Do More Negative Samples Necessarily Hurt in Contrastive Learning?



Pranjal Awasthi

Nishanth Dikkala*

Pritish Kamath

Google Research

Self-Supervised Learning

- ***Objective***: Learn *useful* groupings/representations of complex unlabeled data.

- Harder than supervised learning but larger potential
  - Labeling is expensive. Unlabeled data is cheap

- Of late, deep unsupervised learning approaches made significant strides
  - *Game playing* – AlphaZero – learns via self-play
  - *Masked Prediction in NLP* - Large language models – PaLM, GPT-3, OPT-175B etc.
  - *Contrastive Approaches in Vision* - SimCLR [2020] – Achieves AlexNet level downstream classification performance

Contrastive Learning

- Introduced in early 2000s
- Gaining popularity for deep unsupervised learning
    - CLIP encoder in DALL.E and DALL.E 2.
    - SimCLR [2020] – 76.5% top-1 accuracy on Imagenet (downstream)

- **High-Level Idea**: *Contrast* between different inputs.
    - Similar examples have representations close to each other.
    - Dissimilar examples have representations far from each other.

Bear in mind, digital art

A phone made of grass, digital art

3D render of a fish that looks like a corgi in an aquarium

Images generated by DALL.E 2 model of OpenAI.

The math of Contrastive Learning

The math of Contrastive Learning

$x \sim \mathcal{D}$

The math of Contrastive Learning

$x \sim \mathcal{D}$

$x^+ \sim \mathcal{A}(.\,|x)$

The math of Contrastive Learning

$x \sim \mathcal{D}$

$x^+ \sim \mathcal{A}(.\,|x)$

$x^- \sim \mathcal{D}$

The math of Contrastive Learning

(x, x^+) - positive pair

$x \sim \mathcal{D}$

$x^+ \sim \mathcal{A}(.|x)$

$x^- \sim \mathcal{D}$

The math of Contrastive Learning

(x, x^+) - positive pair

$x \sim \mathcal{D}$

$x^+ \sim \mathcal{A}(.|x)$

negative sample

$x^- \sim \mathcal{D}$

The math of Contrastive Learning

(x, x^+) - positive pair

$x \sim \mathcal{D}$

$f_\theta(x) \in \mathbb{R}^d$

$x^+ \sim \mathcal{A}(.\,|x)$

$f_\theta(x^+) \in \mathbb{R}^d$

negative sample

$x^- \sim \mathcal{D}$

$f_\theta(x^-) \in \mathbb{R}^d$

The math of Contrastive Learning

(x, x^+) - positive pair

$x \sim \mathcal{D}$

$f_\theta(x) \in \mathbb{R}^d$

$x^+ \sim \mathcal{A}(.|x)$

$f_\theta(x^+) \in \mathbb{R}^d$

negative sample

$x^- \sim \mathcal{D}$

$f_\theta(x^-) \in \mathbb{R}^d$

Assume $\|f_\theta(x)\|_2 = 1$
(standard in practice)
Hence, representations are vectors on the sphere $\mathbb{S}^{d-1}$

The math of Contrastive Learning

(x, x^+) - positive pair

$x \sim \mathcal{D}$

$f_\theta(x) \in \mathbb{R}^d$

$x^+ \sim \mathcal{A}(.\,|x)$

$f_\theta(x^+) \in \mathbb{R}^d$

negative sample

$x^- \sim \mathcal{D}$

$f_\theta(x^-) \in \mathbb{R}^d$

Assume $\|f_\theta(x)\|_2 = 1$ (standard in practice)
Hence, representations are vectors on the sphere $\mathbb{S}^{d-1}$

<u>Noise contrastive estimation</u>

$$\min_\theta \text{ NCE Loss} \equiv \mathbb{E}_{x,x^+,x^-} \log\left(1 + \frac{\exp\left(f(x)^T f(x^-)\right)}{\exp\left(f(x)^T f(x^+)\right)}\right)$$

The math of Contrastive Learning

(x, x^+) - positive pair

$x \sim \mathcal{D}$

$f_\theta(x) \in \mathbb{R}^d$

$x^+ \sim \mathcal{A}(.|x)$

$f_\theta(x^+) \in \mathbb{R}^d$

negative sample

$x^- \sim \mathcal{D}$

$f_\theta(x^-) \in \mathbb{R}^d$

Assume $\|f_\theta(x)\|_2 = 1$ (standard in practice)

Hence, representations are vectors on the sphere $\mathbb{S}^{d-1}$

<u>Noise contrastive estimation</u>

$$\min_\theta \text{ NCE Loss} \equiv \mathbb{E}_{x,x^+,x^-} \log\left(1 + \frac{\boldsymbol{exp}\left(\boldsymbol{f}(\boldsymbol{x})^{\boldsymbol{T}} \boldsymbol{f}(\boldsymbol{x}^-)\right)}{\boldsymbol{exp}\left(\boldsymbol{f}(\boldsymbol{x})^{\boldsymbol{T}} \boldsymbol{f}(\boldsymbol{x}^+)\right)}\right)$$

Multiple Negative Samples

- $(x, x^+), (x_1^-, x_2^-, \dots, x_k^-) \sim \mathcal{D}^k$
- NCE Loss = $\mathbb{E}_{x,x^+,x_{1:k}^-} \log\left(1 + \frac{\sum_{i=1}^{k} \exp\left(f(x)^T f(x_i^-)\right)}{\exp\left(f(x)^T f(x^+)\right)}\right)$

- Intuition: More contrastive signal for the learner
- Simulates batches in practice.
- SimCLR uses up to 4096 negative samples per positive pair!

Three central questions

1. *Why does minimizing the contrastive loss help with downstream inference tasks?*
2. *What is the effect of increasing the number of negative samples?*
3. *What is the geometry of the representations optimizing the population contrastive loss?*

A Theoretical Model *[Saunshi et al 2019]*

$\mathcal{C}$- set of latent classes, $|\mathcal{C}| = C$. Distribution over classes - ρ

1. Sample $c \sim \rho$.
2. Sample $(x, x^+) \sim \mathcal{D}_c^2$
3. Sample $x_1^-, x_2^-, \ldots, x_k^- \sim \mathcal{D}^k$ where $\mathcal{D} = \sum_c \mathcal{D}_c \rho_c$ is marginal distribution.

A Theoretical Model *[Saunshi et al 2019]*

$\mathcal{C}$- set of latent classes, $|\mathcal{C}| = C$. Distribution over classes - ρ

1. Sample $c \sim \rho$.
2. Sample $(x, x^+) \sim \mathcal{D}_c^2$
3. Sample $x_1^-, x_2^-, \ldots, x_k^- \sim \mathcal{D}^k$ where $\mathcal{D} = \sum_c \mathcal{D}_c \rho_c$ is marginal distribution.

A Theoretical Model *[Saunshi et al 2019]*

$\mathcal{C}$- set of latent classes, $|\mathcal{C}| = C$. Distribution over classes - ρ

1. Sample $c \sim \rho$.
2. Sample $(x, x^+) \sim \mathcal{D}_c^2$
3. Sample $x_1^-, x_2^-, \dots, x_k^- \sim \mathcal{D}^k$ where $\mathcal{D} = \sum_c \mathcal{D}_c \rho_c$ is marginal distribution.

$x =$  $x^+ =$ 

A Theoretical Model *[Saunshi et al 2019]*

$\mathcal{C}$- set of latent classes, $|\mathcal{C}| = C$. Distribution over classes - ρ

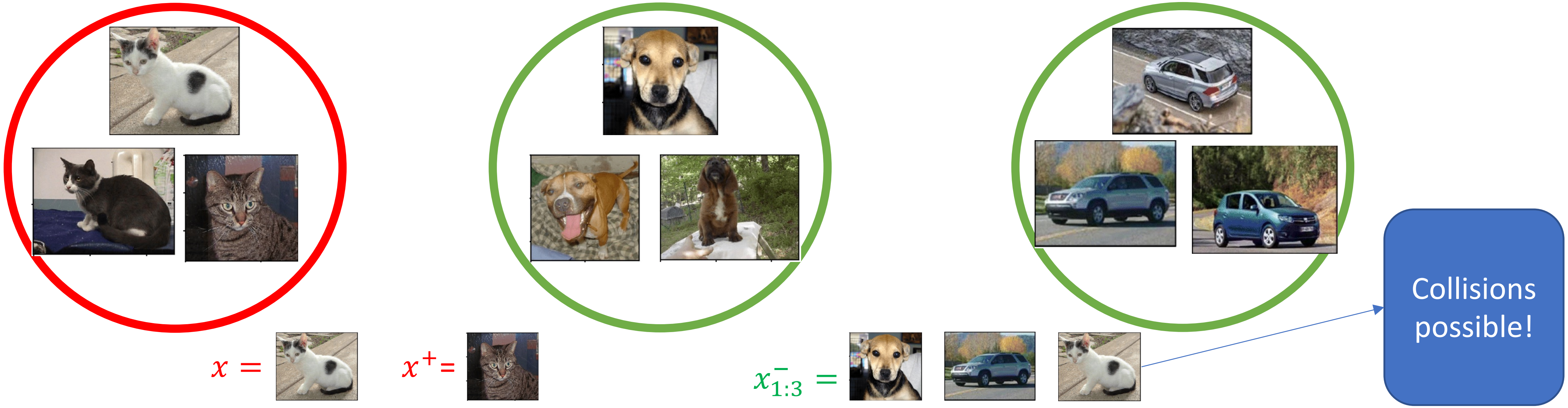
1. Sample $c \sim \rho$.
2. Sample $(x, x^+) \sim \mathcal{D}_c^2$
3. Sample $x_1^-, x_2^-, \ldots, x_k^- \sim \mathcal{D}^k$ where $\mathcal{D} = \sum_c \mathcal{D}_c \rho_c$ is marginal distribution.

$x =$  $x^+ =$  $x_{1:3}^- =$ 

A Theoretical Model *[Saunshi et al 2019]*

$\mathcal{C}$- set of latent classes, $|\mathcal{C}| = C$. Distribution over classes - ρ

1. Sample $c \sim \rho$.
2. Sample $(x, x^+) \sim \mathcal{D}_c^2$
3. Sample $x_1^-, x_2^-, \ldots, x_k^- \sim \mathcal{D}^k$ where $\mathcal{D} = \sum_c \mathcal{D}_c \rho_c$ is marginal distribution.

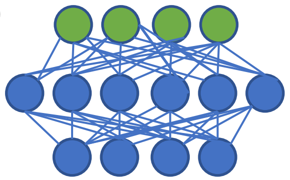


A Theoretical Model *[Saunshi et al 2019]*

- **NCE Loss**: $\mathcal{L}_{NCE}^{(k)}(f) = \mathbb{E}_{\mathcal{D}_{NCE}}\left[\ell\left(\left\langle f(x)^T\left(f(x^+) - f(x_i^-)\right)\right\rangle_{i=1}^{k}\right)\right]$

  when $\ell(v) = \log\left(1 + \sum_{i=1}^{k} \exp(-v_i)\right)$ we recover logistic loss.

Learn representation
(green layer)



A Theoretical Model *[Saunshi et al 2019]*

- **NCE Loss**: $\mathcal{L}_{NCE}^{(k)}(f) = \mathbb{E}_{\mathcal{D}_{NCE}}\left[\ell\left(\left\langle f(x)^T\left(f(x^+) - f(x_i^-)\right)\right\rangle_{i=1}^{k}\right)\right]$

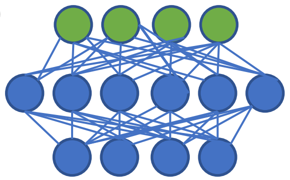
  when $\ell(v) = \log\left(1 + \sum_{i=1}^{k} \exp(-v_i)\right)$ we recover logistic loss.

- **Downstream Supervised Task**: Classify examples from $\mathcal{C}$ using a linear predictor over the representations.

  For any d-dimensional representation $f : \mathcal{X} \to \mathbb{S}^{d-1}$,

$$\mathcal{L}_{sup}(f) = \inf_{\{w_c |\ c \in \mathcal{C}, \|w_c\| = 1\}} \mathbb{E}_{(x,c)\sim\mathcal{D}_{sup}} \ell(\langle f(x)^T(w_c - w_{c'})\rangle_{c' \neq c})$$

Learn representation
(green layer)



A Theoretical Model *[Saunshi et al 2019]*

- **NCE Loss**: $\mathcal{L}_{NCE}^{(k)}(f) = \mathbb{E}_{\mathcal{D}_{NCE}}\left[\ell\left(\langle f(x)^T\left(f(x^+) - f(x_i^-)\right)\rangle_{i=1}^{k}\right)\right]$

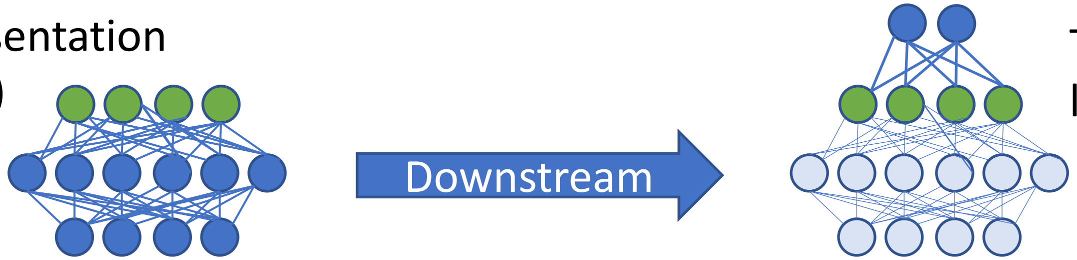
  when $\ell(v) = \log\left(1 + \sum_{i=1}^{k} \exp(-v_i)\right)$ we recover logistic loss.

- **Downstream Supervised Task**: Classify examples from $\mathcal{C}$ using a linear predictor over the representations.

  For any d-dimensional representation $f : \mathcal{X} \rightarrow \mathbb{S}^{d-1}$,

$$\mathcal{L}_{sup}(f) = \inf_{\{w_c |\ c \in \mathcal{C}, \|w_c\|=1\}} \mathbb{E}_{(x,c)\sim\mathcal{D}_{sup}}\, \ell(\langle f(x)^T(w_c - w_{c'})\rangle_{c' \neq c})$$

Learn representation (green layer)



Train linear predictor on top of learnt representations

Results of Prior Work

Results of Prior Work

- *[Saunshi et al 2019], [Ash et al 2021]:*
  For any representation f,

$$\mathcal{L}_{sup}(f) \leq \alpha(k,\rho)\left(\mathcal{L}_{NCE}^{(k)}(f) - \tau(k)\right)$$

  where α initially decreases with k, then grows exponentially with k.

  ($\frac{\eta^k}{k}$ for $\eta = 1 + \frac{1}{C-1}$)

Results of Prior Work

- *[Saunshi et al 2019], [Ash et al 2021]:*
  For any representation f,

  $$\mathcal{L}_{sup}(f) \leq \alpha(k,\rho)\left(\mathcal{L}_{NCE}^{(k)}(f) - \tau(k)\right)$$

  where α initially decreases with k, then grows exponentially with k.

  ($\frac{\eta^k}{k}$ for $\eta = 1 + \frac{1}{C-1}$)

- For small k, minimizing $\mathcal{L}_{NCE}^{(k)}(f) \Longrightarrow$ minimizing $\mathcal{L}_{sup}(f)$.

Results of Prior Work

- *[Saunshi et al 2019], [Ash et al 2021]:*
  For any representation f,

$$\mathcal{L}_{sup}(f) \leq \alpha(k,\rho)\left(\mathcal{L}_{NCE}^{(k)}(f) - \tau(k)\right)$$

  where α initially decreases with k, then grows exponentially with k.

  ($\frac{\eta^k}{k}$ for $\eta = 1 + \frac{1}{C-1}$)
- For small k, minimizing $\mathcal{L}_{NCE}^{(k)}(f) \Longrightarrow$ minimizing $\mathcal{L}_{sup}(f)$.
- However, performance of contrastive learning seems to degrade exponentially fast with increasing k.

Results of Prior Work

- *[Saunshi et al 2019], [Ash et al 2021]:*
  For any representation f,

$$\mathcal{L}_{sup}(f) \leq \alpha(k, \rho) \left(\mathcal{L}_{NCE}^{(k)}(f) - \tau(k) \right)$$

  where α initially decreases with k, then grows exponentially with k.

  ($\frac{\eta^k}{k}$ for $\eta = 1 + \frac{1}{C-1}$)
- For small k, minimizing $\mathcal{L}_{NCE}^{(k)}(f) \Longrightarrow$ minimizing $\mathcal{L}_{sup}(f)$.
- However, performance of contrastive learning seems to degrade exponentially fast with increasing k.
- **Collision-Coverage Tradeoff**: Collisions are negative samples which are drawn from same class as x. As k increases,
  - First the coverage of all classes in $\mathcal{C}$ increases, so performance improves.
  - As k increases further, #collisions increase leading to degradation in performance.

Our Results

- ***We prove that collision-coverage tradeoff doesn't exist for the minimizer of a certain family of contrastive losses (includes logistic loss)!***
- ***The downstream classification accuracy of the optimal representation doesn't necessarily degrade with increasing*** k.

Our Results

- ***We prove that collision-coverage tradeoff doesn't exist for the minimizer of a certain family of contrastive losses (includes logistic loss)!***
- ***The downstream classification accuracy of the optimal representation doesn't necessarily degrade with increasing k.***

- Focus on population loss to decouple issues of generalization.

Our Results

- ***We prove that collision-coverage tradeoff doesn't exist for the minimizer of a certain family of contrastive losses (includes logistic loss)!***
- ***The downstream classification accuracy of the optimal representation doesn't necessarily degrade with increasing k.***

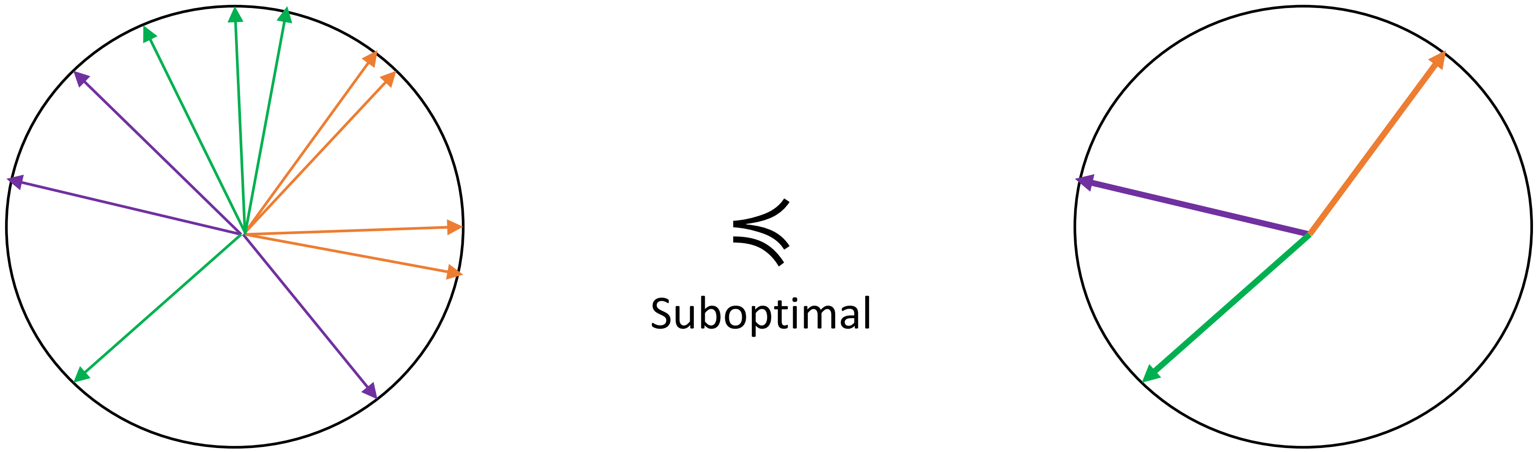
- Focus on population loss to decouple issues of generalization.
- ***(Assumption) Non-overlapping latent classes:*** Distributions $\{\mathcal{D}_c : c \in \mathcal{C}\}$ have disjoint supports.

Our Results

***Lemma 1*: (A Structural Property of the Optimal Representation)**

The optimal representation is *latent-indistinguishable for any k*.
For a strictly convex loss ℓ, and any $f: \mathcal{X} \to \mathbb{S}^{d-1}$, there exists $\tilde{f}: \mathcal{X} \to \mathbb{S}^{d-1}$, such that

- $\tilde{f}(x) = \tilde{f}(x')$ for any x, x' from the same class and
- $\mathcal{L}_{NCE}^{(k)}(\tilde{f}) < \mathcal{L}_{NCE}^{(k)}(f)$.



Our Results

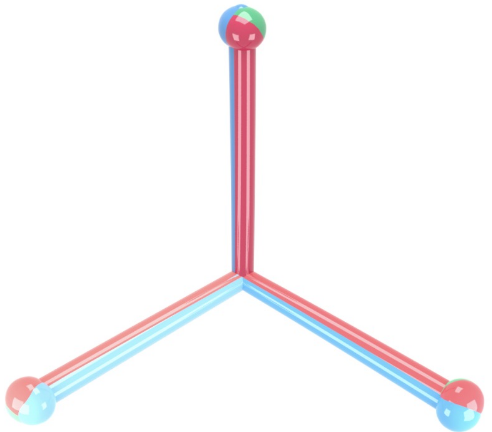
Balanced class distribution $\left(\rho_c = \frac{1}{C}\right)$ - For **any k**, optimal representation is Simplex ETF (when $d \geq C - 1$).

- Simplex Equiangular Triangular Frame (ETF):
  - $f(x) = f(x')$ for any x, x' from the same class
  - Angle between representations of any two distinct classes is same.
  - $f(x)^T f(x') = -\frac{1}{C-1}$ for any x, x' $s.t.\, c(x) \neq c(x')$.

- Therefore, downstream classification loss is non-increasing with increasing k.

Unbalanced class distributions

- Optimal representation is latent-indistinguishable with separation between representations determined by class distribution ρ.
- **Conjecture**: Downstream classification loss is non-increasing with k.
- provide evidence from simulations



Example of Simplex ETF with 3 vectors in 3+ dimensional space. [Papyan, Han and Donoho (2020)]
Figure from: https://medium.com/mlearning-ai/what-is-neural-collapse-de1decf83f48

Proof Outline –(1) Optimality of Latent-Indistinguishable Representations

Proof Outline –(1) Optimality of Latent-Indistinguishable Representations

For any representation f, and a fixed latent class $c^* \in \mathcal{C}$, sample $x^* \sim \mathcal{D}_{c^*}$ and define f_{x^*} as follows:

$f_{x^*}(x) = f(x^*)$ if $x \in c^*$,
$f_{x^*}(x) = f(x)$ if $x \notin c^*$.

Proof Outline –(1) Optimality of Latent-Indistinguishable Representations

For any representation f, and a fixed latent class $c^* \in \mathcal{C}$, sample $x^* \sim \mathcal{D}_{c^*}$ and define f_{x^*} as follows:

$f_{x^*}(x) = f(x^*)$ if $x \in c^*$,
$f_{x^*}(x) = f(x)$ if $x \notin c^*$.

- We show that $\mathbb{E}_{x^* \sim \mathcal{D}_{c^*}}\left[\mathcal{L}_{NCE}^{(k)}(f_{x^*})\right] < \mathcal{L}_{NCE}^{(k)}(f)$.

Proof Outline –(1) Optimality of Latent-Indistinguishable Representations

For any representation f, and a fixed latent class $c^* \in \mathcal{C}$, sample $x^* \sim \mathcal{D}_{c^*}$ and define f_{x^*} as follows:

$f_{x^*}(x) = f(x^*)$ if $x \in c^*$,
$f_{x^*}(x) = f(x)$ if $x \notin c^*$.

- We show that $\mathbb{E}_{x^* \sim \mathcal{D}_{c^*}}\left[\mathcal{L}_{NCE}^{(k)}(f_{x^*})\right] < \mathcal{L}_{NCE}^{(k)}(f)$.
- Implies there exists x^* such that $\mathcal{L}_{NCE}^{(k)}(f_{x^*}) < \mathcal{L}_{NCE}^{(k)}(f)$.

Proof Outline –(1) Optimality of Latent-Indistinguishable Representations

For any representation f, and a fixed latent class $c^* \in \mathcal{C}$, sample $x^* \sim \mathcal{D}_{c^*}$ and define f_{x^*} as follows:

$f_{x^*}(x) = f(x^*)$ if $x \in c^*$,
$f_{x^*}(x) = f(x)$ if $x \notin c^*$.

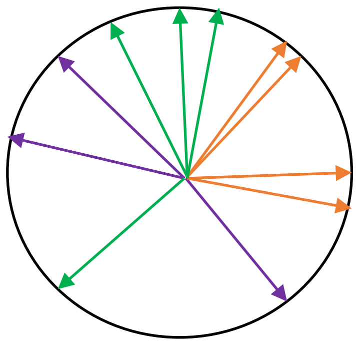
- We show that $\mathbb{E}_{x^* \sim \mathcal{D}_{c^*}}\left[\mathcal{L}_{NCE}^{(k)}(f_{x^*})\right] < \mathcal{L}_{NCE}^{(k)}(f)$.
- Implies there exists x^* such that $\mathcal{L}_{NCE}^{(k)}(f_{x^*}) < \mathcal{L}_{NCE}^{(k)}(f)$.
- Iterate over all $c^* \in \mathcal{C}$ one by one.

Proof Outline –(1) Optimality of Latent-Indistinguishable Representations

For any representation f, and a fixed latent class $c^* \in \mathcal{C}$, sample $x^* \sim \mathcal{D}_{c^*}$ and define f_{x^*} as follows:

$f_{x^*}(x) = f(x^*)$ if $x \in c^*$,
$f_{x^*}(x) = f(x)$ if $x \notin c^*$.

- We show that $\mathbb{E}_{x^* \sim \mathcal{D}_{c^*}}\left[\mathcal{L}_{NCE}^{(k)}(f_{x^*})\right] < \mathcal{L}_{NCE}^{(k)}(f)$.
- Implies there exists x^* such that $\mathcal{L}_{NCE}^{(k)}(f_{x^*}) < \mathcal{L}_{NCE}^{(k)}(f)$.
- Iterate over all $c^* \in \mathcal{C}$ one by one.

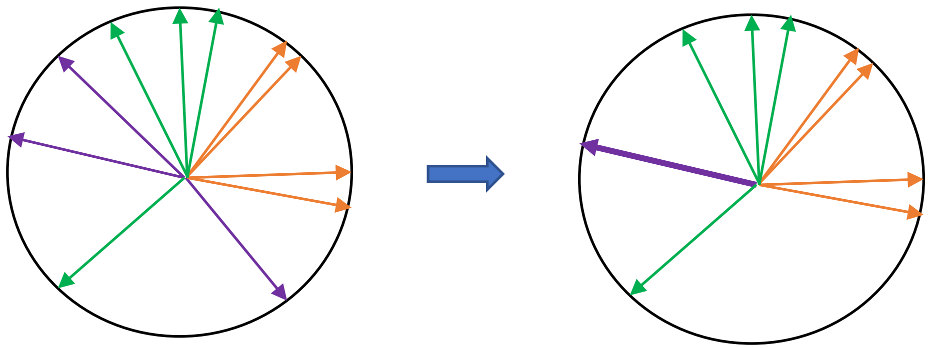


Proof Outline –(1) Optimality of Latent-Indistinguishable Representations

For any representation f, and a fixed latent class $c^* \in \mathcal{C}$, sample $x^* \sim \mathcal{D}_{c^*}$ and define f_{x^*} as follows:

$f_{x^*}(x) = f(x^*)$ if $x \in c^*$,
$f_{x^*}(x) = f(x)$ if $x \notin c^*$.

- We show that $\mathbb{E}_{x^* \sim \mathcal{D}_{c^*}}\left[\mathcal{L}_{NCE}^{(k)}(f_{x^*})\right] < \mathcal{L}_{NCE}^{(k)}(f)$.
- Implies there exists x^* such that $\mathcal{L}_{NCE}^{(k)}(f_{x^*}) < \mathcal{L}_{NCE}^{(k)}(f)$.
- Iterate over all $c^* \in \mathcal{C}$ one by one.

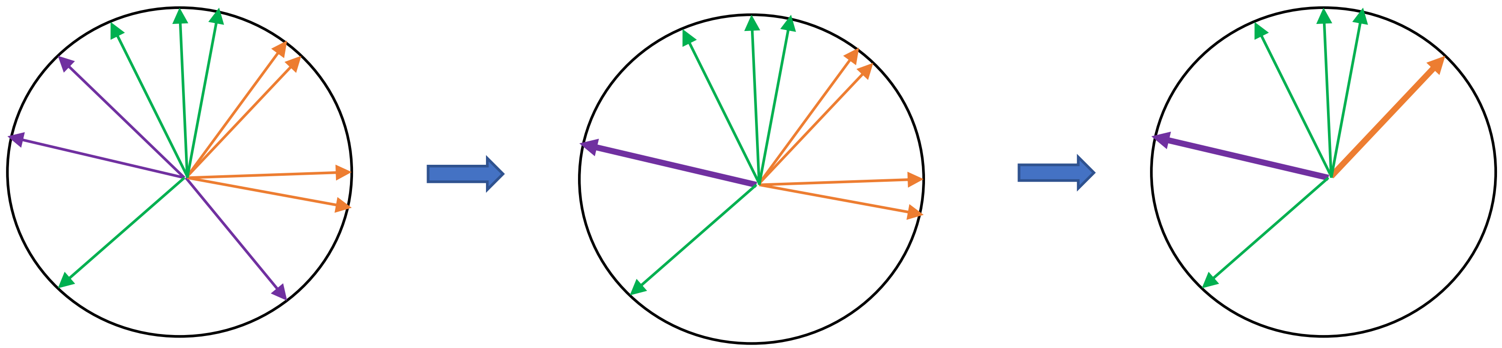


Proof Outline –(1) Optimality of Latent-Indistinguishable Representations

For any representation f, and a fixed latent class $c^* \in \mathcal{C}$, sample $x^* \sim \mathcal{D}_{c^*}$ and define f_{x^*} as follows:

$$f_{x^*}(x) = f(x^*) \text{ if } x \in c^*,$$
$$f_{x^*}(x) = f(x) \text{ if } x \notin c^*.$$

- We show that $\mathbb{E}_{x^* \sim \mathcal{D}_{c^*}}\left[\mathcal{L}_{NCE}^{(k)}(f_{x^*})\right] < \mathcal{L}_{NCE}^{(k)}(f)$.
- Implies there exists x^* such that $\mathcal{L}_{NCE}^{(k)}(f_{x^*}) < \mathcal{L}_{NCE}^{(k)}(f)$.
- Iterate over all $c^* \in \mathcal{C}$ one by one.

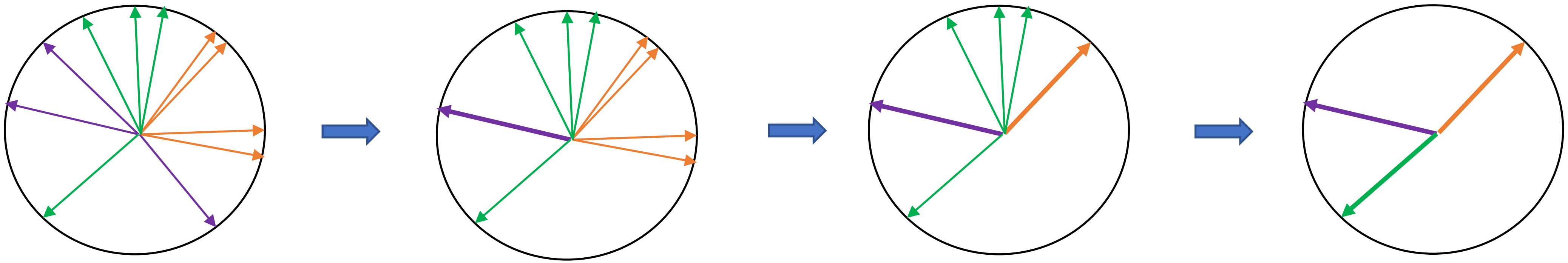


Proof Outline –(1) Optimality of Latent-Indistinguishable Representations

For any representation f, and a fixed latent class $c^* \in \mathcal{C}$, sample $x^* \sim \mathcal{D}_{c^*}$ and define f_{x^*} as follows:

$f_{x^*}(x) = f(x^*)$ if $x \in c^*$,
$f_{x^*}(x) = f(x)$ if $x \notin c^*$.

- We show that $\mathbb{E}_{x^* \sim \mathcal{D}_{c^*}}\left[\mathcal{L}_{NCE}^{(k)}(f_{x^*})\right] < \mathcal{L}_{NCE}^{(k)}(f)$.
- Implies there exists x^* such that $\mathcal{L}_{NCE}^{(k)}(f_{x^*}) < \mathcal{L}_{NCE}^{(k)}(f)$.
- Iterate over all $c^* \in \mathcal{C}$ one by one.

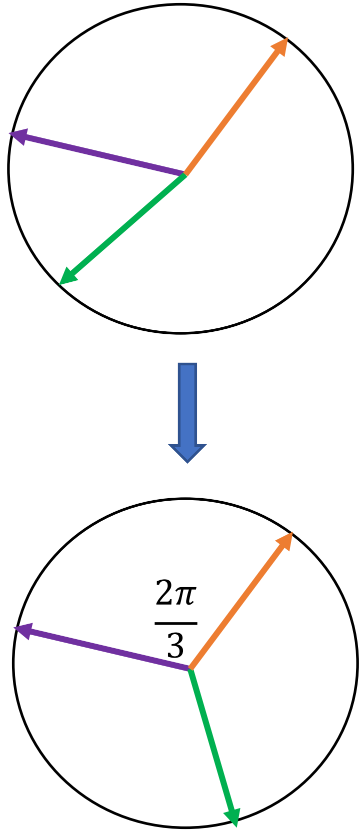


Proof Outline –(2) Simplex ETF Optimality for Balanced Class Distribution

Established that optimal representation is latent-indistinguishable.

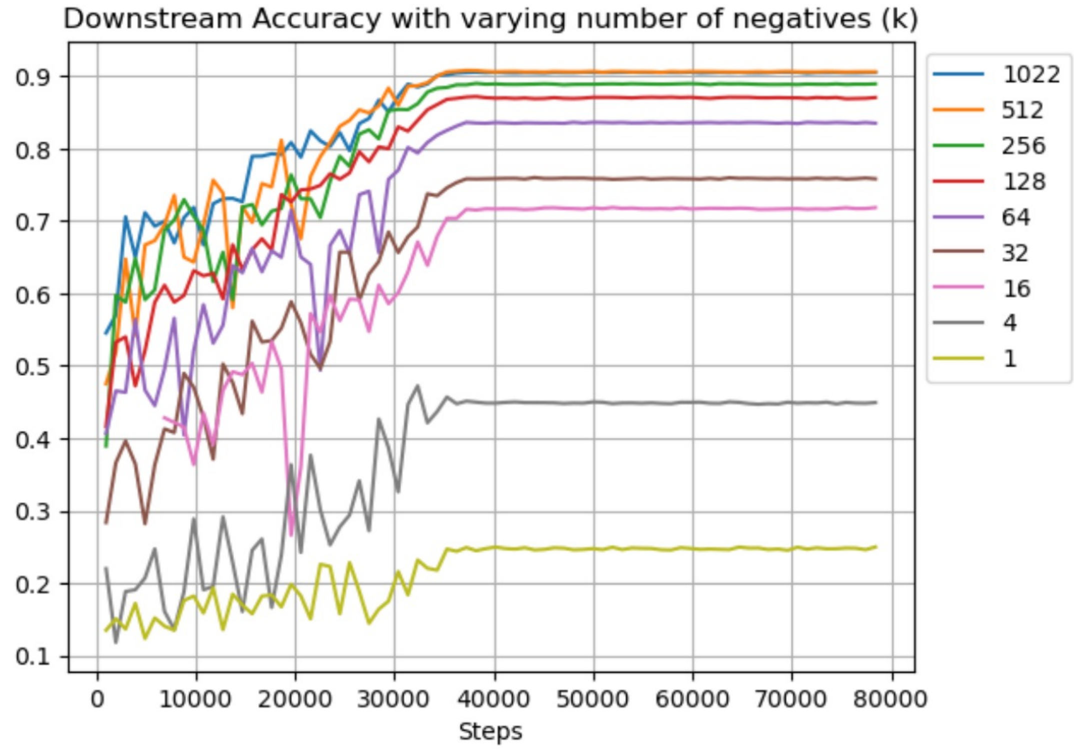
- **Step 1**: Equiangularity is optimal (Jensen's inequality).

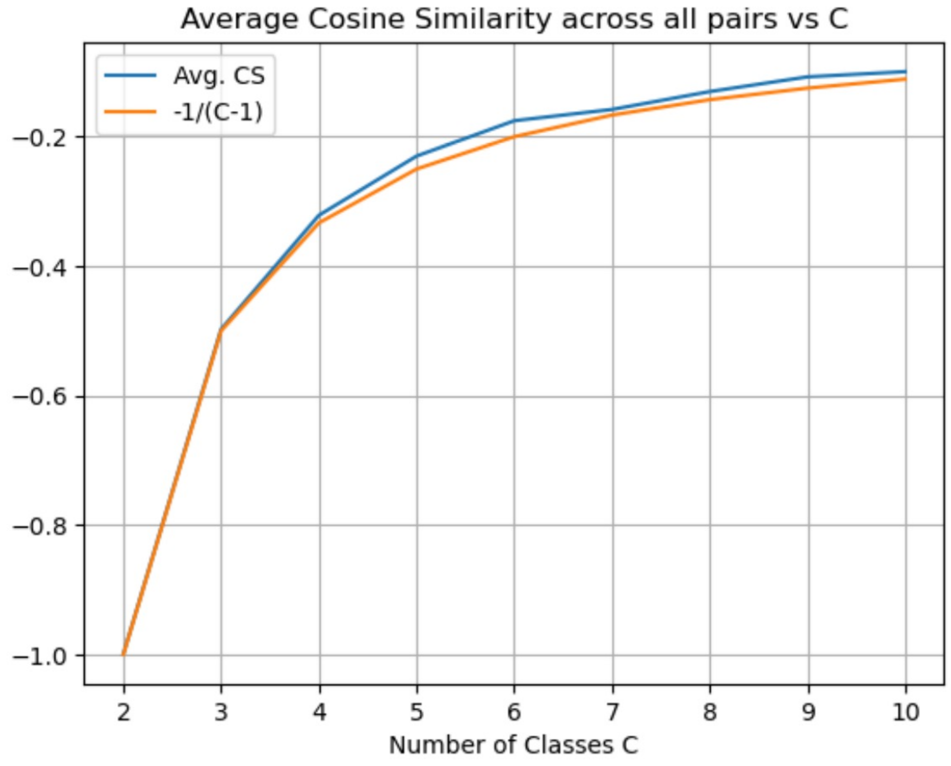
- **Step 2**: Among equiangular representations, simplex ETF is optimal.
  - $\|\sum_c u_c\|_2^2 = C + \sum_{c' \neq c} u_c^T u_{c'} \implies \mathbb{E}_{c,c' \sim \rho}[u_c^T u_{c'} \mid c \neq c'] \geq -\frac{1}{C-1}$
  - Simplex ETF achieves the $-\frac{1}{C-1}$ lower bound.



Experiments

CIFAR 10/100 – Balanced datasets





$$\text{Cosine Similarity}(u, v) = \frac{u^\top v}{\|u\|_2 \|v\|_2}$$

Concluding Remarks

Concluding Remarks

- **Main conclusion**: More negative examples need not be harmful in contrastive learning.
  - Also supported by results of [Bao et al (2021), Nozawa and Sato (2021)].
    Show upper and lower bounds of form

$$\mathcal{L}_{sup}(f) \in \left[\mathcal{L}_{NCE}^{(k)}(f) - \Delta_L, \mathcal{L}_{NCE}^{(k)}(f) + \Delta_U\right]$$

    where $\Delta_U - \Delta_L = \alpha + 2\log\left(1 + \frac{1}{k}\right)$ where α does not depend on k.
    (Don't give structural characterization we do and do not imply monotonicity for balanced classes)

Concluding Remarks

- **Main conclusion**: More negative examples need not be harmful in contrastive learning.
    - Also supported by results of [Bao et al (2021), Nozawa and Sato (2021)]. Show upper and lower bounds of form

$$\mathcal{L}_{sup}(f) \in \left[\mathcal{L}_{NCE}^{(k)}(f) - \Delta_L, \mathcal{L}_{NCE}^{(k)}(f) + \Delta_U\right]$$

      where $\Delta_U - \Delta_L = \alpha + 2\log\left(1 + \frac{1}{k}\right)$ where α does not depend on k.
      (Don't give structural characterization we do and do not imply monotonicity for balanced classes)

- **Open directions**:
    - **Conjecture**: Show downstream supervised loss is non-decreasing with increasing number of negatives k even for unbalanced class distributions.
    - More Realistic Models
        - Realistic Augmentation Distributions – no class knowledge, minimal overlap
        - Inductive bias of encoder – paper in this ICML by Saunshi et al (2022)
        - Multiple downstream tasks

Concluding Remarks

- **Main conclusion**: More negative examples need not be harmful in contrastive learning.
  - Also supported by results of [Bao et al (2021), Nozawa and Sato (2021)].
    Show upper and lower bounds of form

$$\mathcal{L}_{sup}(f) \in \left[\mathcal{L}_{NCE}^{(k)}(f) - \Delta_L, \mathcal{L}_{NCE}^{(k)}(f) + \Delta_U\right]$$

    where $\Delta_U - \Delta_L = \alpha + 2\log\left(1 + \frac{1}{k}\right)$ where α does not depend on k.
    (Don't give structural characterization we do and do not imply monotonicity for balanced classes)

- **Open directions**:
  - **Conjecture**: Show downstream supervised loss is non-decreasing with increasing number of negatives k even for unbalanced class distributions.
  - More Realistic Models
    - Realistic Augmentation Distributions – no class knowledge, minimal overlap
    - Inductive bias of encoder – paper in this ICML by Saunshi et al (2022)
    - Multiple downstream tasks

Thank you!