Appendix

Table of Contents

A 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:

$$W_p(\mathbb{P}_r, \mathbb{P}_g) = \Big(\inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}\big[\|x-y\|^p\big]\Big)^{1/p}. \tag{9}$$

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 2 and Lemma 3 respectively. Both distances require full-rank covariance matrices, making them impropriate to conduct dimensional collapse analysis. In contrast, our quadratic Wasserstein distance-based uniformity metric is free of such a requirement.

Lemma 2 (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 3 (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:*

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

B PROOF OF THEOREM 2

We first need the following lemma, whose proof is collected in the end of this section.

Lemma 4. *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:*

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

We are ready to prove Theorem 2.

Proof of Theorem 2. According to the Lemma 4, 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:

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

□

B.1 Proofs for supporting lemmas

Proof of Lemma 4. 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:

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

□

C EXAMINING THE FOUR PROPERTIES FOR TWO UNIFORMITY METRICS

C.1 PROOF OF THEOREM 1: EXAMINING THE FOUR PROPERTIES FOR $-\mathcal{L}_\mathcal{U}$

Property 1 can be easily verified for $-\mathcal{L}_\mathcal{U}$ 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 $-\mathcal{L}_\mathcal{U}$ does not satisfy Property 2. Due to the definition of $\mathcal{L}_\mathcal{U}$ in Eqn. (2), we have

$$\begin{aligned}
\mathcal{L}_\mathcal{U}(\mathcal{D} \uplus \mathcal{D}) &:= \log \frac{1}{2n(2n-1)/2}\left(4\sum_{i=2}^{n}\sum_{j=1}^{i-1} e^{-t\|\mathbf{z}_i-\mathbf{z}_j\|_2^2} + \sum_{i=1}^{n} e^{-t\|\mathbf{z}_i-\mathbf{z}_i\|_2^2}\right) \\
&= \log \frac{1}{2n(2n-1)/2}\left(4\sum_{i=2}^{n}\sum_{j=1}^{i-1} e^{-t\|\mathbf{z}_i-\mathbf{z}_j\|_2^2} + n\right).
\end{aligned} \tag{10}$$

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

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

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 $-\mathcal{L}_{\mathcal{U}}$ does not 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

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

By the definitions of $\widehat{\mathbf{z}}_i$ and $\widehat{\mathbf{z}}_j$, we have $\|\widehat{\mathbf{z}}_i\|_2 = \sqrt{2}\|\mathbf{z}_i\|_2$, $\|\widehat{\mathbf{z}}_j\|_2 = \sqrt{2}\|\mathbf{z}_j\|_2$, and $\langle \widehat{\mathbf{z}}_i, \widehat{\mathbf{z}}_j \rangle = 2\langle \mathbf{z}_i, \mathbf{z}_j \rangle$. Thus

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

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

Third, we prove that the existing metric $-\mathcal{L}_{\mathcal{U}}$ does not 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

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

By the definitions of $\widehat{\mathbf{z}}_i$ and $\widehat{\mathbf{z}}_j$, we have $\|\widehat{\mathbf{z}}_i\|_2 = \|\mathbf{z}_i\|_2$, $\|\widehat{\mathbf{z}}_j\|_2 = \|\mathbf{z}_j\|_2$, $\langle \widehat{\mathbf{z}}_i, \widehat{\mathbf{z}}_j \rangle = \langle \mathbf{z}_i, \mathbf{z}_j \rangle$, and thus

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

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

C.2 Proof of Theorem 3: Examining the four properties for $-\mathcal{W}_2$

Property 1 can be easily verified for $-\mathcal{W}_2$, and thus the proof is skipped. 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

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

Thus

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

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.

Third, 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

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

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

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

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 $\widehat{\mathbf{Y}}$

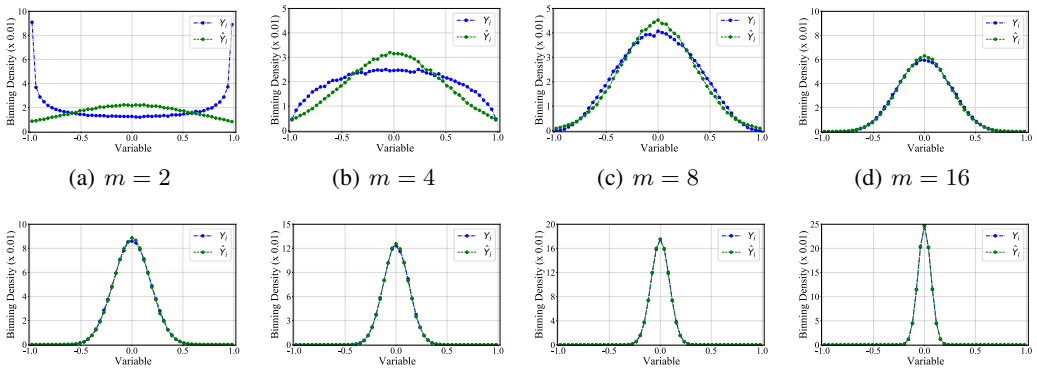
This section further compares the distributions of $\mathbf{Y}$ and $\widehat{\mathbf{Y}}$.

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 $m \geq 16$.

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., 32), the densities of the two distributions are close.

E ADDITIONAL SYNTHETIC STUDIES

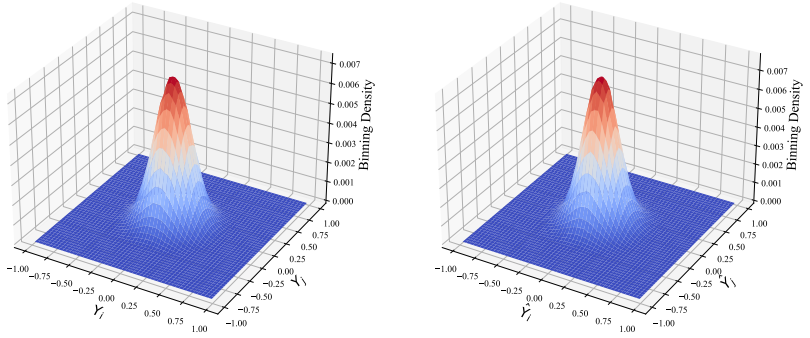
E.1 CORRELATION BETWEEN $-\mathcal{L}_{\mathcal{U}}$ AND $-\mathcal{W}_2$



(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.



(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 $m = 32$.

We employ synthetic experiments to study the uniformity metrics across different distributions. Specifically, we sample 50,000 data vectors ($m = 256$) from different distributions, such as the isotropic Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, the uniform distribution on the hyperrectangle $[\mathbf{0}, \mathbf{1}]$, and the mixture of Gaussians, etc. Then we normalize these data vectors, and estimate the uniformity of different distributions by two metrics. As shown in Fig. 10, isotropic Gaussian distribution achieves the maximum values for both $-\mathcal{W}_2$ and $-\mathcal{L}_\mathcal{U}$, which indicates that isotropic Gaussian distribution achieves larger uniformity than other distributions. This empirical result is consistent with Fact 1 that the isotropic Gaussian distribution (approximately) achieves the maximum uniformity.

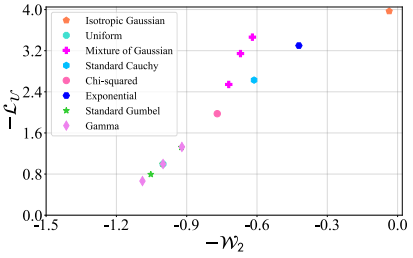


Figure 10: Uniformity analysis for various distributions by two metrics.

E.2 On Instance Cloning Constraint

In this section, we compare the two metrics in terms of Property 2 (ICC). Specifically, we randomly sample 1,000 data vectors from the isotropic Gaussian distribution ($m = 32$) and then mask 50% of their coordinates with zeros, 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, in-

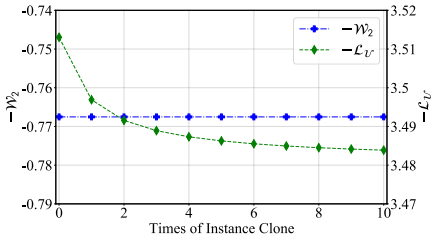
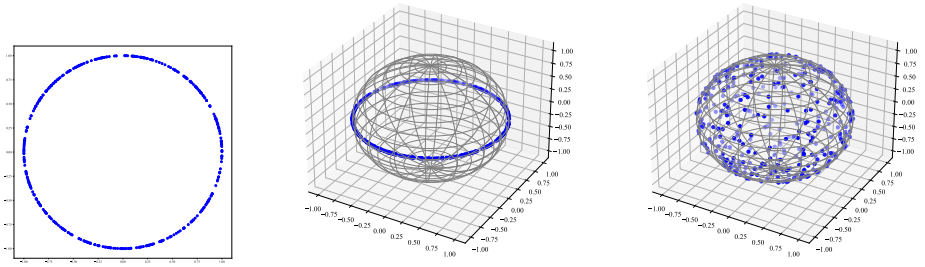


Figure 11: ICC analysis.



(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.

dicating that $-\mathcal{L}_{\mathcal{U}}$ violates the equality in Equation 4. In contrast, our proposed metric $-\mathcal{W}_2$ remains constant, satisfying the equality.

E.3 Understanding Property 4: Why does it relate to dimensional collapse?

This section delves into Property 4 through case studies. Let us begin with a thought experiment. Consider a dataset $\mathcal{D}$ with instances uniformly distributed on the unit hypersphere, thereby possessing (almost) maximal uniformity. When additional coordinates with zeros are inserted to each instance of $\mathcal{D}$, forming a new dataset $\mathcal{D} \oplus \mathbf{0}^k$, it can no longer maintain maximal uniformity. This is because, the new dataset only occupies a small area of the unit hypersphere. Consequently, as k increases, the uniformity of the dataset would decrease significantly.

Let us visualize this thought experiment using synthetic studies. In Figure 12(a), we present 400 data vectors ($\mathcal{D}_1$) sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I}_2)$, which are also nearly uniformly distributed on $\mathcal{S}^1$. By inserting one zero-coordinate to each instance of $\mathcal{D}_1$, we obtain a new dataset $\mathcal{D}_1 \oplus \mathbf{0}^1$, as depicted in Figure 12(b). We also construct another dataset $\mathcal{D}_2$ consisting of 400 data vectors sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I}_3)$, visualized in Figure 12(c). Notably, $\mathcal{D}_1 \oplus \mathbf{0}^1$ forms a ring on $\mathcal{S}^2$, while $\mathcal{D}_2$ is almost uniformly distributed over $\mathcal{S}^2$. Naturally, $\mathcal{U}(\mathcal{D}_2) > \mathcal{U}(\mathcal{D}_1 \oplus \mathbf{0}^1)$. If $\mathcal{U}(\mathcal{D}_1) = \mathcal{U}(\mathcal{D}_2)$[4], then $\mathcal{U}(\mathcal{D}_1) = \mathcal{U}(\mathcal{D}_2) > \mathcal{U}(\mathcal{D}_1 \oplus \mathbf{0}^1)$. This partially confirms the validity of Property 4.

Additionally, increasing the value of k in Property 4 exacerbates the degree of dimensional collapse. To illustrate, consider a dataset $\mathcal{D}$ sampled from a multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_m/m)$, exhibiting a collapse degree close to 0%. However, upon inserting m-dimensional zero-value vectors to each instance of $\mathcal{D}$, denoted as $\mathcal{D} \oplus \mathbf{0}^m$, half of the dimensions collapse. Consequently, the collapse degree increases to 50%. Figure 13 visually represents the collapse of $\mathcal{D} \oplus \mathbf{0}^k$ using the singular value spectra of the representations. It is evident that a larger k results in a more pronounced dimensional collapse. In summary, Property 4 corresponds to dimensional collapse.

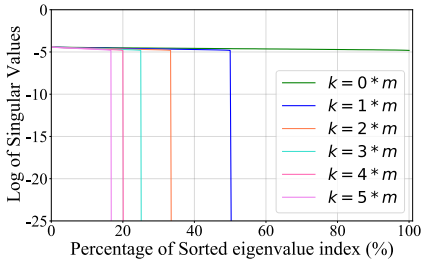
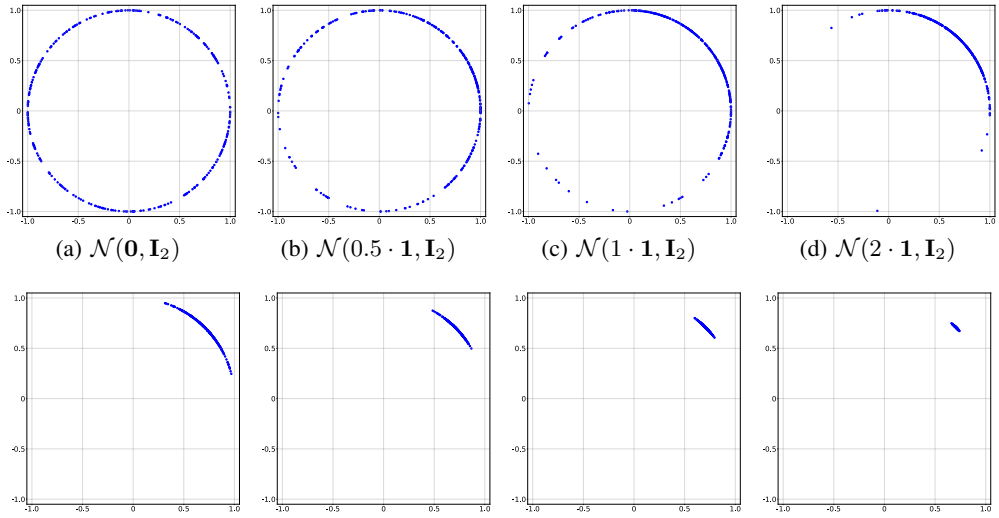


Figure 13: Singular value spectrum of $\mathcal{D} \oplus \mathbf{0}^k$

E.4 Understanding $\mathcal{W}_2$: Large means may lead to collapse

In this section, we explore our uniformity loss $\mathcal{W}_2$. This loss embodies two primary constraints. Firstly, it promotes the covariance matrix to be isotropic (specifically $\mathbf{I}_m/m$). Secondly, it enforces the mean to be zero. The latter constraint on the mean is crucial. To illustrate, we present a case study demonstrating that deviating the mean from zero compromises uniformity, even if the covariance matrix is precisely $\mathbf{I}_m/m$ and thus isotropic. Means deviating from zero may result in dimensional collapse and even constant collapse.

[4]Intuitively, maximal uniformity should stay constant regardless of dimensions; otherwise the corresponding uniformity metric exhibit a preference for larger or smaller dimensions.



(a) $\mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ (b) $\mathcal{N}(0.5 \cdot \mathbf{1}, \mathbf{I}_2)$ (c) $\mathcal{N}(1 \cdot \mathbf{1}, \mathbf{I}_2)$ (d) $\mathcal{N}(2 \cdot \mathbf{1}, \mathbf{I}_2)$

(e) $\mathcal{N}(4 \cdot \mathbf{1}, \mathbf{I}_2)$ (f) $\mathcal{N}(8 \cdot \mathbf{1}, \mathbf{I}_2)$ (g) $\mathcal{N}(16 \cdot \mathbf{1}, \mathbf{I}_2)$ (h) $\mathcal{N}(32 \cdot \mathbf{1}, \mathbf{I}_2)$

Figure 14: Visualizing ℓ_2 normalized Gaussian vectors with different means.

Table 3: Parameter settings for various models in the 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

Assuming $\mathbf{X} \in \mathbb{R}^2$ follows a Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_2)$, let $\mathbf{Y} = \mathbf{X} + k \cdot \mathbf{1}$ such that $\mathbf{Y} \sim \mathcal{N}(k \cdot \mathbf{1}, \mathbf{I}_2)$, where $\mathbf{1} \in \mathbb{R}^k$ represents a vector of all ones. We vary k from 0 to 32 and visualize the ℓ_2-normalized $\mathbf{Y}$'s in Figure 14 (by generating multiple independent copies). It is clear that an excessively large means will cause representations to collapse to a single point, even if the covariance matrix is isotropic.

F Experiment Settings and Convergence Analysis

F.1 Experiment Settings

To ensure fair comparisons, all experiments in Section 6 are conducted on a single 1080 GPU. Additionally, we maintain consistency in network architecture across all models, utilizing ResNet-18 (He et al., 2016) as the backbone and a three-layer MLP as the projector. The LARS optimizer (You et al., 2017) is employed with a base learning rate of 0.2, accompanied by a cosine decay learning rate schedule (Loshchilov & Hutter, 2017) for all models. Evaluation follows a linear evaluation protocol, where models are pre-trained for 500 epochs. Evaluation involves adding a linear classifier and training the classifier for 100 epochs while preserving the learned representations. The same augmentation strategy is deployed across all models, encompassing various operations such as color distortion, rotation, and cutout. Following da Costa et al. (2022), we set the temperature $t = 0.2$ for all contrastive learning methods. For MoCo (He et al., 2020) and NNCLR (Dwibedi et al., 2021), which require an additional queue to store negative samples, we set the queue size to 2^{12}. Regarding the linear decay for weighting the quadratic Wasserstein distance, refer to Table 3 for the parameter settings.

F.2 Convergence Analysis for Top-1 Accuracy

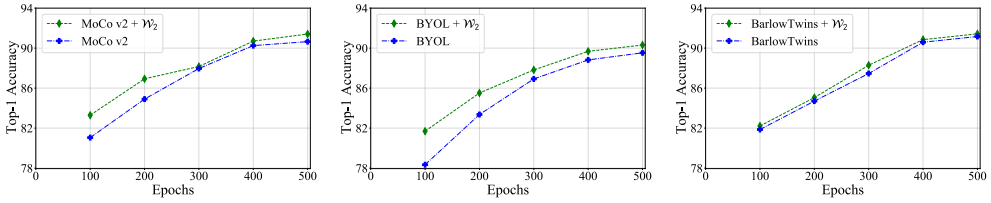
Here we illustrate the convergence of Top-1 accuracy across all training epochs in Fig 15. Throughout the training, we capture the model checkpoint at the end of each epoch to train a linear classifier. We subsequently evaluate the Top-1 accuracy on unseen images from the test set (either CIFAR-10 or CIFAR-100).

For both CIFAR-10 and CIFAR-100, we observe that integrating the proposed uniformity metric as an auxiliary loss significantly enhances the Top-1 accuracy, particularly in the initial stages of training.

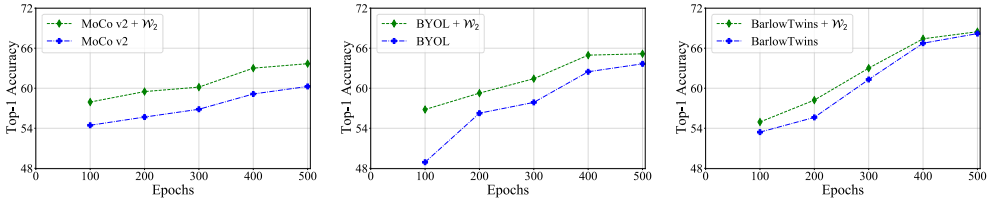
F.3 Convergence analysis for uniformity and alignment

This section presents the convergence of the uniformity metric and alignment loss across all training epochs in Figure 16 and Figure 17, respectively. Throughout the training, we record the model checkpoint at the end of each epoch to evaluate the uniformity using the proposed metric $\mathcal{W}_2$ and alignment (Wang & Isola, 2020) on unseen images from the test set (either CIFAR-10 or CIFAR-100).

For both CIFAR-10 and CIFAR-100, we observe that integrating the proposed uniformity metric as an auxiliary loss significantly improves uniformity. However, it also slightly compromises alignment (where a smaller alignment loss indicates better alignment). It should be noted that improved uniformity often leads to worse alignment.

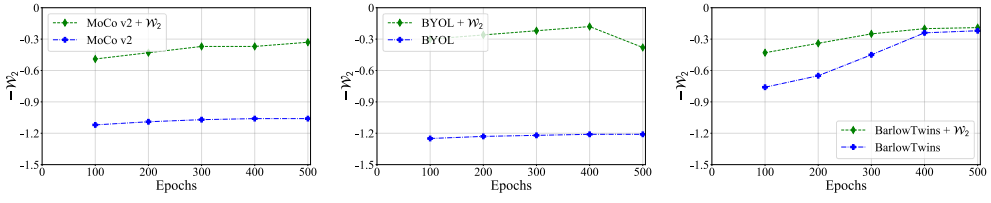


(a) MoCo v2 on CIFAR-10 (b) BYOL on CIFAR-10 (c) BarlowTwins on CIFAR-10

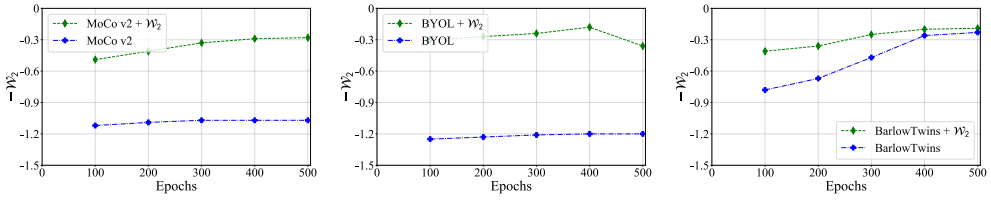


(d) MoCo v2 on CIFAR-100 (e) BYOL on CIFAR-100 (f) BarlowTwins on CIFAR-100

Figure 15: Convergence analysis for Top-1 accuracy during training.

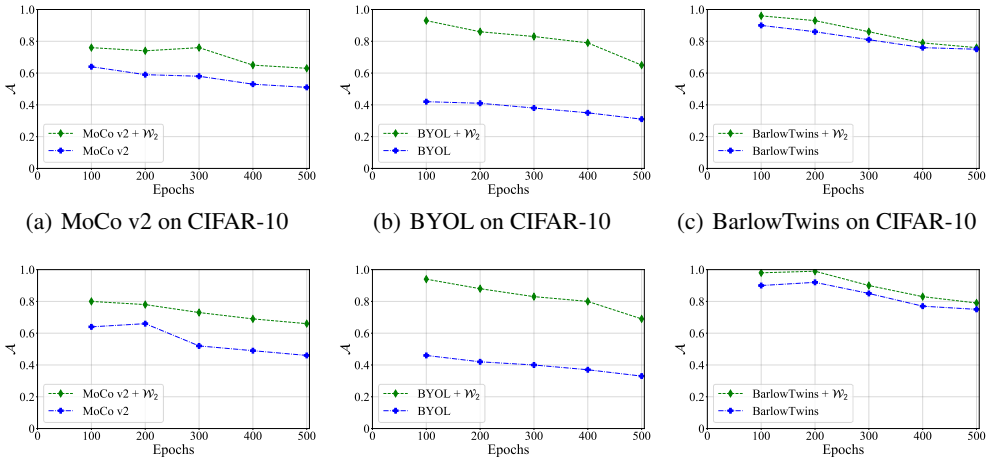


(a) MoCo v2 on CIFAR-10 (b) BYOL on CIFAR-10 (c) BarlowTwins on CIFAR-10



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

Figure 16: Visualizing 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 17: Visualizing alignment during training.