# AggNCE: Asymptotically Identifiable Contrastive Learning

**Jingyi Cui**[1*] **Weiran Huang**[2*] **Yifei Wang**[3] **Yisen Wang**[1,4†]
[1] Key Lab. of Machine Perception (MoE)
School of Intelligence Science and Technology, Peking University
[2] Huawei Noah's Ark Lab
[3] School of Mathematical Sciences, Peking University
[4] Institute for Artificial Intelligence, Peking University

## Abstract

Self-supervised contrastive learning has attracted great attention for its promising empirical performances. However, the identifiability theory of contrastive learning is still under-exploited. In this paper, we propose the *Aggregated InfoNCE (AggNCE)* loss function to deal with this problem. The latent variable learned by AggNCE is proved to be asymptotically identifiable to the ground truth latent under the original data distribution. In experiments, we verify the theoretical findings and show the empirical advantage of AggNCE over InfoNCE.

## 1 Introduction

With great data efficiency and generalization ability, self-supervised contrastive learning has attracted great attention for its promising empirical performances without relying on human annotations. The common approach of self-supervised contrastive learning in computer vision tasks is to treat each image as a separate class, and to learn the embedding through aligning different *views* of an image. The aligned views are called positive samples, which are usually constructed through hand-crafted data augmentations and are semantically similar to the original image. However, the embeddings learned through this approach are not guaranteed to reveal the underlying latents from which the images are generated.

In this paper, we investigate the identifiability theory of contrastive learning, an important property that guarantees the learned representation to reveal the underlying data generating process. By formulating the data generating process as a latent generative model, we prove that the mean latent of the multiple views learned by the InfoNCE loss is identifiable up to an invertible matrix, inspired by which, we propose the *Aggregated InfoNCE (AggNCE)* loss function. In experiments, we verify the performance of AggNCE and show that AggNCE empirically outperforms InfoNCE.

## 2 Preliminaries

**Notations.** Throughout this paper, we denote $x \in \mathbb{R}^d$ as an instance sample, i.e. images or views, and denote $z \in \mathbb{R}^n$ as the latent variable of $x$, where $d \in \mathbb{N}^+$ and $n \in \mathbb{N}^+$ are dimensions of the input data and latent respectively. The generating function of $x$ from $z$ is denoted as $g : \mathbb{R}^n \to \mathbb{R}^d$. We let $s : \mathbb{R}^n \to \mathbb{R}$ be a similarity measure. For the sake of simplicity, we let $s$ be the cosine similarity, i.e. $s(z, z') = z^\top z'$ for $z, z' \in \mathbb{R}^n$, where $z^\top$ stands for the transpose of $z$. Moreover, we denote $f : \mathbb{R}^d \to \mathbb{R}^n$ as the encoding function.

[*] Equal Contribution.
[†] Corresponding author: Yisen Wang (yisen.wang@pku.edu.cn).

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

**Unsupervised contrastive representation learning.** We consider a prevailing form of contrastive loss called InfoNCE loss function [10], with the form

$$\mathcal{L}_{con} = \mathbb{E}_{x,x' \sim \mathrm{P}_{pos},\, x_i^- \sim \mathrm{P}_{data}} - \log \frac{e^{f(x)^\top f(x')/\tau}}{e^{f(x)^\top f(x')/\tau} + \sum_{i=1}^{M} e^{f(x)^\top f(x_i^-)/\tau}}, \tag{1}$$

where $\tau > 0$ is a temperature parameter, $x$ and $x'$ stands for positive samples and $\{x_i^-\}_{i=1}^M$ are negative samples. This loss has been empirically shown to be effective by numerous works [2, 6]. [15] proved that the InfoNCE loss function optimizes the alignment of features from positive pairs and uniformity of the induced distribution of the (normalized) features on the hypersphere. In this paper, we focus on the uniformity of InfoNCE, and consider the asymptotic form proved in [15], i.e.

$$\lim_{M\to\infty} \mathcal{L}_{con} = \log M - \frac{1}{\tau}\mathbb{E}_{x,x' \sim \mathrm{P}_{pos}} f(x)^\top f(x') + \mathbb{E}_{x \sim \mathrm{P}_{data}} \log \mathbb{E}_{x^- \sim \mathrm{P}_{data}} e^{f(x)^\top f(x_i^-)/\tau}. \tag{2}$$

**Identifiability theory.** Suppose that a random variable $X$ is distributed according to the probability $\mathrm{P}_\theta$, where $\theta \in \Theta$ is the parameter of the generating probability. Then we say that $\theta$ is *identifiable* on the basis of $x$ if $\mathrm{P}_\theta$ is an injective function of $\theta$, i.e. $\mathrm{P}_\theta = \mathrm{P}_{\theta'}$ implies $\theta = \theta'$ [9]. In other words, identifiability guarantees that the generating process is unique based on observations $X$. Following [7], we define the identifiability up to an equivalence relation on $\Theta$ as follows.

**Definition 2.1** (Identifiability)**.** Let $\sim$ be an equivalence relation on $\Theta$. We say that a deep latent variable model $p_\theta(x, z)$ is identifiable up to $\sim$ if for all $\theta, \tilde{\theta} \in \Theta$, we have

$$p_\theta(x) = p_{\tilde{\theta}}(x) \Rightarrow \tilde{\theta} \sim \theta. \tag{3}$$

[1] shows that InfoNCE loss is an upper bound of the negative marginal probability of $x$ and $x'$, i.e.

$$-\log \mathrm{P}(x, x') \le \mathcal{L}_{con}(x, x') = -\mathbb{E}_{\mathrm{Q}(z,z')}\big[\log s(z, z')\big] + \mathbb{E}_{\mathrm{Q}(z)}\big[\log \mathbb{E}_{\mathrm{Q}(z')}[s(z, z')]\big]$$

That is, the optimal latents $(z^*, z'^*) = \arg\min_{z,z' \in \mathbb{R}^n} \mathcal{L}_{con}(x, x')$ guarantee the marginal probability $\mathrm{P}(x^*, x'^*)$ to be large and thus nearly maximized. Therefore, to study the identifiability of $(z^*, z'^*)$ learned by contrastive learning, we only need to study the identifiability of optimal latents that minimizes the joint marginal probability of the views, i.e. $\mathrm{P}(x^*, x'^*)$.

# 3 AggNCE: Asymptotically Identifiable Contrastive Learning

In this section, we assume that there are multiple augmented views $x'_1, \ldots, x'_B$, and the corresponding latents are $z'_1, \ldots, z'_B$.

## 3.1 Generating Process of Multiple Views

We assume the joint probability of $x$, $x'_1, \ldots, x'_B$, $z$, and $z'_1, \ldots, z'_B$ satisfies

$$\mathrm{P}(x, x'_1, \ldots, x'_B, z, z'_1, \ldots, z'_B) = \mathrm{P}(x|z) \prod_{i=1}^{b} \mathrm{P}(x'_i|z'_i) \mathrm{P}(z, z'_1, \ldots, z'_B). \tag{4}$$

Eq. (4) indicates that the instance $x$ depends only on its own latent $z$, and is independent of the augmented $x'$ and its corresponding latent $z'$. Then inspired by [7], we formulate the joint distribution of $x$ and $z$ (and $x'_i$ and $z'_i$) as

$$p_\theta(x, z|u) = p_g(x|z) p_{T,\lambda}(z|u), \text{ and } p_\theta(x'_i, z'_i|u) = p_g(x'_i|z'_i) p_{T,\lambda}(z'_i|u),\ i = 1, \ldots, B \tag{5}$$

where $g : \mathbb{R}^n \to \mathbb{R}^d$ is a generative function such that $x = g(z) + \varepsilon$, $x'_i = g(z'_i) + \varepsilon$, $i = 1, \ldots, b$, and $\varepsilon$ is a mean-zero noise variable independent of $z$, and $p_{T,\lambda}$ denotes the exponential family distribution, $z'_i \sim \exp(T, \lambda(u))$, with sufficient statistics $T$.

**Assumption 3.1.** Assume that for $i \in \{1, \ldots, B\}$,

*(i)* $p(z'_i|z) = \frac{1}{C_z} \mathrm{Q}(z'_i) s(z, z'_i)$.

*(ii)* For all $i \ne j$, $z'_i \perp z'_j \mid z$.

The latent variable model is shown in Figure 1.

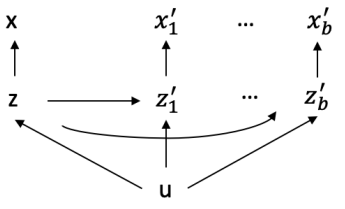


Figure 1: Latent variable model of multi-views.

### 3.2 Identifiability of Contrastive Learning with Multiple Views

To show the identifiability results, we make the same assumptions as in Theorem 1 of [7]. Then we have our main theorem.

**Theorem 3.2.** *Let Assumption 3.1 hold. Then if*

$$\mathrm{P}_{f,T,\lambda}(x'_1,\ldots,x'_B|u) = \mathrm{P}_{\tilde{f},\tilde{T},\tilde{\lambda}}(x'_1,\ldots,x'_B|u), \tag{6}$$

*we have*

$$\frac{1}{B}\sum_{i=1}^{B} T(f^{-1}(x'_i)) = A\frac{1}{B}\sum_{i=1}^{B}\tilde{T}(\tilde{f}^{-1}(x'_i)) + c, \tag{7}$$

*where $A$ is an invertible matrix and $c$ is a constant vector.*

Theorem 3.2 shows that the mean latent of the $B$ augmented views is identifiable up to an invertible matrix. Eq. (7) could be simply understood that the "mean" latent of views $z'_i$ are identifiable up to an invertible matrix, i.e.

$$\frac{1}{B}\sum_{i=1}^{B} z'_i = A\frac{1}{B}\sum_{i=1}^{B}\tilde{z}'_i + c. \tag{8}$$

In Lemma A.1, we show that $\mathbb{E}(z'|z) = z$. Note that the mean $\bar{z}'_B := \frac{1}{B}\sum_{i=1}^{B} z'_i$ is an unbiased estimator of $\mathbb{E}(z'|z)$. As $B \to \infty$, $z$ is asymptotically identifiable. Intuitively, it could be understood as $z'_i = z_0 + \epsilon_i$, where $z_0$ is the content factor sharing over $z'_i$, $i \in [B]$, and $\epsilon_i$ is the view-specific style or noise. By averaging over $z'_i$, we maintain the content only while removing the style or noise. By contrast, we also demonstrate the unidentifiability of contrastive learning with two augmented views in Appendix A.

Then we briefly discuss how many views $B$ are enough for the mean latent $\bar{z}'_B$ to have a good approximation to $z$. According to a simple analysis by Chebyshev's inequality, we have for $\varepsilon > 0$

$$\mathrm{P}\left(\left\|\frac{1}{B}\sum_{i=1}^{B} z'_i - z\right\|_2 \geq \varepsilon\right) \leq \frac{n\sigma_z^2}{B\varepsilon^2}, \tag{9}$$

where $\sigma_z^2$ denotes the variance of $z'_i$. In practice, the variance $\sigma_z^2$ can be easily bounded by some constant through normalization. Therefore, if we want to achieve an $\varepsilon$-robust estimation of $z$ with a constant probability, we require the number of views $B$ to be $O(1/\varepsilon^2)$.

### 3.3 Aggregated InfoNCE (AggNCE) Loss Function

Motivated by Theorem 3.2 , we propose the Aggregated InfoNCE (AggNCE) loss function, where we replace the representation of one view with the aggregated representation of multiple augmented positive samples. The mathematical form is shown as follows.

$$L^B(f) = -\log\frac{\exp[f(x)^\top \bar{f}(x^+)]}{\exp f(x)^\top \bar{f}(x^+) + \sum_{i=1}^{M}\exp[f(x)^\top f(x_i^-)/\tau]}, \tag{10}$$

where $\bar{f}(x^+) := \frac{1}{B}\sum_{i=1}^{B} f(x_i^+)/\tau$.

In the AggNCE loss function, we use the mean representation to learn the mean latent $\bar{z}'$, which is proved to be asymptotically identifiable to the ground true latent $z$ of the original input data. That is, the distribution gap between $z'$ and $z$ can be alleviated by using AggNCE as the number of views $B \to \infty$. Note that when $B = 1$, our AggNCE loss function degenerates to the standard InfoNCE.

## 4 Comparisons with Related Works

**Comparisons with multi-view contrastive learning methods.** The most similar form of AggNCE is to use blocks of similar samples [12]. However, the essence of the block loss is different, since the way it aggregates the negative samples cannot be understood as an optimization over uniformity as is discussed in [15], even if we transfer the idea of "block" to the InfoNCE loss function. Some other methods of using multiple views are to aggregate the loss function induced by different views instead of aggregating the representations, e.g. [14, 11]. Thus, the learned representations through these methods fail to be identifiable to the latent of the original input data.

**Comparisons with identifiability results.** [16] first shows the link between contrastive learning and nonlinear ICA, and proves that the learned representation of contrastive learning recovers the source latent up to an orthogonal linear transformation. However, [16] assumes that the ground-truth marginal distribution of the latents is uniform, which is only true when perfect uniformity is achieved. [8] assumes a partition of the latent space into an invariant content block and a view-specific style block, and proves the block identifiability of contrastive learning. The intuition behind this result is that the learned representations capture the content shared across all augmented views, which turns out to be similar to ours in Theorem 3.2. However, it analyzes a contrastive loss function that aligns the exact content block (with the dimension unknown in real applications), whereas our proposed AggNCE loss function can be easily implemented for practical use by using multiple augmented views. [5] studies a semiparametric model for learning from pairwise measurements, and proves convergence of the estimated parameters. The modeling of data generating process is inherently different from ours.

# 5 Experiments

**Setup.** We follow the setting of SimCLR [2]. We use the SGD optimizer and use ResNet-50 as the encoder and a 2-layer MLP as the projection head. We run experiments on 4 NVIDIA Tesla v100 32GB GPUs. The data augmentations we use are random crop and resize (with random flip), color distortion, and color dropping. The models are trained using both InfoNCE and AggNCE loss with batch size 128 and 500 epochs for each model. We evaluate the self-supervised learned representation by linear evaluation protocol, where a linear classifier is trained on the top of the encoder, and regard its test accuracy as the performance of the encoder.

## 5.1 Performance Comparison

We first show the advantage of AggNCE over InfoNCE through linear evaluations, where we fix the batch size of both methods to be 128. In AggNCE, we vary the number of views $B$ from 2 to 20 to study the effect of the number of views for AggNCE. Note that InfoNCE can be viewed as a special case of AggNCE with $B = 1$. From Table 1 on CIFAR-10 and CIFAR-100 datasets, we observe that the performance gets better with more augmented views, and gets saturated when $B \geq 10$. In Table 2, AggNCE outperforms InfoNCE by $2.11\%$ when $B = 3$ on the ImageNet-100 dataset.

Table 1: Performance comparison of InfoNCE and AggNCE with various numbers of views.

| Method | $B$ | CIFAR-10 | | CIFAR-100 | |
|---|---|---|---|---|---|
| | | acc | advantage | acc | advantage |
| InfoNCE | 1 | 91.47 | / | 82.50 | / |
| AggNCE | 2 | 92.37 | $+0.90\%$ | 84.77 | $+2.27\%$ |
| | 3 | 92.78 | $+1.31\%$ | 84.90 | $+2.40\%$ |
| | 5 | 93.10 | $+1.63\%$ | 85.37 | $+2.87\%$ |
| | 10 | **93.39** | $\mathbf{+1.92\%}$ | 85.70 | $+3.20\%$ |
| | 20 | 93.32 | $+1.85\%$ | **85.89** | $\mathbf{+3.39\%}$ |

* The best results are marked in **bold**.

Table 2: Performance comparisons on ImageNet-100.

| Method | $B$ | batch size | acc(%) | advantage |
|---|---|---|---|---|
| InfoNCE | 1 | 256 | $69.40(0.17)$ | / |
| AggNCE | 3 | 256 | $\mathbf{71.51(0.12)}$ | $+2.11\%$ |

* The best results are marked in **bold**.

For further fair comparisons, we adjust the batch size of InfoNCE so that the same number of views are used in InfoNCE and AggNCE. The results are shown in Appendix B.

# 6 Conclusion

In this paper, we proposed the AggNCE loss function, which aligns the representation of one view with the mean representation of multiple views. Through theoretical discussions, we show that the representations learned by AggNCE can be asymptotically identifiable. Furthermore, we also demonstrate the advantage of AggNCE over InfoNCE in linear evaluation through numerical comparisons.

## Acknowledgment

Yisen Wang is partially supported by the NSF China (No. 62006153), Project 2020BD006 supported by PKU-Baidu Fund, and Huawei Technologies Inc.

## References

[1] Laurence Aitchison. Infonce is a variational autoencoder. *arXiv preprint arXiv:2107.02495*, 2021.

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

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[4] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

[5] Yi Chen, Zhuoran Yang, Yuchen Xie, and Zhaoran Wang. Contrastive learning from pairwise measurements. *Advances in Neural Information Processing Systems*, 31, 2018.

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

[7] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *AISTATS*. PMLR, 2020.

[8] Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *NeurIPS*, 2021.

[9] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[10] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[11] Shuvendu Roy and Ali Etemad. Self-supervised contrastive learning of multi-view facial expressions. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021.

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

[13] Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18, 2017.

[14] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*. Springer, 2020.

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

[16] Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. *arXiv preprint arXiv:2102.08850*, 2021.

# A Unidentifiability of InfoNCE

In this appendix, we first formulate the generating process of the input data and its augmentations as a latent generative model. Then we conduct an identifiability analysis of latents (representations) learned by contrastive learning. Finally, we discuss that the identifiable mean latent does not coincide with the ground truth latent of the original input data, that is, the representations learned by ordinary contrastive learning fail to be identifiable.

## A.1 Data Generating Process: Latent Variable Model

Let $x' \in \mathbb{R}^d$ be the augmented version of $x \in \mathbb{R}^d$. Assume that $z \in \mathbb{R}^n$ and $z' \in \mathbb{R}^n$ is the latent variables of $x$ and $x'$ respectively (usually $n < d$), and that $u$ is an observable prior (instance index). The latent variable model is illustrated in Figure 2.

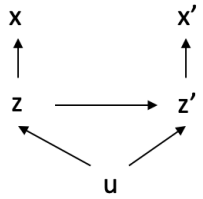


Figure 2: Latent variable model of two views.

We assume the joint probability of $x$, $x'$, $z$, and $z'$ satisfies

$$\mathrm{P}(x, x', z, z') = \mathrm{P}(x|z)\mathrm{P}(x'|z')\mathrm{P}(z, z'). \tag{11}$$

This assumption is natural and straightforward. (11) shows that the instance $x$ depends only on its own latent $z$, and is independent of the augmented $x'$ and its corresponding latent $z'$.

Then inspired by [7], we formulate the joint distribution of $x$ and $z$ (and $x'$ and $z'$) as

$$\begin{aligned} p_\theta(x, z|u) &= p_g(x|z)p_{T,\lambda}(z|u), \\ p_\theta(x', z'|u) &= p_g(x'|z')p_{T,\lambda}(z'|u), \end{aligned} \tag{12}$$

where the definition of $p_g$ and $p_{T,\lambda}$ will be delayed to (13) and (18).

**Generative function:** We naturally assume that $x$ and $x'$ are generated by different latent variables following the same ground-truth generative function. To be specific, we assume that there exists an underlying generative function $g : \mathbb{R}^n \to \mathbb{R}^d$ such that

$$x = g(z) + \varepsilon \text{ and } x' = g(x') + \varepsilon, \tag{13}$$

where $\varepsilon$ is a mean-zero noise variable independent of $z$.

**Latent variables:** We assume that the latent variable of the augmented view $x'$ lies around that of the instance $x$. Intuitively, we would like $x$ and $x'$ to have similar semantic representations. Therefore, we reasonably assume that the conditional distribution of $z'$ on $z$ is

$$p(z'|z) = \frac{1}{C_z}\mathrm{Q}(z')s(z, z'), \tag{14}$$

where $s : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a similarity measure of $z$ and $z'$, and

$$C_z := \int \mathrm{Q}(z')s(z, z')\, dz' \tag{15}$$

is the normalizing constant. By (14), given $z$, $z'$ has larger conditional probability if $z'$ is more similar to $z$ (larger $s(z, z')$). We could without loss of generality assume that the components of $z'$ have the same expectation, i.e. $\mathbb{E}z' = (1, \ldots, 1)^\top$, and that the sum of all components of $z$ is 1, i.e. $(1, \ldots, 1)z = 1$. Note that these assumptions are quite mild since they can be easily achieved by normalizing $z$ and $z'$. We propose them only for the simplicity of the proofs.

**Lemma A.1.** *If $z$ is an isotropic random vector,* (14) *yields that*

$$\mathbb{E}(z'|z) = z. \tag{16}$$

Lemma A.1 shows that the conditional expectation of the augmented latent $z'$ lies exactly at $z$. Intuitively, if there are multiple augmented samples $\{z_i\}_{i=1}^B$, we could understand that the augmented latents $z_i'$ are centered at $z$.

Following [7], we assume that $z$ and $z'$ are drawn from the exponential family distribution, i.e.

$$z, z' \sim \exp(T, \lambda(u)), \tag{17}$$

where $T$ denotes the sufficient statistics and $u$ is an observable prior. Since exponential families have universal approximation capabilities, so this assumption is not very restrictive [13]. To be specific, we assume that the exponents of $z$ (and $z'$) are independent, i.e. under mathematical formulations

$$p_{T,\lambda}(z \mid u) = \prod_{i=1}^{n} \frac{Q_i(z_i)}{Z_i(u)} \exp\left[\sum_{j=1}^{k} T_{i,j}(z_i)\lambda_{i,j}(u)\right], \tag{18}$$

where $Q_i$ is the base measure, $Z_i(u)$ is the normalizing constant, $T_i = (T_{i,j})_{j=1}^k$ are sufficient statistics, and $\lambda_i(u) = (\lambda_{i,j}(u))$ are the corresponding parameters.

### A.2 Identifiability Analysis

In this part, based on the theoretical link between InfoNCE loss and the latent generative model (Proposition A.2), we show the identifiability analysis of representations learned with InfoNCE.

**Proposition A.2** ([1])**.** *Under the latent generative model* (11) *and* (14)*, there holds*

$$-\log \mathrm{P}(x, x') \le \mathcal{L}_{con}(x, x') = -\mathbb{E}_{\mathrm{Q}(z,z')}\big[\log s(z, z')\big] + \mathbb{E}_{\mathrm{Q}(z)}\big[\log \mathbb{E}_{\mathrm{Q}(z')}[s(z, z')]\big]$$

Proposition A.2 shows that the InfoNCE loss is an upper bound of the negative marginal probability of $x$ and $x'$. In contrastive representation learning, we are optimizing the InfoNCE loss function over latents $z$ and $z'$. Then according to Proposition A.2, the optimal latents

$$(z^*, z'^*) = \operatorname*{arg\,min}_{z,z' \in \mathbb{R}^n} \mathcal{L}_{con}(x, x') \tag{19}$$

also guarantee the marginal probability $\mathrm{P}(x^*, x'^*)$ to be large and thus nearly maximized, where $(x^*, x'^*) := \big(g(z^*) + \varepsilon, g(z'^*) + \varepsilon\big)$. Therefore, to study the identifiability of $(z^*, z'^*)$ learned by optimizing the InfoNCE loss, we only need to study the identifiability of optimal latents that minimizes the joint marginal probability of the views, i.e. $\mathrm{P}(x^*, x'^*)$. To be specific, by the definition of identifiability, we need to prove the following claim:

*Let $\sim$ be an equivalence relation on $\Theta$. If for all $\theta, \tilde{\theta} \in \Theta$, then there holds*

$$\mathrm{P}_\theta(x, x') = \mathrm{P}_{\tilde{\theta}}(x, x') \Rightarrow \tilde{\theta} \sim \theta. \tag{20}$$

If (20) holds, the optimal latents $(z^*, z'^*)$ learned by contrastive representation learning are uniquely identified with respect to $\Theta$.

We consider the case where two augmented views are used as positive samples, which is commonly adopted in real-world applications [6, 2, 3, 4].

To show the identifiability results, we make the same assumptions as in Theorem 1 of [7].

**Assumption A.3.** Assume the following holds

- *(i)* The set $\{x \in \mathbb{R}^d \mid \phi_\varepsilon(x) = 0\}$ has measure zero, where $\phi_\varepsilon$ is the characteristic function of the density $p_\varepsilon(x - g(z)) := p_g(x|z)$.
- *(ii)* The generative function $g$ is injective.
- *(iii)* The sufficient statistics $T_{i,j}$ are differentiable almost everywhere, and $(T_{i,j})_{1 \le j \le k}$ are linearly independent on any subset of $\mathbb{R}^d$ of measure greater than zero.
- *(iv)* There exist $nk + 1$ distinct points $u^0, \ldots, u^{nk}$ such that the matrix

$$L = (\lambda(u_1) - \lambda(u_0), \ldots, \lambda(u_{nk}) - \lambda(u_0)) \tag{21}$$

of size $nk \times nk$ is invertible.

The generating process is shown in Assumption A.4.

**Assumption A.4.** Assume that for $i \in \{1, 2\}$,

*(i)* The latent variables $z_i'$ follows the exponential distribution, i.e.

$$z_i' \sim \exp(T, \lambda(u)). \tag{22}$$

*(ii)*

$$p(z_i'|z) = \frac{1}{C_z} \mathrm{Q}(z_i') s(z, z_i'). \tag{23}$$

*(iii)* For all $i \neq j$,

$$z_i' \perp z_j' \,|\, z. \tag{24}$$

We see that *(i)* and *(ii)* directly extend the generating process in (14) and (17) to the case of using two augmented vies. *(iii)* is also an intuitive assumption, since the augmented views are independently generated based on the original input data. By the following lemma, we show that the generating process of $x_1'$ and $x_2'$ is essentially the same as that of $x$ and $x'$.

The latent variable model is illustrated in Figure 1.

**Lemma A.5.** *Let Assumption 3.1 (ii) and (iii) hold. Assume that $z$ is an isotropic random vector. Then for all $i \neq j$, there holds*

$$p(z_i', z_j') = \mathrm{Q}(z_i')\mathrm{Q}(z_j') s(z_i', z_j'). \tag{25}$$

Note that the mathematical form of (25) is similar to (14) by replacing $z$ with $z_i'$. In other words, the probabilistic relation between two augmented latents $z_i'$ and $z_j'$ is the same as that between the latent $z$ of the input data and the augmented latent $z'$.

Next, we consider the case where $B = 2$, and prove the identifiability result of InfoNCE in the following theorem.

**Theorem A.6.** *Let Assumptions A.3 and 3.1 hold. Then if*

$$\mathrm{P}_{f,T,\lambda}(x_i', x_j'|u) = \mathrm{P}_{\tilde{f},\tilde{T},\tilde{\lambda}}(x_i', x_j'|u), \tag{26}$$

*we have*

$$T(f^{-1}(x_i')) + T(f^{-1}(x_j')) = A\Big(\tilde{T}(\tilde{f}^{-1}(x_i')) + \tilde{T}(\tilde{f}^{-1}(x_j'))\Big) + c, \tag{27}$$

*where $A$ is an invertible matrix and $c$ is a constant vector.*

Theorem A.6 shows that the mean latent of two augmented views $(z_i' + z_j')/2$ is identifiable up to an invertible matrix. Note that in many examples of the exponential distribution family, there holds $T(z) = z$, e.g. Gaussian distribution, Bernoulli distribution, Poisson distribution, etc. In these cases, (27) could be understood as

$$z_i' + z_j' = A(\tilde{z}_i' + \tilde{z}_j') + c, \tag{28}$$

where $\tilde{z_i'}$ and $\tilde{z_j'}$ are generated following the latent variable model introduced in Section A.1 under parameter $\tilde{\theta} \in \Theta$.

### A.3 Discussions

From Theorem A.6, we show that the ientifiability of the mean latent $\bar{z}' = (z_1' + z_2')/2$ learned by optimizing the InfoNCE loss that aligns two augmented views. However, notice that the distribution of the augmented latents $z_1'$ and $z_2'$ is similar to but does not coincide with that of the latent $z$ of the original input data. Although Lemma A.1 indicates that $\mathbb{E}(\bar{z}'|z) = z$, there still remains a significant distribution gap between $\bar{z}'$ and $z$.

On the other hand, recall that the latent of an augmented sample $z'$ is assumed to be distributed around the underlying latent of the original input data $z$. Thus, an intuitive way to narrow the distribution gap is to increase the number of views, so that the mean latent of multiple views can be a more accurate estimate of the ground truth latent $z$ of the original input data $x$. Therefore, in the next section, we extend the identifiability result of contrastive learning to a multi-view scenario.

# B Additional Experiments

## B.1 Performance Comparison with Adjusted Batch Size

The comparisons in Table 1 may seem unfair to InfoNCE, since for a batch of training data, InfoNCE accesses $2 \times$ bsz augmented views, whereas AggNCE can access $(B+1) \times$ bsz augmented views. Thus, to make the comparison more fair, we fix the batch size of 128 for AggNCE, and set the batch size of $128 \times (B+1)/2$ for InfoNCE to allow them to access the same number of augmented views. The results are shown in Table 3.

Table 3: Performance comparisons with fixed batch size 128 for AggNCE.

| Dataset | InfoNCE | | AggNCE | | Advantage |
|---|---|---|---|---|---|
| | bsz | acc | $B$ | acc | |
| CIFAR-10 | 192 | 92.50 | 2 | **92.58** | +0.08% |
| | 256 | 92.48 | 3 | **92.78** | +0.30% |
| | 384 | 92.45 | 5 | **93.10** | +0.65% |
| | 1024 | 92.23 | 15 | **93.23** | +1.00% |
| | 1344 | 92.01 | 20 | **93.32** | +1.31% |
| CIFAR-100 | 192 | 84.57 | 2 | **84.77** | +0.20% |
| | 256 | 84.74 | 3 | **84.90** | +0.16% |
| | 384 | 84.96 | 5 | **85.37** | +0.41% |
| | 704 | 83.26 | 10 | **85.70** | +2.44% |
| | 1024 | 83.30 | 15 | **85.93** | +2.63% |

* The best results are marked in **bold**.

In Table 3, we realize a fair comparison between AggNCE and InfoNCE by using the same amount of information in each batch. We show that the AggNCE loss consistently outperforms InfoNCE in the linear evaluation with $B$ varying from 2 to 20. Moreover, we observe an increasing advantage in the performance of AggNCE over InfoNCE as the number of views $B$ increases. Therefore, we could reasonably speculate that by using AggNCE, we asymptotically recover the true underlying latents as $B \to \infty$, which results in the advantage of AggNCE over the ordinary InfoNCE loss function in empirical performance.

We run additional experiments on the ImageNet-100 dataset [14], a random 100-class subset of ImageNet. We follow the details in [2]: crop size from $0.08$ to $1.0$, stronger jittering $(0.8, 0.8, 0.8, 0.2)$, grayscaling probability $0.2$, and Gaussian blurring with $0.5$ probability. We train with the backbone of ResNet-50 for 240 epochs with learning rate $2 \times 10^{-3}$. We report the mean top-1 accuracy of the last five epochs of each method. The corresponding standard deviation is also reported in the parenthesis.

Table 4: Performance comparisons on ImageNet-100.

| Method | $B$ | batch size | acc(%) |
|---|---|---|---|
| InfoNCE | 1 | 256 | 69.40(0.17) |
| | 1 | 768 | 70.28(0.08) |
| AggNCE | 3 | 256 | **71.51(0.12)** |

* The best results are marked in **bold**.

We show that in Table 4, under the same batch size, AggNCE outperforms InfoNCE by $2.11\%$ when $B = 3$ on the ImageNet-100 dataset. We further realize a fair comparison between AggNCE and InfoNCE by using the same amount of information in each batch, where we show that AggNCE still outperforms InfoNCE by $1.23\%$. This promising performance of AggNCE results from the asymptotic identifiability of its underlying data generation process.

# C Proofs

*Proof of Lemma A.1.* By definition, we have

$$\mathbb{E}(z'|z) = \int_{z'} z' p(z'|z)\, dz'$$

$$
\begin{aligned}
&= \int_{z'} z' \frac{1}{C_z} \mathrm{Q}(z') s(z, z') dz' \\
&= \frac{1}{C_z} \int_{z'} z' \mathrm{Q}(z') z^\top z' dz' \\
&= \frac{1}{C_z} \int_{z'} z' \mathrm{Q}(z') z'^\top dz' \cdot z \\
&= \frac{1}{C_z} \int_{z'} \mathrm{Q}(z') z' z'^\top dz' \cdot z \\
&= \frac{1}{C_z} \mathbb{E} z' z'^\top \cdot z \\
&= \frac{1}{C_z} \mathbb{E} z' z'^\top \cdot z \\
&= \frac{1}{C_z} I_n z \\
&= \frac{1}{C_z} z, \qquad (29)
\end{aligned}
$$

where

$$
\begin{aligned}
C_z &= \int \mathrm{Q}(z') s(z, z') \, dz' \\
&= \int \mathrm{Q}(z') z'^\top z \, dz' \\
&= \int \mathrm{Q}(z') z'^\top \, dz' \cdot z \\
&= (\mathbb{E} z')^\top z. \qquad (30)
\end{aligned}
$$

By assumption, $Ez' = (1, \ldots, 1)$ and $\sum_{i=1}^{d} z_i = 1$. Therefore, we have $C_z = (\mathbb{E} z')^\top z = 1$, and thus $\mathbb{E}(z'|z) = z$. □

*Proof of Lemma A.5.* By Assumption 3.1 *(ii)* and *(iii)*, we have

$$
\begin{aligned}
p(z'_i, z'_j|z) &= p(z'_i|z) p(z'_j|z) \\
&= \mathrm{Q}(z'_i) \mathrm{Q}(z'_j) f(z, z'_i) f(z, z'_j). \qquad (31)
\end{aligned}
$$

Then there holds

$$
\begin{aligned}
p(z'_i, z'_j) &= \int_z p(z'_i, z'_j|z) \mathrm{Q}(z) \, dz \\
&= \int_z \mathrm{Q}(z'_i) \mathrm{Q}(z'_j) f(z, z'_i) f(z, z'_j) \mathrm{Q}(z) \, dz \\
&= \mathrm{Q}(z'_i) \mathrm{Q}(z'_j) \int_z \mathrm{Q}(z) f(z, z'_i) f(z, z'_j) \, dz. \\
&= \mathrm{Q}(z'_i) \mathrm{Q}(z'_j) \int_z \mathrm{Q}(z) z'^\top_i z z^\top z'_j \, dz \\
&= \mathrm{Q}(z'_i) \mathrm{Q}(z'_j) z'^\top_i \left( \int_z \mathrm{Q}(z) z z^\top \, dz \right) z'_j \\
&= \mathrm{Q}(z'_i) \mathrm{Q}(z'_j) z'^\top_i \mathbb{E} z z^\top z'_j \qquad (32)
\end{aligned}
$$

Since $z$ is isotropic, then we have

$$
\begin{aligned}
p(z'_i, z'_j) &= \mathrm{Q}(z'_i) \mathrm{Q}(z'_j) z'^\top_i I_n z'_j \\
&= \mathrm{Q}(z'_i) \mathrm{Q}(z'_j) z'^\top_i z'_j \\
&= \mathrm{Q}(z'_i) \mathrm{Q}(z'_j) f(z_i, z_j). \qquad (33)
\end{aligned}
$$

□

*Proof of Theorem A.6.*

$$\begin{aligned}
\mathrm{P}_\theta(x'_i, x'_j|u) &= \iint \mathrm{P}_\theta(x'_i, x'_j, z'_i, z'_j|u)\, dz'_i\, dz'_j \\
&= \iint p_g(x'_i|z'_i) p_g(x'_j|z'_j) \frac{1}{C_z} p_{T,\lambda}(z'_i|u) p_{T,\lambda}(z'_j|u) s(z'_i, z'_j)\, dz'_i\, dz'_j \\
&= \int p_g(x'_i|z'_i) p_{T,\lambda}(z'_i|u) \bigg( \int p_g(x'_j|z'_j) \frac{1}{C_z} p_{T,\lambda}(z'_j|u) s(z'_i, z'_j)\, dz'_j \bigg) dz'_i. \qquad (34)
\end{aligned}$$

Let $q_{T,\lambda,z'_i}(z'_j|u) = \frac{1}{C_z} p_{T,\lambda}(z'_j|u) s(z'_i, z'_j)$. Obviously, $q_{T,\lambda,z'_i}$ is a probability density function. Then

$$\begin{aligned}
\mathrm{P}_\theta(x'_i, x'_j|u) &= \int p_\varepsilon(x'_i - f(z'_i)) p_{T,\lambda}(z'_i|u) \bigg( \int p_\varepsilon(x'_j - f(z'_j)) q_{T,\lambda,z'_i}(z'_j|u)\, dz'_j \bigg) dz'_i \\
&= \int p_\varepsilon(x'_i - \bar{x'_i}) p_{T,\lambda}(f^{-1}(\bar{x'_i})|u) \mathrm{vol} J_{f^{-1}}(\bar{x'_i}) \\
&\quad \cdot \bigg( \int p_\varepsilon(x'_j - \bar{x'_j}) q_{T,\lambda,z'_i}(f^{-1}(\bar{x'_j})|u) \mathrm{vol} J_{f^{-1}}(\bar{x'_j})\, d\bar{x'_j} \bigg) d\bar{x'_i} \\
&= \int p_\varepsilon(x'_i - \bar{x'_i}) \tilde{p}_{T,\lambda,f,u}(\bar{x'_i}) \bigg( \int p_\varepsilon(x'_j - \bar{x'_j}) \tilde{p}_{T,\lambda,f,u}(\bar{x'_j}) f(g^{-1}(x'_i), g^{-1}(x'_j))\, d\bar{x'_j} \bigg) d\bar{x'_i} \\
&= \int p_\varepsilon(x'_i - \bar{x'_i}) \tilde{p}_{T,\lambda,f,u}(\bar{x'_i}) \bigg( (\tilde{q}_{T,\lambda,f,u,\bar{x'_i}} * p_\varepsilon)(\bar{x'_j}) \bigg) d\bar{x'_i} \\
&= \iint p_\varepsilon(x - \bar{x'_i}) p_\varepsilon(x'_j - \bar{x'_j}) \tilde{p}_{T,\lambda,g,u}(\bar{x'_i}) \tilde{p}_{T,\lambda,g,u}(\bar{x'_j}) f(g^{-1}(\bar{x'_i}), g^{-1}(\bar{x'_j}))\, d\bar{x'_j}\, d\bar{x'_i}.
\end{aligned} \qquad (35)$$

Since (35) holds for arbitrary $x'_i$ and $x'_j$. we have

$$\tilde{p}_{T,\lambda,f,u}(x'_i) \cdot \tilde{p}_{T,\lambda,f,u}(x'_j) = \tilde{p}_{\tilde{T},\tilde{\lambda},\tilde{f},u}(x'_i) \cdot \tilde{p}_{\tilde{T},\tilde{\lambda},\tilde{f},u}(x'_j) \qquad (36)$$

By definition and the assumption of exponential family,

$$\begin{aligned}
&\tilde{p}_{T,\lambda,f,u}(x'_i) \cdot \tilde{p}_{T,\lambda,f,u}(x'_j) \\
&= p_{T,\lambda}(f^{-1}(x'_i)|u) \mathrm{vol} J_{f^{-1}}(x'_i) \cdot p_{T,\lambda}(f^{-1}(x'_j)|u) \mathrm{vol} J_{f^{-1}}(x'_j) \\
&= \mathrm{vol} J_{f^{-1}}(x'_i) \cdot \prod_{\iota=1}^{n} \frac{Q_\iota(f_\iota^{-1}(x'_i))}{Z_\iota(u)} \exp\bigg\{ \sum_{\nu=1}^{k} T_{\iota,\nu}(f_\iota^{-1}(x'_i)) \lambda_{\iota,\nu}(u) \bigg\} \\
&\cdot \mathrm{vol} J_{f^{-1}}(x'_j) \cdot \prod_{\iota=1}^{n} \frac{Q_\iota(f_\iota^{-1}(x'_j))}{Z_\iota(u)} \exp\bigg\{ \sum_{\nu=1}^{k} T_{\iota,\nu}(f_\iota^{-1}(x'_j)) \lambda_{\iota,\nu}(u) \bigg\}. \qquad (37)
\end{aligned}$$

Take logarithm,

$$\begin{aligned}
&\log \tilde{p}_{T,\lambda,f,u}(x'_i) \cdot \tilde{p}_{T,\lambda,f,u}(x'_j) \\
&= \log \mathrm{vol} J_{f^{-1}}(x'_i) + \sum_{\iota=1}^{n} \log Q_\iota(f_\iota^{-1}(x'_i)) - \log Z_\iota(u) + \sum_{\nu=1}^{k} T_{\iota,\nu}(f_\iota^{-1}(x'_i)) \lambda_{\iota,\nu}(u) \\
&+ \log \mathrm{vol} J_{f^{-1}}(x'_j) + \sum_{\iota=1}^{n} \log Q_\iota(f_\iota^{-1}(x'_j)) - \log Z_\iota(u) + \sum_{\nu=1}^{k} T_{\iota,\nu}(f_\iota^{-1}(x'_j)) \lambda_{\iota,\nu}(u) \qquad (38)
\end{aligned}$$

Then (36) yields

$$\begin{aligned}
&\log \mathrm{vol} J_{f^{-1}}(x'_i) + \log \mathrm{vol} J_{f^{-1}}(x'_j) + \sum_{\iota=1}^{n} \log Q_\iota(f_\iota^{-1}(x'_i)) + \log Q_\iota(f_\iota^{-1}(x'_j)) \\
&- 2 \log Z_\iota(u) + < T_\iota(f_\iota^{-1}(x'_i)), \lambda_\iota(u) > + < T_\iota(f_\iota^{-1}(x'_j)), \lambda_\iota(u) >
\end{aligned}$$

$$= \log \mathrm{vol} J_{\tilde{f}^{-1}}(x'_i) + \log \mathrm{vol} J_{\tilde{f}^{-1}}(x'_j) + \sum_{\iota=1}^{n} \log Q_\iota(\tilde{f}_\iota^{-1}(x'_i)) + \log Q_\iota(\tilde{f}_\iota^{-1}(x'_j))$$
$$- 2\log \tilde{Z}_\iota(u) + < \tilde{T}_\iota(\tilde{f}_\iota^{-1}(x'_i)), \tilde{\lambda}_\iota(u) > + < \tilde{T}_\iota(\tilde{f}_\iota^{-1}(x'_j)), \tilde{\lambda}_\iota(u) > \tag{39}$$

Define $\bar{\lambda}(u) = \lambda u - \lambda u_0$. We have for $l = 1, \ldots, nk$,

$$< T(f_\iota^{-1}(x'_i)), \lambda(u) > + < T(f_\iota^{-1}(x'_j)), \lambda(u) > + \sum_{\iota=1}^{n} \log \frac{Z_\iota(u_0)}{Z_\iota(u_l)}$$
$$= < \tilde{T}(\tilde{f}_\iota^{-1}(x'_i)), \tilde{\lambda}(u) > + < \tilde{T}(\tilde{f}_\iota^{-1}(x'_j)), \tilde{\lambda}(u) > + \sum_{\iota=1}^{n} \log \frac{\tilde{Z}_\iota(u_0)}{\tilde{Z}_\iota(u_l)} \tag{40}$$

Then we have

$$L^\top \Big( T(f_\iota^{-1}(x'_i)) + T(f_\iota^{-1}(x'_i)) \Big) = \tilde{L}^\top \Big( \tilde{T}(\tilde{f}_\iota^{-1}(x'_i)) + \tilde{T}(\tilde{f}_\iota^{-1}(x'_j)) \Big) + b, \tag{41}$$

where $b_l = \sum_{\iota=1}^{n} 2 \log \frac{\tilde{Z}_\iota(u_0) Z_\iota(u_l)}{Z_\iota(u_0) \tilde{Z}_\iota(u_l)}$. Since $L$ is invertible, we have

$$T(f^{-1}(x'_i)) + T(f^{-1}(x'_j)) = A\Big( \tilde{T}(\tilde{f}^{-1}(x'_i)) + \tilde{T}(\tilde{f}^{-1}(x'_j)) \Big) + c, \tag{42}$$

where $A = L^{-\top} \tilde{L}^\top$, $c = L^{-\top} b$.

Then by the proof of Step III in the proof of Theorem 1 in [7], $A$ is invertible. □

*Proof of Theorem 3.2.* By the data generative model,

$$\mathrm{P}_{f,T,\lambda}(x'_1, \ldots, x'_B | u) = \int \cdots \int \mathrm{P}_\theta(x'_1, \ldots, x'_B, z'_1, \ldots, z'_B | u)\, dz'_1 \cdots dz'_B. \tag{43}$$

According to

$$\mathrm{P}_{f,T,\lambda}(x'_1, \ldots, x'_B | u) = \mathrm{P}_{\tilde{f},\tilde{T},\tilde{\lambda}}(x'_1, \ldots, x'_B | u), \tag{44}$$

and following the same proof of Theorem A.6, we have

$$\tilde{p}_{T,\lambda,f,u}(x'_1) \cdots \tilde{p}_{T,\lambda,f,u}(x'_B) = \tilde{p}_{\tilde{T},\tilde{\lambda},\tilde{f},u}(x'_1) \cdots \tilde{p}_{\tilde{T},\tilde{\lambda},\tilde{f},u}(x'_B), \tag{45}$$

and finally that

$$\frac{1}{B} \sum_{i=1}^{B} T(f^{-1}(x'_i)) = A \frac{1}{B} \sum_{i=1}^{B} \tilde{T}(\tilde{f}^{-1}(x'_i)) + c, \tag{46}$$

where A is invertible. □