	0.75
BarlowTwins + $\mathcal{L}_{\mathcal{U}}$	256	✗	91.38 $\uparrow_{0.22}$	99.77	0.21 $\uparrow_{0.01}$	-3.92	0.76 $\downarrow_{0.01}$	68.41 $\uparrow_{0.22}$	90.99	0.22 $\uparrow_{0.01}$	-3.91	0.76 $\downarrow_{0.01}$
BarlowTwins + $\mathcal{W}_2$	256	✗	**91.43** $\uparrow_{0.27}$	99.78	0.19 $\uparrow_{0.03}$	-3.92	0.76 $\downarrow_{0.01}$	68.47 $\uparrow_{0.28}$	90.64	0.19 $\uparrow_{0.04}$	-3.91	0.79 $\downarrow_{0.04}$
Zero-CL	256	✗	91.35	99.74	0.15	**-3.94**	0.70	68.50	90.97	0.15	-3.93	0.75
Zero-CL + $\mathcal{L}_{\mathcal{U}}$	256	✗	91.28 $\downarrow_{0.07}$	99.74	0.15	**-3.94**	0.72 $\downarrow_{0.02}$	68.44 $\downarrow_{0.06}$	90.91	0.15	-3.93	0.74 $\uparrow_{0.01}$
Zero-CL + $\mathcal{W}_2$	256	✗	91.42 $\uparrow_{0.07}$	**99.82**	**0.14** $\uparrow_{0.01}$	**-3.94**	0.71 $\downarrow_{0.01}$	**68.55** $\uparrow_{0.05}$	**91.02**	**0.14** $\uparrow_{0.01}$	**-3.94**	0.76 $\downarrow_{0.01}$



(a) Singular Value Spectra (b) MoCo v2 (c) BYOL

Figure 7: Dimensional collapse analysis on CIFAR-100 dataset.

in Figure 7(a), most singular values collapse to zeros in most models, indicating a large number of collapsed coordinates in most models. To further understand how the additional loss $\mathcal{W}_2$ helps present dimensional collapse, we add $\mathcal{W}_2$ as an additional loss for Moco v2 and BYOL, the numbers of collapsed coordinates decrease to zeros in both cases; see Figure 7(b) and Figure 7(c). This verifies that our proposed uniformity loss $\mathcal{W}_2$ can effectively address the dimensional collapse issue for Moco v2 and BYOL. In contrast, $\mathcal{L}_{\mathcal{U}}$ can not effectively present dimensional collapse.

7 Conclusion

In this paper, we have delineated four essential properties that an ideal uniformity metric should fulfill. However, the existing uniformity metric proposed by Wang & Isola (2020) fails to meet three of these properties, indicating its incapacity to address dimensional collapse adequately. Empirical investigations corroborate this observation. To overcome this drawback, we introduce a new uniformity metric that satisfies all four properties, demonstrating a remarkable ability to accurately capture dimensional collapse. When integrated as an auxiliary loss in various self-supervised learning methods, our proposed uniformity metric consistently improves their performance in downstream tasks. Nonetheless, it is worth noting that the four identified properties may not encompass a comprehensive characterization of an ideal uniformity metric, warranting further exploration in future studies.

Acknowledgement

Benyou Wang was partially supported by the Shenzhen Science and Technology Program (JCYJ20220818103001002), Shenzhen Doctoral Startup Funding (RCBS20221008093330065), and Tianyuan Fund for Mathematics of National Natural Science Foundation of China (NSFC) (12326608). Qiang Sun was partially supported in part by the Natural Sciences and Engineering Research Council of Canada under Grant RGPIN-2018-06484 and a Data Sciences Institute Catalyst Grant.

REFERENCES

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. In *ICML*, 2019.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022.

A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 1943.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12:805–849, 2012.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.

Henry Cohn and Abhinav Kumar. Universally optimal distribution of points on spheres. *Journal of the American Mathematical Society*, 2007.

Victor Guilherme Turrisi da Costa, Enrico Fini, Moin Nabi, N. Sebe, and Elisa Ricci. Solo-learn: A library of self-supervised methods for visual representation learning. *JMLR*, 2022.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *ICCV*, 2021.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *ArXiv*, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altch'e, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. On feature decorrelation in self-supervised learning. In *ICCV*, 2021.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *ICLR*, 2022.

Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations. In *ICLR*, 2021.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *ICLR*, 2022.

David Lindley and Solomon Kullback. Information theory and statistics. *Journal of the American Statistical Association*, 54:825, 1959.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017.

Ingram Olkin and Friedrich Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263, 1982.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *ICML*, 2021.

Ramon Van Handel. Probability in high dimension. *Lecture Notes (Princeton University)*, 2016.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, 2021.

Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *ICCV*, 2021.

Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. In *CVPR*, 2021.

Yang You, Igor Gitman, and Boris Ginsburg. Scaling sgd batch size to 32k for imagenet training. *ArXiv*, 2017.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, 2021.

Chaoning Zhang, Kang Zhang, Chenshuang Zhang, Trung X. Pham, Chang D. Yoo, and In So Kweon. How does simsiam avoid collapse without negative samples? a unified understanding with self-supervised contrastive learning. In *ICLR*, 2022a.

Shaofeng Zhang, Feng Zhu, Junchi Yan, Rui Zhao, and Xiaokang Yang. Zero-CL: Instance and feature decorrelation for negative-free symmetric contrastive learning. In *ICLR*, 2022b.

Xiangyu Zhao, Raviteja Vemulapalli, P. A. Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive learning for label efficient semantic segmentation. In *ICCV*, 2021.

Appendix

Table of Contents

A Probability density function of $\mathbf{Y}_i$ **12**

B Proof of Theorem 2 **13**

C Examining the four properties for two uniformity metrics **15**
C.1 Proof for $-\mathcal{L}_{\mathcal{U}}$ 15
C.2 Proof for $-\mathcal{W}_2$ 16

D Further comparisons between $\mathbf{Y}$ and $\widehat{\mathbf{Y}}$ **17**
D.1 Binning Densities of Y_i and $\widehat{Y}_i$ 17
D.2 A two-dimensional visualization for $\mathbf{Y}$ and $\widehat{\mathbf{Y}}$ 17

E Statistical distances over Gaussian distributions **17**

F Additional numerical studies on the two metrics **18**
F.1 Correlation between $-\mathcal{L}_{\mathcal{U}}$ and $-\mathcal{W}_2$ 18
F.2 On Instance Cloning Constraint (ICC) 18
F.3 Understanding Property 4: Why does it relate to dimensional collapse 19

G Experiment settings and convergence analysis **19**
G.1 Experiments Setting 19
G.2 Convergence Analysis for Top-1 Accuracy 20
G.3 Convergence analysis for Uniformity and Alignment 20

H Excessively large means can cause collapsed representations **20**

A Probability density function of $\mathbf{Y}_i$

Lemma 2. *Let $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_m)$ and $\mathbf{Y} = \mathbf{Z}/\|\mathbf{Z}\|_2$. Then the probability density function of Y_i, the i-th coordinate of $\mathbf{Y}$ is:*

$$f_{Y_i}(y_i) = \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}(1-y_i^2)^{(m-3)/2}, \ \ \forall\, y_i \in [-1,1].$$

Proof. Let $\mathbf{Z} = [Z_1, Z_2, \cdots, Z_m] \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_m)$, then $Z_i \sim \mathcal{N}(0, \sigma^2), \forall i \in [1, m]$. Let $U = Z_i/\sigma \sim \mathcal{N}(0,1)$, $V = \sum_{j \neq i}^m (Z_j/\sigma)^2 \sim \mathcal{X}^2(m-1)$, then U and V are independent with each other. The random variable $T = \frac{U}{\sqrt{V/(m-1)}}$ follows the Student's t-distribution with $m-1$ degrees of freedom, and its probability density function (pdf) is:

$$f_T(t) = \frac{\Gamma(m/2)}{\sqrt{(m-1)\pi}\Gamma((m-1)/2)}(1+\frac{t^2}{m-1})^{-m/2}.$$

For random variable Y_i, we have

$$Y_i = \frac{Z_i}{\sqrt{\sum_{i=1}^m Z_i^2}} = \frac{Z_i}{\sqrt{Z_i^2 + \sum_{j \neq i}^m Z_j^2}} = \frac{Z_i/\sigma}{\sqrt{(Z_i/\sigma)^2 + \sum_{j \neq i}^m (Z_j/\sigma)^2}} = \frac{U}{\sqrt{U^2+V}},$$

and then $T = \frac{U}{\sqrt{V/(m-1)}} = \frac{\sqrt{m-1}Y_i}{\sqrt{1-Y_i^2}}$, $Y_i = \frac{T}{\sqrt{T^2+m-1}}$. Therefore, the cumulative distribution function (cdf) of T is:

$$
\begin{aligned}
F_{Y_i}(y_i) = P(\{Y_i \le y_i\}) &= \begin{cases} P(\{Y_i \le y_i\}) & y_i \le 0 \\ P(\{Y_i \le 0\}) + P(\{0 < Y_i \le y_i\}) & y_i > 0 \end{cases} \\
&= \begin{cases} P(\{\frac{T}{\sqrt{T^2+m-1}} \le y_i\}) & y_i \le 0 \\ P(\{\frac{T}{\sqrt{T^2+m-1}} \le 0\}) + P(\{0 < \frac{T}{\sqrt{T^2+m-1}} \le y_i\}) & y_i > 0 \end{cases} \\
&= \begin{cases} P(\{\frac{T^2}{T^2+m-1} \ge y_i^2, T \le 0\}) & y_i \le 0 \\ P(\{T \le 0\} + P(\{\frac{T^2}{T^2+m-1} \le y_i^2, T > 0\}) & y_i > 0 \end{cases} \\
&= \begin{cases} P(\{T \le \frac{\sqrt{m-1}y_i}{\sqrt{1-y_i^2}}\}) & y_i \le 0 \\ P(\{T \le 0\} + P(\{0 < T \le \frac{\sqrt{m-1}y_i}{\sqrt{1-y_i^2}}\}) & y_i > 0 \end{cases} \\
&= P(\{T \le \frac{\sqrt{m-1}y_i}{\sqrt{1-y_i^2}}\}) = F_T(\frac{\sqrt{m-1}y_i}{\sqrt{1-y_i^2}}).
\end{aligned}
$$

The probability density function of Y_i can then be derived as:

$$
\begin{aligned}
f_{Y_i}(y_i) = \frac{d}{dy_i} F_{Y_i}(y_i) &= \frac{d}{dy_i} F_T(\frac{\sqrt{m-1}y_i}{\sqrt{1-y_i^2}}) \\
&= f_T(\frac{\sqrt{m-1}y_i}{\sqrt{1-y_i^2}}) \frac{d}{dy_i}(\frac{\sqrt{m-1}y_i}{\sqrt{1-y_i^2}}) \\
&= [\frac{\Gamma(m/2)}{\sqrt{(m-1)\pi}\Gamma((m-1)/2)}(1-y_i^2)^{m/2}][\sqrt{m-1}(1-y_i^2)^{-3/2}] \\
&= \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}(1-y_i^2)^{(m-3)/2}.
\end{aligned}
$$

□

B PROOF OF THEOREM 2

Proof. According to the Lemma 2, the pdf of Y_i and $\widehat{Y}_i$ are:

$$
f_{Y_i}(y) = \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}(1-y^2)^{(m-3)/2}, \quad f_{\widehat{Y}_i}(y) = \sqrt{\frac{m}{2\pi}} \exp\{-\frac{my^2}{2}\}.
$$

Then the Kullback-Leibler divergence between Y_i and $\widehat{Y}_i$ is

$$
\begin{aligned}
D_{\mathrm{KL}}(Y_i \| \widehat{Y}_i) &= \int_{-1}^{1} f_{Y_i}(y)[\log f_{Y_i}(y) - \log f_{\widehat{Y}_i}(y)]dy \\
&= \int_{-1}^{1} f_{Y_i}(y)[\log \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)} + \frac{m-3}{2}\log(1-y^2) - \log\sqrt{\frac{m}{2\pi}} + \frac{my^2}{2}]dy \\
&= \log\sqrt{\frac{2}{m}}\frac{\Gamma(m/2)}{\Gamma((m-1)/2)} + \int_{-1}^{1} f_{Y_i}(y)[\frac{m-3}{2}\log(1-y^2) + \frac{my^2}{2}]dy.
\end{aligned}
$$

Letting $\mu = y^2$, we have $y = \sqrt{\mu}$ and $dy = \frac{1}{2}\mu^{-\frac{1}{2}}du$. Thus,

$$\begin{aligned}
\mathcal{A} &:= \int_{-1}^{1} f_{Y_i}(y)[\frac{m-3}{2}\log(1-y^2) + \frac{my^2}{2}]dy \\
&= 2\int_0^1 \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}(1-y^2)^{\frac{m-3}{2}}[\frac{m-3}{2}\log(1-y^2) + \frac{my^2}{2}]dy \\
&= \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}\int_0^1 (1-\mu)^{\frac{m-3}{2}}[\frac{m-3}{2}\log(1-\mu) + \frac{m}{2}\mu]\mu^{-\frac{1}{2}}d\mu \\
&= \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}\frac{m-3}{2}\int_0^1 (1-\mu)^{\frac{m-3}{2}}\mu^{-\frac{1}{2}}\log(1-\mu)d\mu \\
&+ \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}\frac{m}{2}\int_0^1 (1-\mu)^{\frac{m-3}{2}}\mu^{\frac{1}{2}}d\mu.
\end{aligned}$$

By using the property of Beta distribution, and the inequality that $\frac{-\mu}{1-\mu} \leq \log(1-\mu) \leq -\mu$, we have

$$\begin{aligned}
\mathcal{A}_1 &:= \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}\frac{m-3}{2}\int_0^1 (1-\mu)^{\frac{m-3}{2}}\mu^{-\frac{1}{2}}\log(1-\mu)d\mu \\
&\leq -\frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}\frac{m-3}{2}\int_0^1 (1-\mu)^{\frac{m-3}{2}}\mu^{\frac{1}{2}}d\mu \\
&= -\frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}\frac{m-3}{2}B(\frac{3}{2}, \frac{m-1}{2}) \text{ and} \\
\mathcal{A}_2, &:= \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}\frac{m}{2}\int_0^1 (1-\mu)^{\frac{m-3}{2}}\mu^{\frac{1}{2}}d\mu \\
&= \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}\frac{m}{2}B(\frac{3}{2}, \frac{m-1}{2}).
\end{aligned}$$

Then, for $\mathcal{A}$, we have

$$\begin{aligned}
\mathcal{A} = \mathcal{A}_1 + \mathcal{A}_2 &\leq -\frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}\frac{m-3}{2}B(\frac{3}{2}, \frac{m-1}{2}) + \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}\frac{m}{2}B(\frac{3}{2}, \frac{m-1}{2}) \\
&= \frac{3}{2}\frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}B(\frac{3}{2}, \frac{m-1}{2}) = \frac{3}{2}\frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}\frac{\Gamma(3/2)\Gamma((m-1)/2)}{\Gamma((m+2)/2)} \\
&= \frac{3}{2}\frac{\Gamma(3/2)\Gamma(m/2)}{\sqrt{\pi}\Gamma((m+2)/2)} = \frac{3}{2}\frac{(\sqrt{\pi}/2)\Gamma(m/2)}{\sqrt{\pi}\Gamma((m+2)/2)} = \frac{3}{4}\frac{\Gamma(m/2)}{\Gamma((m+2)/2)}.
\end{aligned}$$

Using the Stirling formula, we have $\Gamma(x+\alpha) \to \Gamma(x)x^{\alpha}$ as $x \to \infty$ and thus

$$\begin{aligned}
\lim_{m\to\infty} D_{\mathrm{KL}}(Y_i \| \widehat{Y}_i) &= \lim_{m\to\infty} \log\sqrt{\frac{2}{m}}\frac{\Gamma(m/2)}{\Gamma((m-1)/2)} + \lim_{m\to\infty}\mathcal{A} \\
&\leq \lim_{m\to\infty} \log\sqrt{\frac{2}{m}}\frac{\Gamma((m-1)/2)(\frac{m-1}{2})^{1/2}}{\Gamma((m-1)/2)} + \lim_{m\to\infty}\frac{3}{4}\frac{\Gamma(m/2)}{\Gamma((m+2)/2)} \\
&= \lim_{m\to\infty} \log\sqrt{\frac{2}{m}}\sqrt{\frac{m-1}{2}} + \frac{3}{4}\frac{\Gamma(m/2)}{\Gamma(m/2)m} = \lim_{m\to\infty}\log\sqrt{\frac{m-1}{m}} + \frac{3}{4m} = 0.
\end{aligned}$$

We further use T_2 inequality (Van Handel, 2016, Theorem 4.31) to derive the quadratic Wasserstein metric (Van Handel, 2016, Definition 4.29) as:

$$\lim_{m\to\infty}\mathcal{W}_2(Y_i, \widehat{Y}_i) \leq \lim_{m\to\infty}\sqrt{\frac{2}{m}D_{\mathrm{KL}}(Y_i\|\widehat{Y}_i)} = 0.$$

□

C EXAMINING THE FOUR PROPERTIES FOR TWO UNIFORMITY METRICS

C.1 PROOF FOR $-\mathcal{L}_\mathcal{U}$

Property 1 can be easily verified for $-\mathcal{L}_\mathcal{U}$ using the definitions and thus we skip the verification. We only examine the other three properties for the uniformity metric $-\mathcal{L}_\mathcal{U}$.

First, we prove that $-\mathcal{L}_\mathcal{U}$ cannot satisfy Property 2. Due to the definition of $\mathcal{L}_\mathcal{U}$ in Eqn. (2), we have

$$\begin{aligned}\mathcal{L}_\mathcal{U}(\mathcal{D} \uplus \mathcal{D}) &:= \log \frac{1}{2n(2n-1)/2} \left(4 \sum_{i=2}^{n} \sum_{j=1}^{i-1} e^{-t\|\mathbf{z}_i - \mathbf{z}_j\|_2^2} + \sum_{i=1}^{n} e^{-t\|\mathbf{z}_i - \mathbf{z}_i\|_2^2} \right) \\ &= \log \frac{1}{2n(2n-1)/2} \left(4 \sum_{i=2}^{n} \sum_{j=1}^{i-1} e^{-t\|\mathbf{z}_i - \mathbf{z}_j\|_2^2} + n \right).\end{aligned} \tag{9}$$

Letting $G = \sum_{i=2}^{n} \sum_{j=1}^{i-1} e^{-t\|\mathbf{z}_i - \mathbf{z}_j\|_2^2}$, we have

$$G = \sum_{i=2}^{n} \sum_{j=1}^{i-1} e^{-t\|\mathbf{z}_i - \mathbf{z}_j\|_2^2} \leq \sum_{i=2}^{n} \sum_{j=1}^{i-1} e^{-t\|\mathbf{z}_i - \mathbf{z}_i\|_2^2} = n(n-1)/2,$$

and $G = n(n-1)/2$ if and only if $\mathbf{z}_1 = \mathbf{z}_2 = \ldots = \mathbf{z}_n$. Thus

$$\begin{aligned}\mathcal{L}_\mathcal{U}(\mathcal{D} \uplus \mathcal{D}) - \mathcal{L}_\mathcal{U}(\mathcal{D}) &= \log \frac{4G+n}{2n(2n-1)/2} - \log \frac{G}{n(n-1)/2} \\ &= \log \frac{(4G+n)n(n-1)/2}{2nG(2n-1)/2} = \log \frac{(4G+n)(n-1)}{4nG-2G} \\ &= \log \frac{4nG - 4G + n^2 - n}{4nG - 2G} \geq \log 1 = 0.\end{aligned}$$

The above equality holds if and only if $G = n(n-1)/2$, which requires $\mathbf{z}_1 = \mathbf{z}_2 = ... = \mathbf{z}_n$, a trivial case when all representations collapse to one constant point. We have excluded this trivial case, and thus $-\mathcal{L}_\mathcal{U}(\mathcal{D} \uplus \mathcal{D}) < -\mathcal{L}_\mathcal{U}(\mathcal{D})$. Therefore, the uniformity metric $-\mathcal{L}_\mathcal{U}$ does not satisfy Property 2.

Second, we prove that $-\mathcal{L}_\mathcal{U}$ cannot satisfy Property 3. Letting $\widehat{\mathbf{z}}_i = \mathbf{z}_i \oplus \mathbf{z}_i$ and $\widehat{\mathbf{z}}_j = \mathbf{z}_j \oplus \mathbf{z}_j$, we have

$$\mathcal{L}_\mathcal{U}(\mathcal{D} \oplus \mathcal{D}) := \log \frac{1}{n(n-1)/2} \sum_{i=2}^{n} \sum_{j=1}^{i-1} e^{-t\|\widehat{\mathbf{z}}_i - \widehat{\mathbf{z}}_j\|_2^2}.$$

As $\widehat{\mathbf{z}}_i = [z_{i1}, z_{i2}, ..., z_{im}, z_{i1}, z_{i2}, ..., z_{im}]^\mathrm{T}$ and $\widehat{\mathbf{z}}_j = [z_{j1}, z_{j2}, ..., z_{jm}, z_{j1}, z_{j2}, ..., z_{jm}]^\mathrm{T}$, we have $\|\widehat{\mathbf{z}}_i\| = \sqrt{2}$, $\|\widehat{\mathbf{z}}_j\| = \sqrt{2}$, and $\langle \widehat{\mathbf{z}}_i, \widehat{\mathbf{z}}_j \rangle = 4\langle \mathbf{z}_i, \mathbf{z}_j \rangle$. Thus

$$\|\widehat{\mathbf{z}}_i - \widehat{\mathbf{z}}_j\|_2^2 = 4 - 4\langle \mathbf{z}_i, \mathbf{z}_j \rangle = 2\|\mathbf{z}_i - \mathbf{z}_j\|_2^2 > \|\mathbf{z}_i - \mathbf{z}_j\|_2^2.$$

Therefore, $-\mathcal{L}_\mathcal{U}(\mathcal{D} \oplus \mathcal{D}) > -\mathcal{L}_\mathcal{U}(\mathcal{D})$, indicating that the uniformity metric $-\mathcal{L}_\mathcal{U}$ cannot satisfy the Property 3.

Finally, we prove that the baseline metric $-\mathcal{L}_\mathcal{U}$ cannot satisfy the Property 4. Letting $\widehat{\mathbf{z}}_i = \mathbf{z}_i \oplus \mathbf{0}^k$ and $\widehat{\mathbf{z}}_j = \mathbf{z}_j \oplus \mathbf{0}^k$, we have

$$\mathcal{L}_\mathcal{U}(\mathcal{D} \oplus \mathbf{0}^k) := \log \frac{1}{n(n-1)/2} \sum_{i=2}^{n} \sum_{j=1}^{i-1} e^{-t\|\widehat{\mathbf{z}}_i - \widehat{\mathbf{z}}_j\|_2^2}.$$

As $\widehat{\mathbf{z}}_i = [z_{i1}, z_{i2}, ..., z_{im}, 0, 0, ..., 0]^\mathrm{T}$ and $\widehat{\mathbf{z}}_j = [z_{j1}, z_{j2}, ..., z_{jm}, 0, 0, ..., 0]^\mathrm{T}$, we have $\|\widehat{\mathbf{z}}_i\| = \|\mathbf{z}_i\|$ = 1, $\|\widehat{\mathbf{z}}_j\| = \|\mathbf{z}_j\|$ = 1, $\langle \widehat{\mathbf{z}}_i, \widehat{\mathbf{z}}_j \rangle = \langle \mathbf{z}_i, \mathbf{z}_j \rangle$, and thus

$$\|\widehat{\mathbf{z}}_i - \widehat{\mathbf{z}}_j\|_2^2 = 2 - 2\langle \widehat{\mathbf{z}}_i, \widehat{\mathbf{z}}_j \rangle = 2 - 2\langle \mathbf{z}_i, \mathbf{z}_j \rangle = \|\mathbf{z}_i - \mathbf{z}_j\|_2^2.$$

Therefore, $-\mathcal{L}_\mathcal{U}(\mathcal{D} \oplus \mathbf{0}^k) = -\mathcal{L}_\mathcal{U}(\mathcal{D})$, indicating that the uniformity metric $-\mathcal{L}_\mathcal{U}$ cannot satisfy Property 4.

C.2 PROOF FOR $-\mathcal{W}_2$

Property 1 are easily verified for $-\mathcal{W}_2$. We only examine the rest three properties for the proposed uniformity metric $-\mathcal{W}_2$.

First, we prove that our proposed metric $-\mathcal{W}_2$ satisfies Property 2. Let $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ be defined as above, for $\mathcal{D} \uplus \mathcal{D} = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n, \mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n\}$, the mean and covariance estimators are

$$\widetilde{\boldsymbol{\mu}} = \frac{1}{2n}\sum_{i=1}^{n} 2\mathbf{z}_i = \widehat{\boldsymbol{\mu}}, \quad \widetilde{\boldsymbol{\Sigma}} = \frac{1}{2n}\sum_{i=1}^{n} 2(\mathbf{z}_i - \widetilde{\boldsymbol{\mu}})^T(\mathbf{z}_i - \widetilde{\boldsymbol{\mu}}) = \widehat{\boldsymbol{\Sigma}},$$

which agree with those for $\mathcal{D}$. Then we have

$$\mathcal{W}_2(\mathcal{D} \uplus \mathcal{D}) := \sqrt{\|\widehat{\boldsymbol{\mu}}\|_2^2 + 1 + \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}) - \frac{2}{\sqrt{m}}\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}^{1/2})} = \mathcal{W}_2(\mathcal{D}).$$

Therefore, our proposed metric $-\mathcal{W}_2$ satisfies Property 2.

Second, we prove that $-\mathcal{W}_2$ satisfies Property 3. Let $\widetilde{\mathbf{z}}_i = \mathbf{z}_i \oplus \mathbf{z}_i \in \mathbb{R}^{2m}$. For $\mathcal{D} \oplus \mathcal{D}$, the mean and covariance estimators are:

$$\widetilde{\boldsymbol{\mu}} = \begin{pmatrix} \widehat{\boldsymbol{\mu}} \\ \widehat{\boldsymbol{\mu}} \end{pmatrix}, \quad \widetilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \widehat{\boldsymbol{\Sigma}} & \widehat{\boldsymbol{\Sigma}} \\ \widehat{\boldsymbol{\Sigma}} & \widehat{\boldsymbol{\Sigma}} \end{pmatrix}.$$

We easily have

$$\widetilde{\boldsymbol{\Sigma}}^{1/2} = \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}^{1/2}/\sqrt{2} & \widehat{\boldsymbol{\Sigma}}^{1/2}/\sqrt{2} \\ \widehat{\boldsymbol{\Sigma}}^{1/2}/\sqrt{2} & \widehat{\boldsymbol{\Sigma}}^{1/2}/\sqrt{2} \end{pmatrix}, \ \operatorname{tr}(\widetilde{\boldsymbol{\Sigma}}) = 2\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}), \text{ and } \operatorname{tr}(\widetilde{\boldsymbol{\Sigma}}^{1/2}) = \sqrt{2}\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}^{1/2}).$$

Thus

$$\begin{aligned}
\mathcal{W}_2(\mathcal{D} \oplus \mathcal{D}) &:= \sqrt{\|\widetilde{\boldsymbol{\mu}}\|_2^2 + 1 + \operatorname{tr}(\widetilde{\boldsymbol{\Sigma}}) - \frac{2}{\sqrt{2m}}\operatorname{tr}(\widetilde{\boldsymbol{\Sigma}}^{1/2})} \\
&= \sqrt{2\|\widehat{\boldsymbol{\mu}}\|_2^2 + 1 + 2\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}) - \frac{2\sqrt{2}}{\sqrt{2m}}\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}^{1/2})} \\
&> \sqrt{\|\widehat{\boldsymbol{\mu}}\|_2^2 + 1 + \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}) - \frac{2}{\sqrt{m}}\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}^{1/2})} = \mathcal{W}_2(\mathcal{D}).
\end{aligned}$$

Therefore, $-\mathcal{W}_2(\mathcal{D} \oplus \mathcal{D}) < -\mathcal{W}_2(\mathcal{D})$, indicating that our proposed metric $-\mathcal{W}_2$ could satisfy the Property 3.

Finally, we prove that our proposed metric $-\mathcal{W}_2$ satisfies Property 4. Let $\widetilde{\mathbf{z}}_i = \mathbf{z}_i \oplus \mathbf{0}^k \in \mathbb{R}^{m+k}$ with an overload of notation. For $\mathcal{D} \oplus \mathbf{0}^k$, the sample mean and covariance estimators are

$$\widetilde{\boldsymbol{\mu}} = \begin{pmatrix} \widehat{\boldsymbol{\mu}} \\ \mathbf{0}^k \end{pmatrix}, \quad \widetilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \widehat{\boldsymbol{\Sigma}} & \mathbf{0}^{m\times k} \\ \mathbf{0}^{k\times m} & \mathbf{0}^{k\times k} \end{pmatrix},$$

where $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ are defined previously. Therefore, we have $\operatorname{tr}(\widetilde{\boldsymbol{\Sigma}}) = \operatorname{tr}(\widehat{\boldsymbol{\Sigma}})$, $\operatorname{tr}(\widetilde{\boldsymbol{\Sigma}}^{1/2}) = \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}^{1/2})$, and thus

$$\begin{aligned}
\mathcal{W}_2(\mathcal{D} \oplus \mathbf{0}^k) &:= \sqrt{\|\widetilde{\boldsymbol{\mu}}\|_2^2 + 1 + \operatorname{tr}(\widetilde{\boldsymbol{\Sigma}}) - \frac{2}{\sqrt{m+k}}\operatorname{tr}(\widetilde{\boldsymbol{\Sigma}}^{1/2})} \\
&= \sqrt{\|\widehat{\boldsymbol{\mu}}\|_2^2 + 1 + \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}) - \frac{2}{\sqrt{m+k}}\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}^{1/2})} \\
&> \sqrt{\|\widehat{\boldsymbol{\mu}}\|_2^2 + 1 + \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}) - \frac{2}{\sqrt{m}}\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}^{1/2})} = \mathcal{W}_2(\mathcal{D}).
\end{aligned}$$

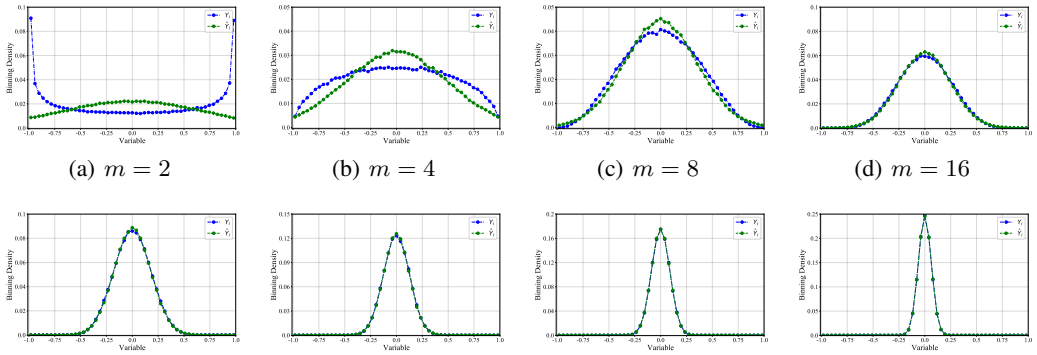
Therefore, $-\mathcal{W}_2(\mathcal{D} \oplus \mathbf{0}^k) < -\mathcal{W}_2(\mathcal{D})$, indicating that our proposed metric $-\mathcal{W}_2$ satisfies the Property 4.

D FURTHER COMPARISONS BETWEEN $\mathbf{Y}$ AND $\widehat{\mathbf{Y}}$

This section further compares the distributions of $\mathbf{Y}$ and $\widehat{\mathbf{Y}}$.

D.1 BINNING DENSITIES OF Y_i AND $\widehat{Y}_i$

We visually compare the distributions of Y_i and $\widehat{Y}_i$. To estimate the distributions of Y_i and $\widehat{Y}_i$, we bin 200,000 sampled data points into 51 groups. Figure 8 compares the binning densities of Y_i and $\widehat{Y}_i$ when $m \in \{2, 4, 8, 16, 32, 64, 128, 256\}$. We can observe that two distributions are highly overlapped when m is moderately large, e.g., $m \geq 8$ or $m \geq 16$.



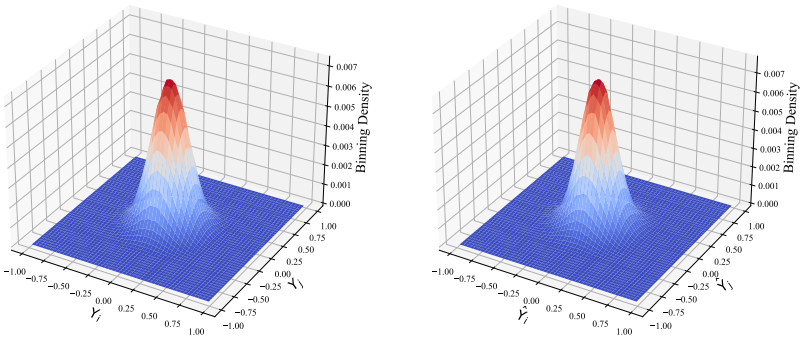
(a) $m = 2$ (b) $m = 4$ (c) $m = 8$ (d) $m = 16$

(e) $m = 32$ (f) $m = 64$ (g) $m = 128$ (h) $m = 256$

Figure 8: Comparing the binning densities of Y_i and $\widehat{Y}_i$ with various dimensions.

D.2 A TWO-DIMENSIONAL VISUALIZATION FOR $\mathbf{Y}$ AND $\widehat{\mathbf{Y}}$

By binning 2,000,000 data points into 51×51 groups in two-axis, we also analyze the joint binning densities and present 2D joint binning densities of (Y_i, Y_j) $(i \neq j)$ in Figure 9(a) and $(\widehat{Y}_i, \widehat{Y}_j)$ $(i \neq j)$ in Figure 9(b). Even if m is relatively small (i.e., 8 or 16), the densities of the two distributions are close.



(a) Density for Y_i and Y_j (b) Density for $\widehat{Y}_i$ and $\widehat{Y}_j$

Figure 9: Visualization of two arbitrary dimensions for $\mathbf{Y}$ and $\widehat{\mathbf{Y}}$ when $m = 32$.

E STATISTICAL DISTANCES OVER GAUSSIAN DISTRIBUTIONS

We first introduce the Wasserstein distance or the earth mover distance.

Definition 1. The Wasserstein distance or earth-mover distance with p norm is defined as below:

$$W_p(\mathbb{P}_r, \mathbb{P}_g) = \Big(\inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma} \big[\|x - y\|^p \big] \Big)^{1/p} . \tag{10}$$

where $\Pi(\mathbb{P}_r, \mathbb{P}_g)$ denotes the set of all joint distributions $\gamma(x, y)$ whose marginals are respectively $\mathbb{P}_r$ and $\mathbb{P}_g$. Intuitively, when viewing each distribution as a unit amount of earth/soil, the Wasserstein distance or earth-mover distance takes the minimum cost of transporting "mass" from x to y to transform the distribution $\mathbb{P}_r$ into the distribution $\mathbb{P}_g$. This distance is also called the quadratic Wasserstein distance when $p = 2$.

In this paper, we mainly exploit the quadratic Wasserstein distance over Gaussian distributions. Besides this distance, we also discuss other distribution distances as uniformity metrics and make comparisons with the Wasserstein distance. Specifically, the Kullback-Leibler divergence and the Bhattacharyya distance over Gaussian distributions are provided in Lemma 3 and Lemma 4 respectively. Both distances require full-rank covariance matrices, making them impropriate to conduct dimensional collapse analysis. In constrast, our quadratic Wasserstein distance based uniformity metric is free of such a requirement.

Lemma 3 (Kullback-Leibler divergence (Lindley & Kullback, 1959))**.** *Suppose two random variables* $\mathbf{Z}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ *and* $\mathbf{Z}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ *obey multivariate normal distributions, then Kullback-Leibler divergence between* $\mathbf{Z}1$ *and* $\mathbf{Z}_2$ *is:*

$$D_{\mathrm{KL}}(\mathbf{Z}_1, \mathbf{Z}_2) = \frac{1}{2}((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \mathrm{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1 - \mathbf{I}) + \ln \frac{\det \boldsymbol{\Sigma}_2}{\det \boldsymbol{\Sigma}_1}).$$

Lemma 4 (Bhattacharyya Distance (Bhattacharyya, 1943))**.** *Suppose two random variables* $\mathbf{Z}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ *and* $\mathbf{Z}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ *obey multivariate normal distributions,* $\boldsymbol{\Sigma} = \frac{1}{2}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$*, then bhattacharyya distance between* $\mathbf{Z}1$ *and* $\mathbf{Z}_2$ *is:*

$$\mathcal{D}_B(\mathbf{Z}_1, \mathbf{Z}_2) = \frac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2} \ln \frac{\det \boldsymbol{\Sigma}}{\sqrt{\det \boldsymbol{\Sigma}_1 \det \boldsymbol{\Sigma}_2}}.$$

F Additional Numerical Studies on the Two Metrics

F.1 Correlation between $-\mathcal{L}_{\mathcal{U}}$ and $-\mathcal{W}_2$

We employ synthetic experiments to study the uniformity metrics across different distributions. Specifically, we sample 50,000 data vectors from different distributions, such as the standard Gaussian distribution $\mathcal{N}(\mathbf{0}, I)$, the uniform Distribution $U(\mathbf{0}, \mathbf{1})$, and the mixture of Gaussians. Using these data vectors, we estimate the uniformity of different distributions by two metrics. As shown in Fig. 10, standard Gaussian distribution achieves the maximum values for both $-\mathcal{W}_2$ and $-\mathcal{L}_{\mathcal{U}}$, which indicates that multivariate standard Gaussian distributions achieve larger uniformity than other distributions. This empirical result is consistent with Fact 1 that standard Gaussian distribution (approximately) achieves the maximum uniformity.

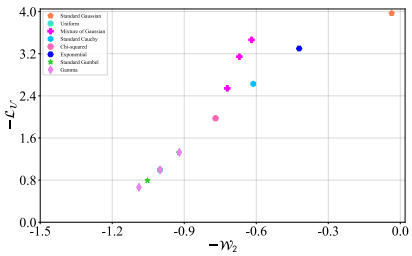


Figure 10: Uniformity analysis for various distributions by two metrics.

F.2 On Instance Cloning Constraint (ICC)

Besides empirical comparison on Feature Cloning Constraint (FCC) and On Feature Baby Constraint (FBC), we also conduct comparison on Instance Cloning Constraint (ICC). Specifically, We randomly sample 10,000 data vectors from a standard Gaussian distribution and mask 50% of their dimensions with zero-vectors, forming a new dataset $\mathcal{D}$ with an overload of notation. To investigate the impact of instance cloning, we create multiple clones of the dataset, such as $\mathcal{D} \uplus \mathcal{D}$ and $\mathcal{D} \uplus \mathcal{D} \uplus \mathcal{D}$, which correspond to one and two times cloning, respectively. We evaluate the two metrics on these datasets. Figure 11 shows that the value of $-\mathcal{L}_{\mathcal{U}}$ slightly decreases as the number of clones increases, indicating that $-\mathcal{L}_{\mathcal{U}}$ violates the equality constraint in Equation 4. In contrast, our proposed metric $-\mathcal{W}_2$ remains constant, satisfying the constraint.

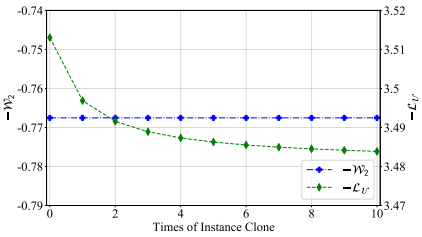
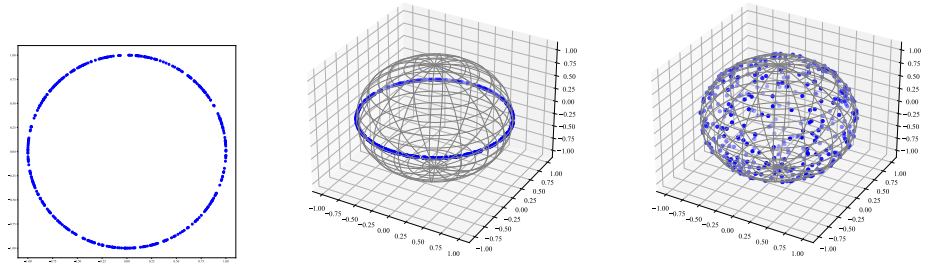


Figure 11: ICC analysis.

F.3 UNDERSTANDING PROPERTY 4: WHY DOES IT RELATE TO DIMENSIONAL COLLAPSE

This section explores Property 4 using a case study. Suppose the dataset $\mathcal{D}$ has maximal uniformity. When more coordinates with zero-values are inserted to $\mathcal{D}$, this new dataset ($\mathcal{D} \oplus \mathbf{0}^k$) cannot achieve maximal uniformity anymore, as they only occupy a small space on the surface of the unit hypersphere. Therefore, the uniformity would decrease significantly with large k.



(a) Two-dimensional visualization with no collapsed dimension

(b) Three-dimensional visualization with one collapsed dimension

(c) Three-dimensional visualization with no collapsed dimension

Figure 12: A case study for Property 4 and blue points are data vectors.

To showcase that Property 4 should involve a strict inequality, we visualize sampled data vectors. In Figure 12(a), we visualize 400 data vectors ($\mathcal{D}_1$) sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I}_2)$, and they are almost uniformly distributed on $\mathcal{S}^1$. We insert one dimension with zero-value to $\mathcal{D}_1$, and denote it as $\mathcal{D}_1 \oplus \mathbf{0}^1$, as shown in Figure 12(b). In comparison with $\mathcal{D}_2$ where 400 data vectors are sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I}_3)$, as visualized in Figure 12(c), $\mathcal{D}_1 \oplus \mathbf{0}^1$ only occupies a ring on $\mathcal{S}^2$, while $\mathcal{D}_2$ are almost uniformly distributed on $\mathcal{S}^2$. Therefore, $\mathcal{U}(\mathcal{D}_2) > \mathcal{U}(\mathcal{D}_1 \oplus \mathbf{0}^1)$. We assume no matter how great/small m is, the maximal uniformity over various dimensions m should be equal to each other. Then, we have $\mathcal{U}(\mathcal{D}_1) = \mathcal{U}(\mathcal{D}_2) > \mathcal{U}(\mathcal{D}_1 \oplus \mathbf{0}^1)$. Therefore, Property 4 should be a strict inequality.

Intuitively, increasing the value of k in Property 4 would enlarge the degree of dimensional collapse. To illustrate this point, we sample dataset $\mathcal{D}$ from an multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_m/m)$. This dataset has collapse degree close to 0%. However, when inserting m dimension zero-value vectors to $\mathcal{D}$, denoted as $\mathcal{D} \oplus \mathbf{0}^m$, half of the dimensions become collapsed. As a result, the collapse degree increases to 50%. Figure 13 visualizes such collapse of $\mathcal{D} \oplus \mathbf{0}^k$ by the singular value spectrum of the representations. We can observe that a larger k would lead to a more serious dimensional collapse. In summary, Property 4 is closely related to the dimensional collapse.

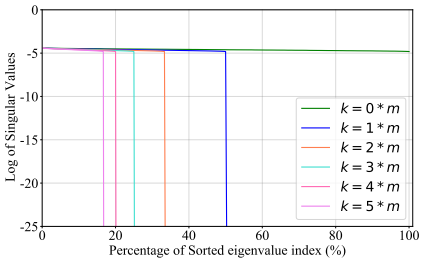


Figure 13: Singular value spectrum of $\mathcal{D} \oplus \mathbf{0}^k$.

G EXPERIMENT SETTINGS AND CONVERGENCE ANALYSIS

G.1 EXPERIMENTS SETTING

To make fair comparisons, we conduct all experiments in Sec. 6 on a single 1080 GPU. Also, we adopt the same network architecture for all models, i.e., ResNet-18 (He et al., 2016) as the encoder and a three-layer MLP as the projector. Besides, we use LARS optimizer (You et al., 2017) with a base learning rate 0.2, along with a cosine decay learning rate schedule (Loshchilov & Hutter, 2017) for all models. We evaluate all models under a linear evaluation protocol. Specifically, models are pre-trained for 500 epochs, evaluated by adding a linear classifier and training the classifier for 100 epochs while keeping the learned representations unchanged. We also deploy the same augmentation strategy for all models, which contains a series of data augmentation operations, such as color distortion, rotation, and cutout. Following da Costa et al. (2022), we set temperature $t = 0.2$ for all contrastive methods. For MoCo (He et al., 2020) and NNCLR (Dwibedi et al., 2021) that

require an extra queue to save negative samples, we set the queue size to be 2^{12}. For the linear decay for weighting the quadratic Wasserstein distance, Table 3 collects the parameter settings.

Table 3: Parameter settings for various models in experiments.

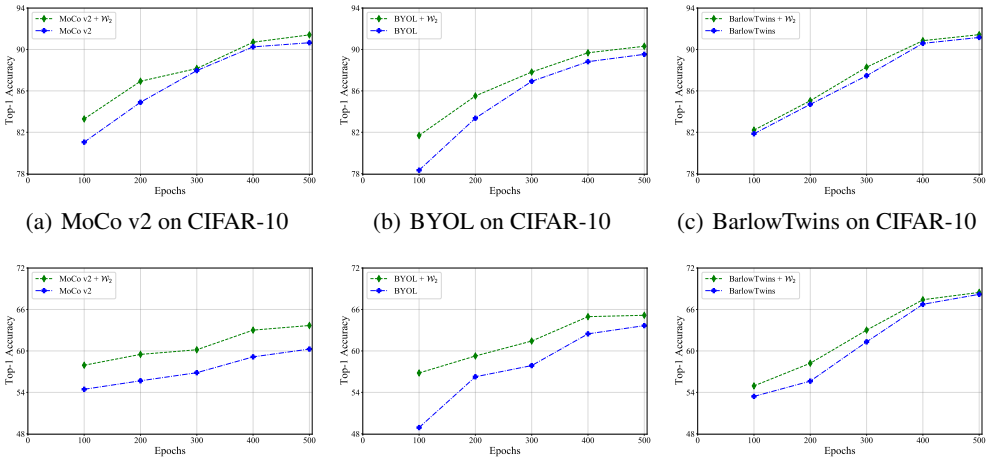
Models	MoCo v2	BYOL	BarlowTwins	Zero-CL
$\alpha_{\max}$	1.0	0.2	30.0	30.0
$\alpha_{\min}$	1.0	0.2	0	30.0

G.2 Convergence Analysis for Top-1 Accuracy

Here we show the convergence of Top-1 accuracy along all the training epochs in Fig 14. During training, we take the model checkpoint after finishing each epoch to train linear classifier, and then evaluate the Top-1 accuracy on the unseen images of the test set (on either CIFAR-10 or CIFAR-100). For both CIFAR-10 and CIFAR-100, we obverse that imposing the proposed uniformity metric as an auxiliary penalty loss helps largely improve the Top-1 accuracy, especially in the early stage.

G.3 Convergence analysis for Uniformity and Alignment

This section shows the convergence of the uniformity metric and alignment loss through all the training epochs in Figure 15 and Figure 16, respectively. During training, we take the model checkpoint after finishing each epoch to evaluate the uniformity (i.e., using the proposed metric $\mathcal{W}_2$) and alignment (Wang & Isola, 2020) on the unseen images of the test set (on either CIFAR-10 or CIFAR-100). For both CIFAR-10 and CIFAR-100, we find that imposing the proposed uniformity metric as an auxiliary penalty loss helps largely improve the uniformity. Consequently, it also lightly damages the alignment (*the smaller alighment loss, the better-aligned*) since a better uniformity usually leads to worse alignment.



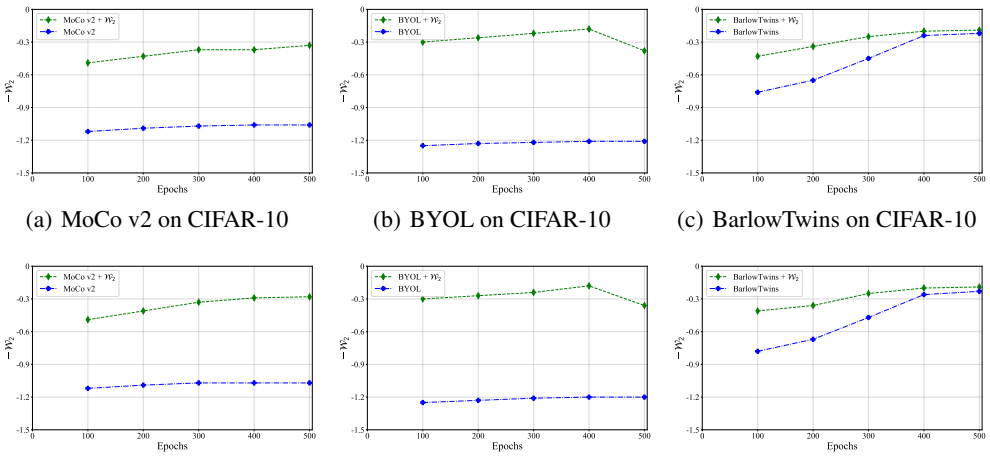
(a) MoCo v2 on CIFAR-10 (b) BYOL on CIFAR-10 (c) BarlowTwins on CIFAR-10

(d) MoCo v2 on CIFAR-100 (e) BYOL on CIFAR-100 (f) BarlowTwins on CIFAR-100

Figure 14: Convergence analysis for Top-1 accuracy during training.

H Excessively large means can cause collapsed representations

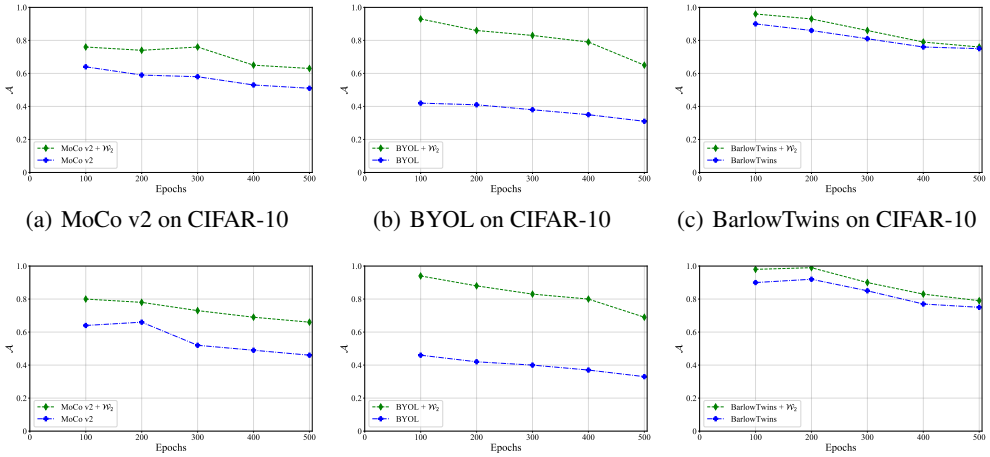
We assume $\mathbf{X}$ follows a Gaussian distribution, $\mathbf{X} \sim \mathcal{N}(0, I_2)$. By adding an additional vector to change its mean, we obtain $\mathbf{Y}$, where $\mathbf{Y} = \mathbf{X} + k\mathbf{I}$ and $\mathbf{Y} \sim \mathcal{N}(k, I_2)$. $\mathbf{I}$ is a vector of all ones, and k is a constant. We vary k from 0 to 32 and visualize ℓ_2-normalized $\mathbf{Y}$ in Figure 17. It is evident that an excessively large mean will cause representations to collapse to a single point even if the covariance matrix is an identity matrix.



(a) MoCo v2 on CIFAR-10 (b) BYOL on CIFAR-10 (c) BarlowTwins on CIFAR-10

(d) MoCo v2 on CIFAR-100 (e) BYOL on CIFAR-100 (f) BarlowTwins on CIFAR-100

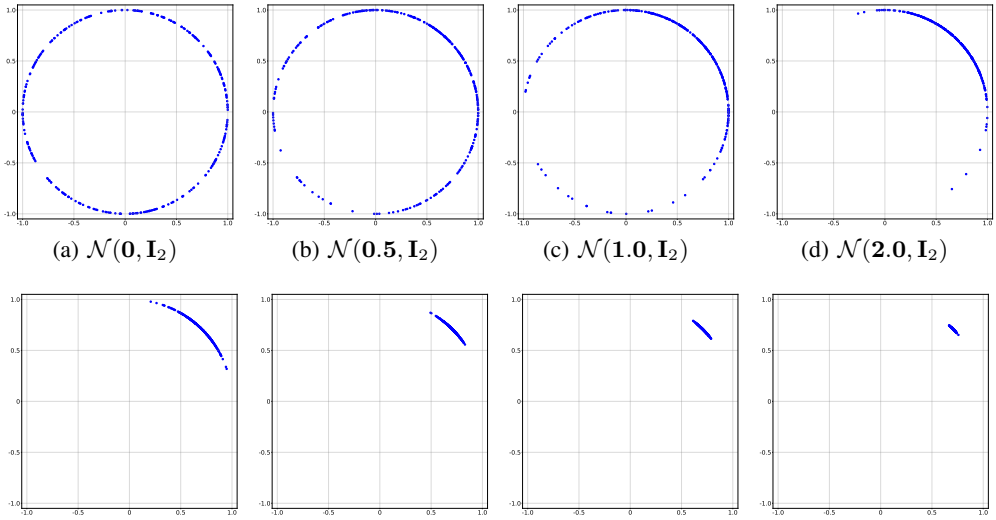
Figure 15: Visualizing uniformity during training



(a) MoCo v2 on CIFAR-10 (b) BYOL on CIFAR-10 (c) BarlowTwins on CIFAR-10

(d) MoCo v2 on CIFAR-100 (e) BYOL on CIFAR-100 (f) BarlowTwins on CIFAR-100

Figure 16: Visualizing alignment during training.



(a) $\mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ (b) $\mathcal{N}(\mathbf{0.5}, \mathbf{I}_2)$ (c) $\mathcal{N}(\mathbf{1.0}, \mathbf{I}_2)$ (d) $\mathcal{N}(\mathbf{2.0}, \mathbf{I}_2)$

(e) $\mathcal{N}(\mathbf{4.0}, \mathbf{I}_2)$ (f) $\mathcal{N}(\mathbf{8.0}, \mathbf{I}_2)$ (g) $\mathcal{N}(\mathbf{16.0}, \mathbf{I}_2)$ (h) $\mathcal{N}(\mathbf{32.0}, \mathbf{I}_2)$

Figure 17: Visualizing ℓ_2 normalized Gaussian distributions.