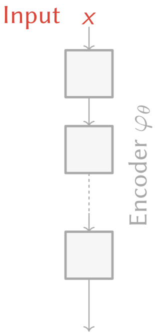
Dissecting Supervised Contrastive Learning

Florian Graf[†], C. Hofer[†], M. Niethammer[‡] and R. Kwitt[†]

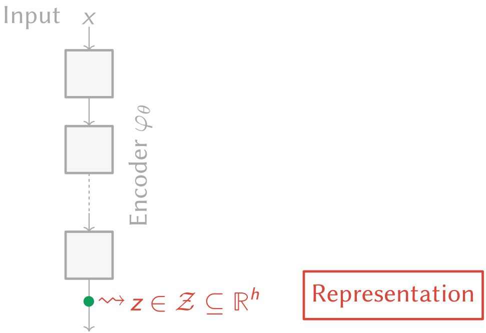
[†]University of Salzburg, [‡]UNC Chapel Hill

Background

Many **(neural) classifiers** can be abstracted as follows:

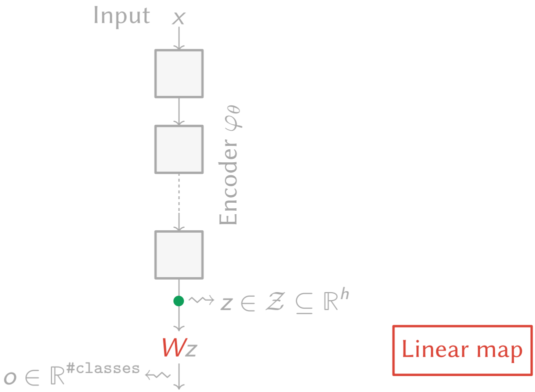


Many **(neural) classifiers** can be abstracted as follows:



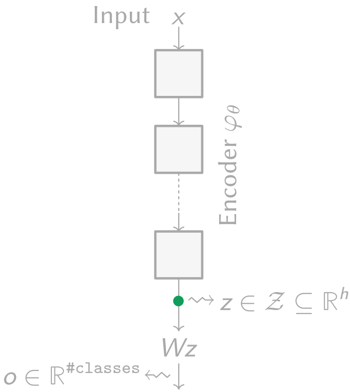
Background

Many **(neural) classifiers** can be abstracted as follows:



Background

Many **(neural) classifiers** can be abstracted as follows:

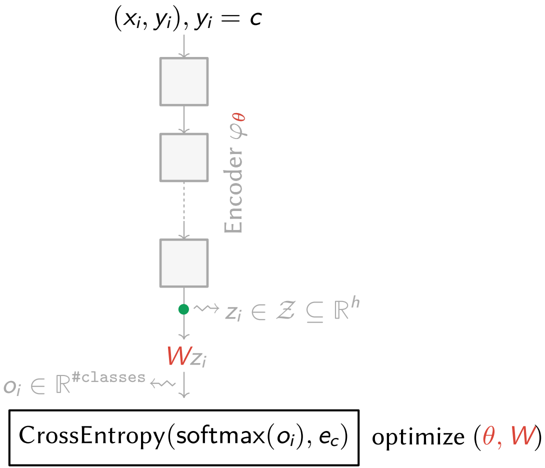


$\widehat{c} = \arg\max(o)$

Decision rule

Background

Predominantly, we minimize **cross-entropy**, wrt. (θ, W), i.e.,



Prior work

Within this regime, several avenues have been pursued.

Prior work

Within this regime, several avenues have been pursued.

This includes (among others):

- Fixing the classifier weights
  (e.g., random or, a-priori, max. spaced on the sphere)
  [Hoffer et al., ICLR '18]
  [Mettes et al., NeurIPS '19]

Prior work

Within this regime, several avenues have been pursued.

This includes (among others):

- Fixing the classifier weights
  (e.g., random or, a-priori, max. spaced on the sphere)
  [Hoffer et al., ICLR '18]
  [Mettes et al., NeurIPS '19]
- Promoting max. spaced classifier weights on the sphere
  (as optimization objective)
  [Liu et al., NeurIPS '18]

Prior work

Within this regime, several avenues have been pursued.

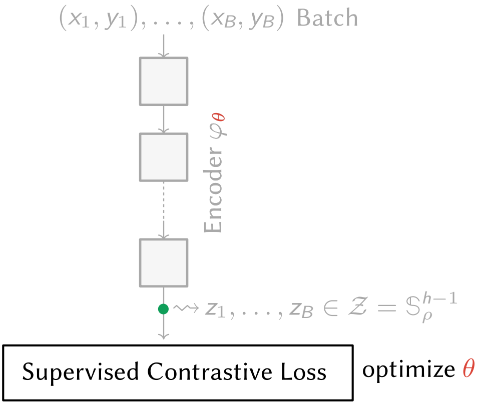
This includes (among others):

- Fixing the classifier weights
  (e.g., random or, a-priori, max. spaced on the sphere)
  [Hoffer et al., ICLR '18]
  [Mettes et al., NeurIPS '19]
- Promoting max. spaced classifier weights on the sphere
  (as optimization objective)
  [Liu et al., NeurIPS '18]
- Theoretically studying *neural collapse* phenomena
  (in the terminal stage of training)
  [Papyan et al., PNAS '20]

Alternatively, [Khosla et al., NeurIPS '20], directly optimize φ_θ:

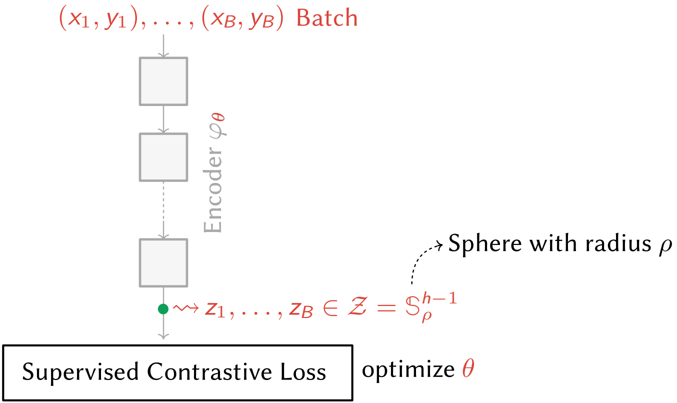
Alternatively, [Khosla et al., NeurIPS '20], directly optimize φ_θ:

Step 1



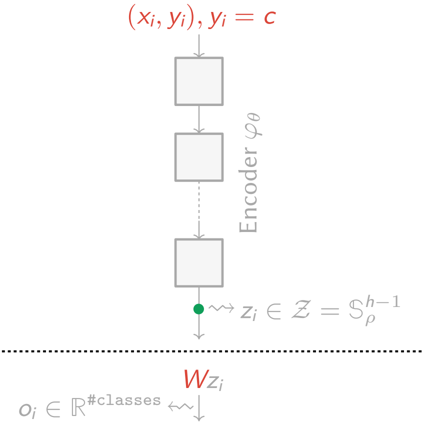
Alternatively, [Khosla et al., NeurIPS '20], directly optimize φ_θ:

Step 1



Alternatively, [Khosla et al., NeurIPS '20], directly optimize φ_θ:

Step 2



$\boxed{\mathrm{CrossEntropy}(\mathrm{softmax}(o_i), e_c)}$ optimize W

Prior work

A recent body of work considers **theoretical aspects** of contrastive learning in an

unsupervised/self-supervised regime.

Prior work

A recent body of work considers **theoretical aspects** of contrastive learning in an

unsupervised/self-supervised regime.

This includes:

- Generalization guarantees for downstream classifiers
  (by formalizing semantic similarity via latent classes)
  [Arora et al., ICML '19]

Prior work

A recent body of work considers **theoretical aspects** of contrastive learning in an

unsupervised/self-supervised regime.

This includes:

- Generalization guarantees for downstream classifiers
  (by formalizing semantic similarity via latent classes)
  [Arora et al., ICML '19]
- Asymptotic geometric properties of representations
  (by studying alignment & uniformity)
  [Wang & Isola, ICML '20]

Problem statement

Question

Are **representations**, learned by φ_θ

1. via the cross-entropy (**CE**), or
2. via the supervised contrastive (**SC**)

objective, geometrically similar?

Problem statement

Question

Are **representations**, learned by φ_θ

1. via the cross-entropy (**CE**), or
2. via the supervised contrastive (**SC**)

objective, geometrically similar?

We study this question at **optimality**, i.e., which N-points

$$Z_\theta = (\varphi_\theta(x_1), \ldots, \varphi_\theta(x_N)) \in \mathcal{Z}^N$$

minimize the **CE** / **SC** loss?

Problem statement

Formally[1],

$$\underset{\theta}{\operatorname{argmin}} \ \mathbf{loss}(\varphi_\theta(x_1), \ldots, \varphi_\theta(x_N); Y)$$

[1]for **CE**, the classifier weights, *W*, need to be included as well

Problem statement

Formally[1],

$$\underset{\theta}{\operatorname{argmin}}\ \mathbf{loss}(\varphi_\theta(x_1), \ldots, \varphi_\theta(x_N); Y)$$

[1] for **CE**, the classifier weights, W, need to be included as well

Problem statement

Formally[1],

$$\underset{\theta}{\operatorname{argmin}}\ \mathbf{loss}(\varphi_\theta(x_1), \ldots, \varphi_\theta(x_N); Y)$$

[1] for **CE**, the classifier weights, *W*, need to be included as well

Problem statement

Formally[1],

$$\operatorname*{argmin}_{\theta} \mathbf{loss}(\varphi_\theta(x_1), \ldots, \varphi_\theta(x_N); Y)$$

Assumption

We assume a **powerful enough**[2] encoder φ_θ.

[1] for **CE**, the classifier weights, W, need to be included as well

[2] capable of yielding any geometric arrangement of $(\varphi_\theta(x_1), \ldots, \varphi_\theta(x_N)) \in \mathcal{Z}^N$

Problem statement

Formally[1],

$$\underset{z_1,\ldots,z_N}{\operatorname{argmin}}\ \mathbf{loss}(z_1, \ldots, z_N; Y)$$

Assumption

We assume a **powerful enough**[2] encoder φ_θ.

Hence, we search for configurations of N (free) points, i.e.,

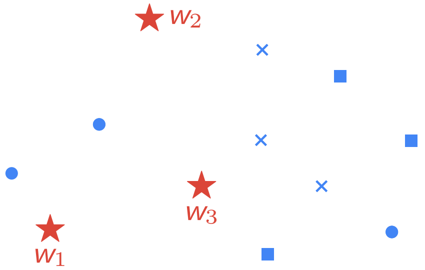
$$Z = (z_1, \ldots, z_N) \in \mathcal{Z}^N$$

minimizing the **CE** and the **SC** loss, respectively.

[1] for **CE**, the classifier weights, W, need to be included as well

[2] capable of yielding any geometric arrangement of $(\varphi_\theta(x_1), \ldots, \varphi_\theta(x_N)) \in \mathcal{Z}^N$

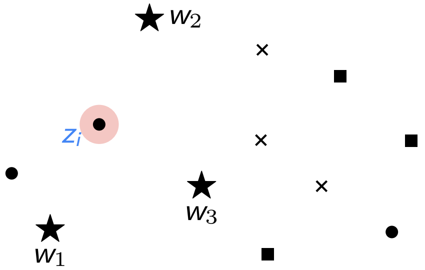
Cross-entropy (CE) loss $\mathcal{L}_{\mathrm{CE}}(Z, W; Y)$



$$\mathcal{L}_{\mathrm{CE}}(Z, W; Y)$$

$$-\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\exp(\langle z_i, w_{y_i}\rangle)}{\sum_{c=1}^{K}\exp(\langle z_i, w_c\rangle)}\right)$$

Cross-entropy (CE) loss $\mathcal{L}_{\mathrm{CE}}(Z, W; Y)$

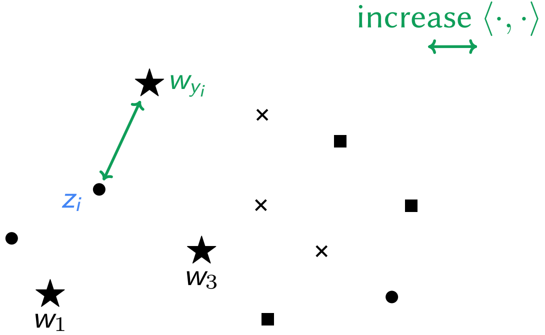


$\mathcal{L}_{\mathrm{CE}}(Z, W; Y)$

$$-\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\exp(\langle z_i, w_{y_i}\rangle)}{\sum_{c=1}^{K}\exp(\langle z_i, w_c\rangle)}\right)$$

$\Rightarrow$ Contribution of a **single** z_i

Cross-entropy (CE) loss $\mathcal{L}_{\mathrm{CE}}(Z, W; Y)$

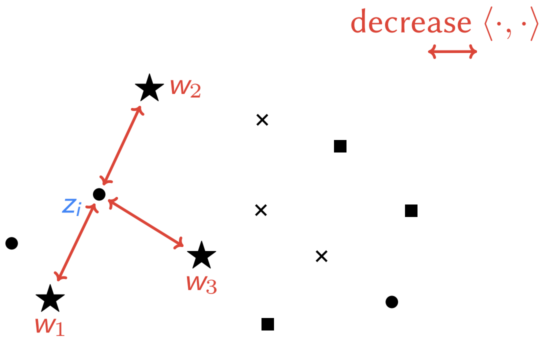


$\mathcal{L}_{\mathrm{CE}}(Z, W; Y)$

$$-\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\exp(\langle z_i, w_{y_i}\rangle)}{\sum_{c=1}^{K}\exp(\langle z_i, w_c\rangle)}\right)$$

$\Rightarrow$ Contribution of a **single** z_i

Cross-entropy (CE) loss $\mathcal{L}_{\mathrm{CE}}(Z, W; Y)$



$\mathcal{L}_{\mathrm{CE}}(Z, W; Y)$

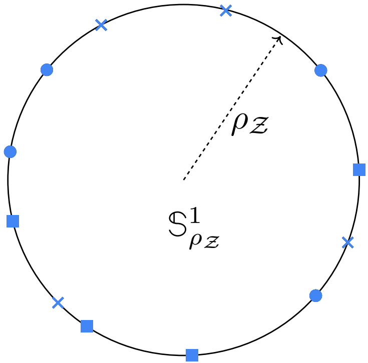
$$-\frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{\exp(\langle z_i, w_{y_i} \rangle)}{\sum_{c=1}^{K} \exp(\langle z_i, w_c \rangle)} \right)$$

$\Rightarrow$ Contribution of a **single** z_i

Supervised contrastive (SC) loss $\mathcal{L}_{\text{SC}}(Z; Y)$

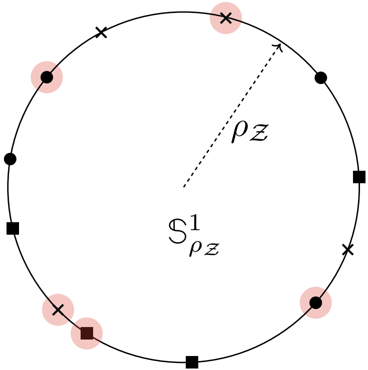
$\mathcal{L}_{\text{SC}}(Z; Y)$
$-\sum_{B \in \mathcal{B}} \sum_{i \in B} \frac{\mathbb{1}_{\lvert B_{y_i} \rvert > 1}}{\lvert B_{y_i} \rvert - 1} \sum_{j \in B_{y_i} \setminus \{\!\{ i \}\!\}} \log \left(\frac{\exp(\langle z_i, z_j \rangle)}{\sum_{k \in B \setminus \{\!\{ i \}\!\}} \exp(\langle z_i, z_k \rangle)} \right)$

Supervised contrastive (SC) loss $\mathcal{L}_{\mathrm{SC}}(Z; Y)$



$$\mathcal{L}_{\mathrm{SC}}(Z; Y)$$

$$-\sum_{B\in\mathcal{B}}\sum_{i\in B}\frac{\mathbb{1}_{|B_{y_i}|>1}}{|B_{y_i}|-1}\sum_{j\in B_{y_i}\setminus\{\!\{i\}\!\}}\log\left(\frac{\exp(\langle z_i,z_j\rangle)}{\sum_{k\in B\setminus\{\!\{i\}\!\}}\exp(\langle z_i,z_k\rangle)}\right)$$

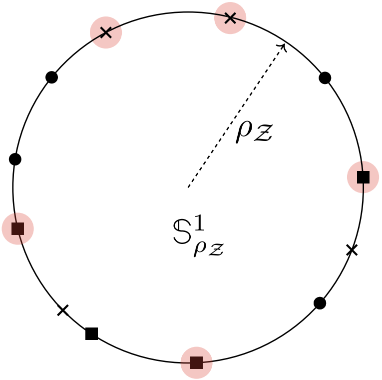


$\mathcal{L}_{\mathrm{SC}}(Z;Y)$

$$-\sum_{B\in\mathcal{B}}\sum_{i\in B}\frac{\mathbb{1}_{|B_{y_i}|>1}}{|B_{y_i}|-1}\sum_{j\in B_{y_i}\setminus\{\!\{i\}\!\}}\log\left(\frac{\exp(\langle z_i,z_j\rangle)}{\sum_{k\in B\setminus\{\!\{i\}\!\}}\exp(\langle z_i,z_k\rangle)}\right)$$

$\Rightarrow$ Contribution of **all** batches of size b

Supervised contrastive (SC) loss $\mathcal{L}_{\mathrm{SC}}(Z;Y)$

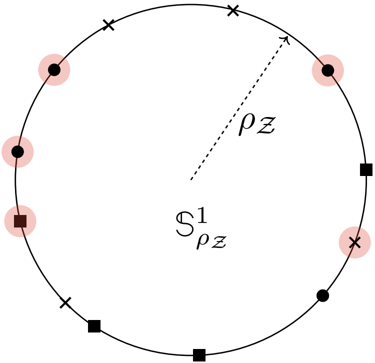


$\mathcal{L}_{\mathrm{SC}}(Z;Y)$

$$-\sum_{B\in\mathcal{B}}\sum_{i\in B}\frac{\mathbb{1}_{|B_{y_i}|>1}}{|B_{y_i}|-1}\sum_{j\in B_{y_i}\setminus\{\!\{i\}\!\}}\log\left(\frac{\exp(\langle z_i,z_j\rangle)}{\sum_{k\in B\setminus\{\!\{i\}\!\}}\exp(\langle z_i,z_k\rangle)}\right)$$

$\Rightarrow$ Contribution of **all** batches of size b

Supervised contrastive (SC) loss $\mathcal{L}_{\mathrm{SC}}(Z; Y)$

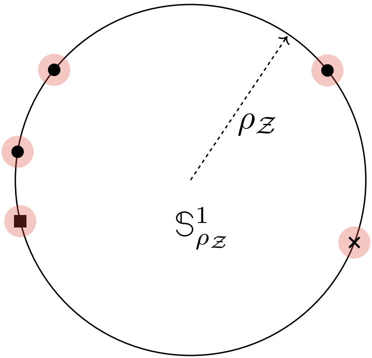


$$\mathcal{L}_{\mathrm{SC}}(Z; Y)$$

$$-\sum_{B\in\mathcal{B}}\sum_{i\in B}\frac{\mathbb{1}_{|B_{y_i}|>1}}{|B_{y_i}|-1}\sum_{j\in B_{y_i}\setminus\{\!\{i\}\!\}}\log\left(\frac{\exp(\langle z_i, z_j\rangle)}{\sum_{k\in B\setminus\{\!\{i\}\!\}}\exp(\langle z_i, z_k\rangle)}\right)$$

$\Rightarrow$ Contribution of **all** batches of size b

Supervised contrastive (SC) loss $\mathcal{L}_{\mathrm{SC}}(Z; Y)$

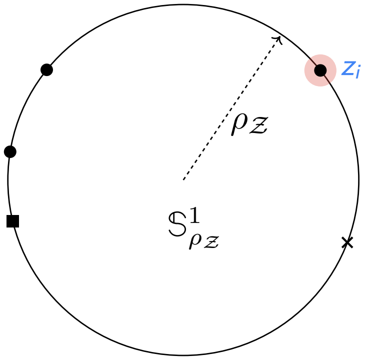


$\mathcal{L}_{\mathrm{SC}}(Z; Y)$

$$-\sum_{B\in\mathcal{B}} \sum_{i\in B} \frac{\mathbb{1}_{|B_{y_i}|>1}}{|B_{y_i}|-1} \sum_{j\in B_{y_i}\setminus\{\!\{i\}\!\}} \log\left(\frac{\exp(\langle z_i, z_j\rangle)}{\sum_{k\in B\setminus\{\!\{i\}\!\}} \exp(\langle z_i, z_k\rangle)}\right)$$

$\Rightarrow$ Contribution of **all** instances in batch

Supervised contrastive (SC) loss $\mathcal{L}_{\mathrm{SC}}(Z;Y)$

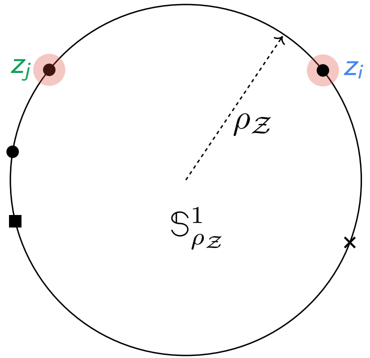


$$\mathcal{L}_{\mathrm{SC}}(Z;Y)$$

$$-\sum_{B\in\mathcal{B}}\sum_{i\in B}\frac{\mathbb{1}_{|B_{y_i}|>1}}{|B_{y_i}|-1}\sum_{j\in B_{y_i}\setminus\{\!\{i\}\!\}}\log\left(\frac{\exp(\langle z_i,z_j\rangle)}{\sum_{k\in B\setminus\{\!\{i\}\!\}}\exp(\langle z_i,z_k\rangle)}\right)$$

$\Rightarrow$ Contribution of **single** instance in batch

Supervised contrastive (SC) loss $\mathcal{L}_{\mathrm{SC}}(Z; Y)$

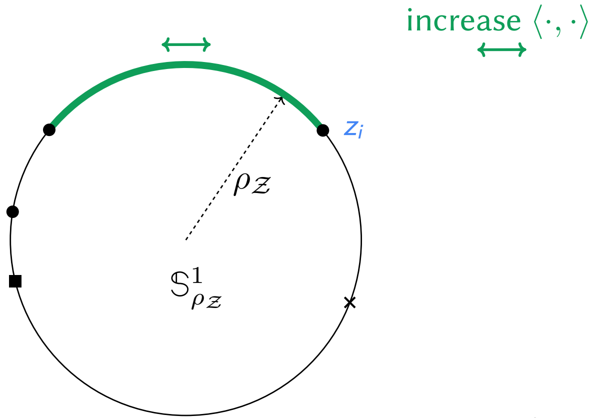


$\mathcal{L}_{\mathrm{SC}}(Z; Y)$

$$-\sum_{B\in\mathcal{B}}\sum_{i\in B}\frac{\mathbb{1}_{|B_{y_i}|>1}}{|B_{y_i}|-1}\sum_{j\in B_{y_i}\setminus\{\!\{i\}\!\}}\log\left(\frac{\exp(\langle z_i,z_j\rangle)}{\sum_{k\in B\setminus\{\!\{i\}\!\}}\exp(\langle z_i,z_k\rangle)}\right)$$

$\Rightarrow$ Contribution of **single** instance in batch

Supervised contrastive (SC) loss $\mathcal{L}_{\mathrm{SC}}(Z; Y)$

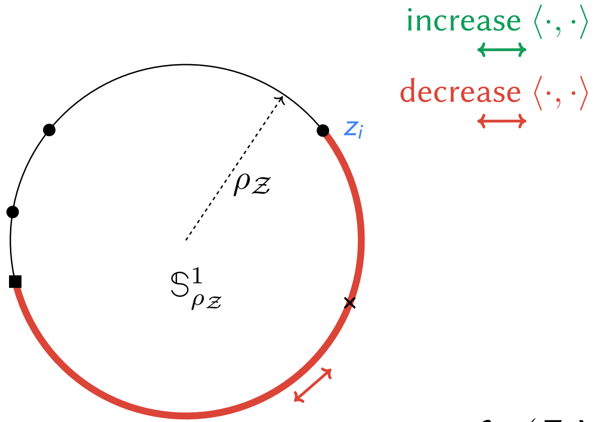


$\mathcal{L}_{\mathrm{SC}}(Z; Y)$

$$-\sum_{B \in \mathcal{B}} \sum_{i \in B} \frac{\mathbb{1}_{|B_{y_i}|>1}}{|B_{y_i}|-1} \sum_{j \in B_{y_i} \setminus \{\!\{ i \}\!\}} \log \left(\frac{\exp(\langle z_i, z_j \rangle)}{\sum_{k \in B \setminus \{\!\{ i \}\!\}} \exp(\langle z_i, z_k \rangle)} \right)$$

$\Rightarrow$ Contribution of **single** instance in batch

Supervised contrastive (SC) loss $\mathcal{L}_{\mathrm{SC}}(Z; Y)$

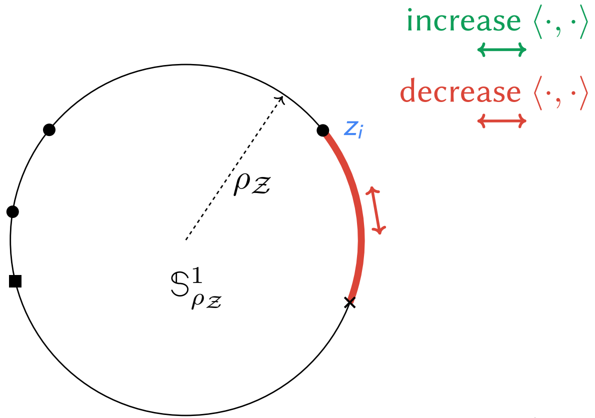


$$\mathcal{L}_{\mathrm{SC}}(Z; Y)$$

$$-\sum_{B\in\mathcal{B}}\sum_{i\in B}\frac{\mathbb{1}_{|B_{y_i}|>1}}{|B_{y_i}|-1}\sum_{j\in B_{y_i}\setminus\{\!\{i\}\!\}}\log\left(\frac{\exp(\langle z_i,z_j\rangle)}{\sum_{k\in B\setminus\{\!\{i\}\!\}}\exp(\langle z_i,z_k\rangle)}\right)$$

$\Rightarrow$ Contribution of **single** instance in batch

Supervised contrastive (SC) loss $\mathcal{L}_{SC}(Z; Y)$

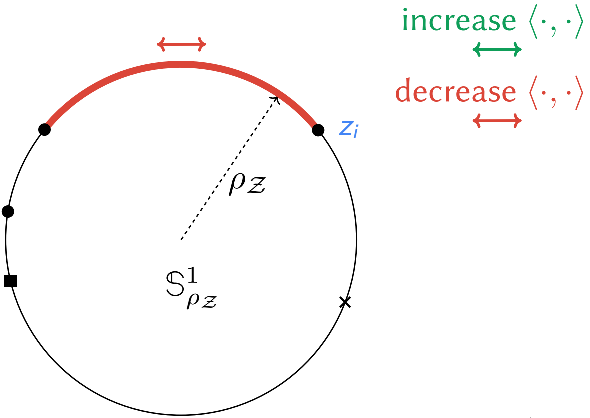


$$\mathcal{L}_{SC}(Z; Y)$$

$$-\sum_{B \in \mathcal{B}} \sum_{i \in B} \frac{\mathbb{1}_{|B_{y_i}|>1}}{|B_{y_i}|-1} \sum_{j \in B_{y_i} \setminus \{\!\{ i \}\!\}} \log \left(\frac{\exp(\langle z_i, z_j \rangle)}{\sum_{k \in B \setminus \{\!\{ i \}\!\}} \exp(\langle z_i, z_k \rangle)} \right)$$

$\Rightarrow$ Contribution of **single** instance in batch

Supervised contrastive (SC) loss $\mathcal{L}_{\mathrm{SC}}(Z; Y)$

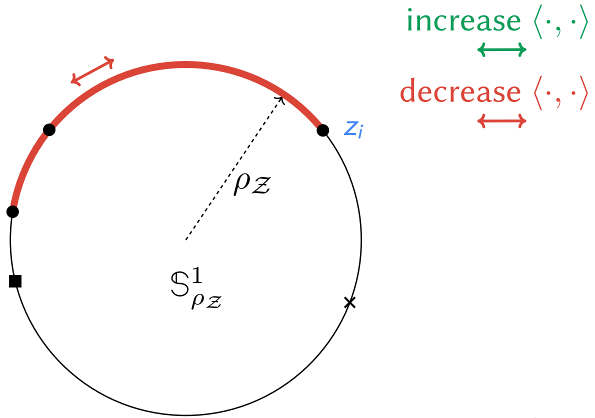


$\mathcal{L}_{\mathrm{SC}}(Z; Y)$

$$-\sum_{B \in \mathcal{B}} \sum_{i \in B} \frac{\mathbb{1}_{|B_{y_i}|>1}}{|B_{y_i}|-1} \sum_{j \in B_{y_i} \setminus \{\!\{ i \}\!\}} \log \left(\frac{\exp(\langle z_i, z_j \rangle)}{\sum_{k \in B \setminus \{\!\{ i \}\!\}} \exp(\langle z_i, z_k \rangle)} \right)$$

$\Rightarrow$ Contribution of **single** instance in batch

Supervised contrastive (SC) loss $\mathcal{L}_{\mathrm{SC}}(Z; Y)$

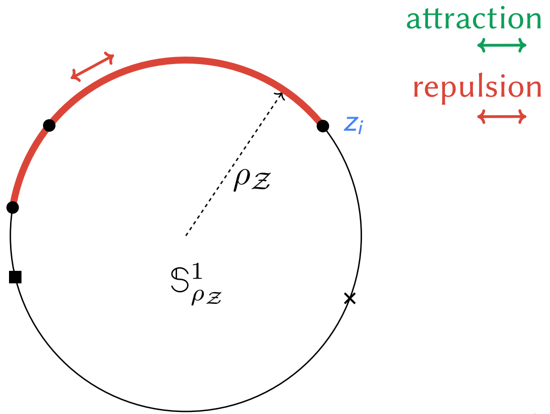


$\mathcal{L}_{\mathrm{SC}}(Z; Y)$

$$-\sum_{B\in\mathcal{B}}\sum_{i\in B}\frac{\mathbb{1}_{|B_{y_i}|>1}}{|B_{y_i}|-1}\sum_{j\in B_{y_i}\setminus\{\!\{i\}\!\}}\log\left(\frac{\exp(\langle z_i,z_j\rangle)}{\sum_{k\in B\setminus\{\!\{i\}\!\}}\exp(\langle z_i,z_k\rangle)}\right)$$

$\Rightarrow$ Contribution of **single** instance in batch

Supervised contrastive (SC) loss $\mathcal{L}_{\mathrm{SC}}(Z; Y)$



$$\mathcal{L}_{\mathrm{SC}}(Z; Y)$$

$$-\sum_{B\in\mathcal{B}}\sum_{i\in B}\frac{\mathbb{1}_{|B_{y_i}|>1}}{|B_{y_i}|-1}\sum_{j\in B_{y_i}\setminus\{\!\{i\}\!\}}\log\left(\frac{\exp(\langle z_i,z_j\rangle)}{\sum_{k\in B\setminus\{\!\{i\}\!\}}\exp(\langle z_i,z_k\rangle)}\right)$$

$\Rightarrow$ Contribution of **single** instance in batch

Problem statement - Revisited

Question

How do the representations (z_i) arrange?

Goal: compare $\underset{z_1,\ldots,z_N}{\operatorname{argmin}} \mathcal{L}_{\mathrm{SC}}(Z; Y)$ vs. $\underset{z_1,\ldots,z_N,W}{\operatorname{argmin}} \mathcal{L}_{\mathrm{CE}}(Z, W; Y)$

Problem statement - Revisited

Question

How do the representations (z_i) arrange?

Goal: compare $\underset{z_1,\dots,z_N}{\operatorname{argmin}} \mathcal{L}_{\mathrm{SC}}(Z; Y)$ vs. $\underset{z_1,\dots,z_N,W}{\operatorname{argmin}} \mathcal{L}_{\mathrm{CE}}(Z, W; Y)$

(Regular simplex) conjecture

Problem statement - Revisited

Question

How do the representations (z_i) arrange?

Goal: compare $\underset{z_1,\ldots,z_N}{\operatorname{argmin}} \mathcal{L}_{\mathrm{SC}}(Z; Y)$ vs. $\underset{z_1,\ldots,z_N, W}{\operatorname{argmin}} \mathcal{L}_{\mathrm{CE}}(Z, W; Y)$

(Regular simplex) conjecture

- Classes collapse to a point

Problem statement - Revisited

Question

How do the representations (z_i) arrange?

Goal: compare $\underset{z_1,\ldots,z_N}{\operatorname{argmin}}\, \mathcal{L}_{\mathrm{SC}}(Z; Y)$ vs. $\underset{z_1,\ldots,z_N, W}{\operatorname{argmin}}\, \mathcal{L}_{\mathrm{CE}}(Z, W; Y)$

(Regular simplex) conjecture

- Classes collapse to a point
- These points are maximally separated

Problem statement - Revisited

Question

How do the representations (z_i) arrange?

Goal: compare $\underset{z_1,\dots,z_N}{\operatorname{argmin}} \mathcal{L}_{\mathrm{SC}}(Z; Y)$ vs. $\underset{z_1,\dots,z_N, W}{\operatorname{argmin}} \mathcal{L}_{\mathrm{CE}}(Z, W; Y)$

(Regular simplex) conjecture

- Classes collapse to a point
- These points are maximally separated

Problem statement - Revisited

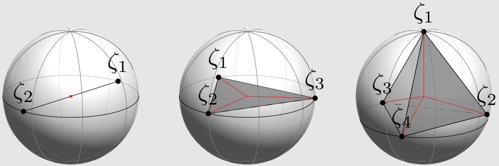
Question

How do the representations (z_i) arrange?

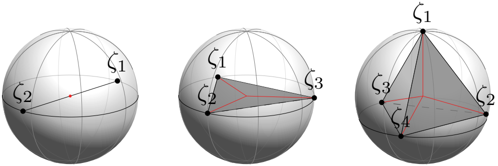
Goal: compare $\underset{z_1,\dots,z_N}{\operatorname{argmin}} \mathcal{L}_{\mathrm{SC}}(Z; Y)$ vs. $\underset{z_1,\dots,z_N,W}{\operatorname{argmin}} \mathcal{L}_{\mathrm{CE}}(Z, W; Y)$

(Regular simplex) conjecture

- Classes collapse to a point
- These points are maximally separated



Recap: Regular simplex

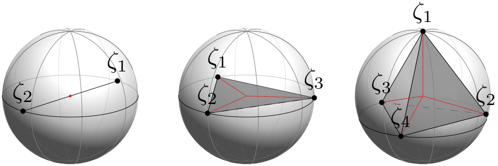


$\zeta_1, \ldots, \zeta_K \in \mathbb{R}^h$ form (the vertices of) an

origin-centered regular simplex inscribed in $\mathbb{S}^{h-1}_{\rho}$

if and only if the following conditions hold:

Recap: Regular simplex



$\zeta_1, \ldots, \zeta_K \in \mathbb{R}^h$ form (the vertices of) an

origin-centered regular simplex inscribed in $\mathbb{S}_\rho^{h-1}$

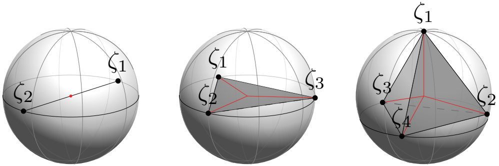
if and only if the following conditions hold:

S1. **origin-centered**: $\frac{1}{K}\sum_{i\in[K]} \zeta_i = 0$

S2. **sphere-inscribed**: $\|\zeta_i\| = \rho$ for $i \in [K]$

S3. **regular**: $\exists d \in \mathbb{R} : d = \|\zeta_i - \zeta_j\|$ for $1 \leq i < j \leq K$

Recap: Regular simplex



$\zeta_1, \ldots, \zeta_K \in \mathbb{R}^h$ form (the vertices of) an

origin-centered regular simplex inscribed in $\mathbb{S}_\rho^{h-1}$

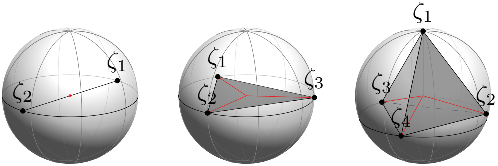
if and only if the following conditions hold:

S1. **origin-centered**: $\frac{1}{K} \sum_{i \in [K]} \zeta_i = 0$

S2. **sphere-inscribed**: $\|\zeta_i\| = \rho$ for $i \in [K]$

S3. **regular**: $\exists d \in \mathbb{R} : d = \|\zeta_i - \zeta_j\|$ for $1 \leq i < j \leq K$

Recap: Regular simplex



$\zeta_1, \ldots, \zeta_K \in \mathbb{R}^h$ form (the vertices of) an

origin-centered regular simplex inscribed in $\mathbb{S}_\rho^{h-1}$

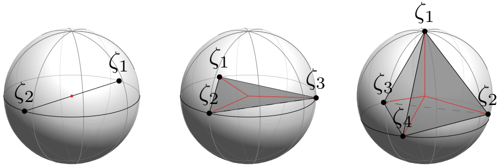
if and only if the following conditions hold:

S1. **origin-centered**: $\frac{1}{K}\sum_{i\in[K]} \zeta_i = 0$

S2. **sphere-inscribed**: $\|\zeta_i\| = \rho$ for $i \in [K]$

S3. **regular**: $\exists d \in \mathbb{R} : d = \|\zeta_i - \zeta_j\|$ for $1 \leq i < j \leq K$

Recap: Regular simplex



$\zeta_1, \ldots, \zeta_K \in \mathbb{R}^h$ form (the vertices of) an

origin-centered regular simplex inscribed in $\mathbb{S}_\rho^{h-1}$

if and only if the following conditions hold:

S1. **origin-centered**: $\frac{1}{K} \sum_{i \in [K]} \zeta_i = 0$

S2. **sphere-inscribed**: $\|\zeta_i\| = \rho$ for $i \in [K]$

S3. **regular**: $\exists d \in \mathbb{R} : d = \|\zeta_i - \zeta_j\|$ for $1 \leq i < j \leq K$

Proof Idea

Proof Idea

Bound the loss functions by a sequence of inequalities
(using Jensen, Cauchy-Schwarz)

Proof Idea

Bound the loss functions by a sequence of inequalities
(using Jensen, Cauchy-Schwarz)

loss

Proof Idea

Bound the loss functions by a sequence of inequalities
(using Jensen, Cauchy-Schwarz)

$$
\begin{aligned}
\textbf{loss} &\overset{(*)}{\geq} \dots \\
&\overset{(**)}{\geq} \dots \\
&\overset{(***)}{\geq}
\end{aligned}
$$

Proof Idea

Bound the loss functions by a sequence of inequalities
(using Jensen, Cauchy-Schwarz)

$$\begin{aligned}\textbf{loss} &\overset{(*)}{\geq} \ldots \\ &\overset{(**)}{\geq} \ldots \\ &\overset{(***)}{\geq} \textbf{tight}\text{ lower bound}\end{aligned}$$

Proof Idea

Bound the loss functions by a sequence of inequalities
(using Jensen, Cauchy-Schwarz)

$$\textbf{loss} \overset{(*)}{\geq} \dots$$
$$\overset{(**)}{\geq} \dots$$
$$\overset{(***)}{\geq} \textbf{tight}\text{ lower bound}$$

$$\boxed{\begin{array}{c}\textbf{necessary}\text{ and }\textbf{sufficient}\\ \text{equality conditions}\\ (*), (**), (***)\end{array}}$$

Proof Idea

Bound the loss functions by a sequence of inequalities
(using Jensen, Cauchy-Schwarz)

$$
\begin{aligned}
\textbf{loss} &\overset{(*)}{\geq} \dots \\
&\overset{(**)}{\geq} \dots \\
&\overset{(***)}{\geq} \textbf{tight} \text{ lower bound}
\end{aligned}
$$

Show that

necessary and **sufficient** equality conditions $(*), (**), (***)$	$\Leftrightarrow$	simplex conditions (S1), (S2), (S3)

- Loss function is not **sample-wise** but **batch-wise**

$$\mathcal{L}_{\mathrm{SC}}(Z;Y)$$

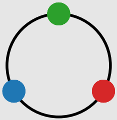
$$-\sum_{B\in\mathcal{B}}\sum_{i\in B}\frac{\mathbb{1}_{|B_{y_i}|>1}}{|B_{y_i}|-1}\sum_{j\in B_{y_i}\setminus\{\!\{i\}\!\}}\log\left(\frac{\exp(\langle z_i,z_j\rangle)}{\sum_{k\in B\setminus\{\!\{i\}\!\}}\exp(\langle z_i,z_k\rangle)}\right)$$

- Loss function is not sample-wise but batch-wise
- **No common minimizer** for all batch-wise contributions

$$\frac{\mathcal{L}_{\mathrm{SC}}(Z;Y)}{-\sum_{B\in\mathcal{B}}\sum_{i\in B}\frac{\mathbb{1}_{|B_{y_i}|>1}}{|B_{y_i}|-1}\sum_{j\in B_{y_i}\setminus\{\!\{i\}\!\}}\log\left(\frac{\exp(\langle z_i,z_j\rangle)}{\sum_{k\in B\setminus\{\!\{i\}\!\}}\exp(\langle z_i,z_k\rangle)}\right)}$$

- Loss function is not sample-wise but batch-wise
- **No common minimizer** for all batch-wise contributions

Example: Batch size **9**



Minimizer

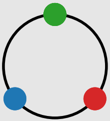
Label configuration

sc$(Z; Y)$

$$-\sum_{B\in\mathcal{B}}\sum_{i\in B}\frac{\mathbb{1}_{|B_{y_i}|>1}}{|B_{y_i}|-1}\sum_{j\in B_{y_i}\setminus\{\!\{i\}\!\}}\log\left(\frac{\exp(\langle z_i,z_j\rangle)}{\sum_{k\in B\setminus\{\!\{i\}\!\}}\exp(\langle z_i,z_k\rangle)}\right)$$

- Loss function is not sample-wise but batch-wise
- **No common minimizer** for all batch-wise contributions

Example: Batch size **9**



Minimizer

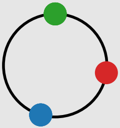
Label configuration

$\mathrm{sc}(Z; Y)$

$$-\sum_{B \in \mathcal{B}} \sum_{i \in B} \frac{\mathbb{1}_{|B_{y_i}|>1}}{|B_{y_i}|-1} \sum_{j \in B_{y_i} \setminus \{\!\{ i \}\!\}} \log \left(\frac{\exp(\langle z_i, z_j \rangle)}{\sum_{k \in B \setminus \{\!\{ i \}\!\}} \exp(\langle z_i, z_k \rangle)} \right)$$

- Loss function is not sample-wise but batch-wise
- **No common minimizer** for all batch-wise contributions

Example: Batch size 9



Minimizer

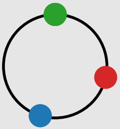
Label configuration

$\text{sc}(Z;Y)$

$$-\sum_{B\in\mathcal{B}}\sum_{i\in B}\frac{\mathbb{1}_{|B_{y_i}|>1}}{|B_{y_i}|-1}\sum_{j\in B_{y_i}\setminus\{\!\{i\}\!\}}\log\left(\frac{\exp(\langle z_i,z_j\rangle)}{\sum_{k\in B\setminus\{\!\{i\}\!\}}\exp(\langle z_i,z_k\rangle)}\right)$$

- Loss function is not sample-wise but batch-wise
- **No common minimizer** for all batch-wise contributions

Example: Batch size 9



Minimizer

Label configuration

sc$(Z; Y)$

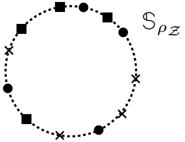
$$-\sum_{B\in\mathcal{B}}\sum_{i\in B}\frac{\mathbb{1}_{|B_{y_i}|>1}}{|B_{y_i}|-1}\sum_{j\in B_{y_i}\setminus\{i\}}\log\left(\frac{\exp(\langle z_i,z_j\rangle)}{\sum_{k\in B\setminus\{i\}}\exp(\langle z_i,z_k\rangle)}\right)$$

- Loss function is not sample-wise but batch-wise
- No common minimizer for all batch-wise contributions
- Attraction and repulsion forces depend on

all other representations in the batch!

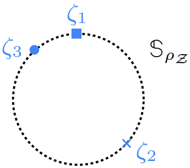
$$\frac{\mathcal{L}_{\mathrm{SC}}(Z;Y)}{-\sum_{B\in\mathcal{B}}\sum_{i\in B}\frac{\mathbb{1}_{|B_{y_i}|>1}}{|B_{y_i}|-1}\sum_{j\in B_{y_i}\setminus\{\!\{i\}\!\}}\log\left(\frac{\exp(\langle z_i,z_j\rangle)}{\sum_{k\in B\setminus\{\!\{i\}\!\}}\exp(\langle z_i,z_k\rangle)}\right)}$$

Theory



Theorem *Supervised Contrastive Loss*

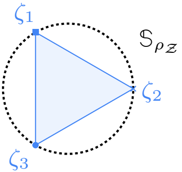
Let $\rho_{\mathcal{Z}} > 0$ and $\mathcal{Z} = \mathbb{S}^{h-1}_{\rho_{\mathcal{Z}}}$. If the labels Y are balanced, then $\mathcal{L}_{\mathrm{SC}}(Z; Y)$ is **minimal** if and only if there $\exists\, \zeta_1, \ldots, \zeta_K \in \mathbb{R}^h$ s.t.



Theorem *Supervised Contrastive Loss*

Let $\rho_{\mathcal{Z}} > 0$ and $\mathcal{Z} = \mathbb{S}^{h-1}_{\rho_{\mathcal{Z}}}$. If the labels Y are balanced, then $\mathcal{L}_{\mathrm{SC}}(Z; Y)$ is **minimal** if and only if there $\exists\ \zeta_1, \ldots, \zeta_K \in \mathbb{R}^h$ s.t.

1. The **classes collapse**: $\forall n \in [N] : z_n = \zeta_{y_n}$



Theorem *Supervised Contrastive Loss*

Let $\rho_{\mathcal{Z}} > 0$ and $\mathcal{Z} = \mathbb{S}^{h-1}_{\rho_{\mathcal{Z}}}$. If the labels Y are balanced, then $\mathcal{L}_{\text{SC}}(Z; Y)$ is **minimal** if and only if there $\exists\, \zeta_1, \dots, \zeta_K \in \mathbb{R}^h$ s.t.

1. The **classes collapse**: $\forall n \in [N] : z_n = \zeta_{y_n}$
2. The $\{\zeta_y\}_y$ form an

   origin centered **regular simplex**

   inscribed in the sphere $\mathbb{S}^{h-1}_{\rho_{\mathcal{Z}}}$ of radius $\rho_{\mathcal{Z}}$.

Theorem *Cross-Entropy*

Let $\mathcal{Z} = \{z \in \mathbb{R}^h : \ \|z\| \leq \rho_{\mathcal{Z}}\}$. If labels Y are balanced, then

$$\mathcal{L}_{\mathrm{CE}}(Z, W;\ Y) \geq \log\left(1 + (K-1)\exp\left(-\rho_{\mathcal{Z}}\frac{\sqrt{K}}{K-1}\|W\|_F\right)\right) \ ,$$

with equality **if and only if** $\exists\ \zeta_1, \ldots, \zeta_K \in \mathbb{R}^h$ s.t.

Theorem *Cross-Entropy*

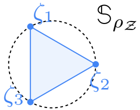
Let $\mathcal{Z} = \{z \in \mathbb{R}^h : \; \|z\| \leq \rho_{\mathcal{Z}}\}$. If labels Y are balanced, then

$$\begin{aligned} &\mathcal{L}_{\mathrm{CE}}(Z, W; \; Y) \\ &\geq \log\left(1 + (K-1)\exp\left(-\rho_{\mathcal{Z}}\frac{\sqrt{K}}{K-1}\|W\|_F\right)\right) \quad , \end{aligned}$$

with equality **if and only if** $\exists \; \zeta_1, \ldots, \zeta_K \in \mathbb{R}^h$ s.t.

1. The **classes collapse**: $\forall n \in [N] : z_n = \zeta_{y_n}$

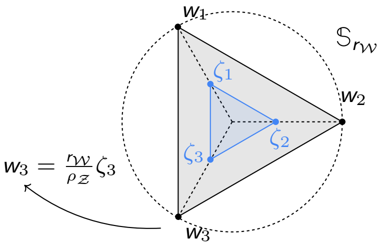
and



2. The $\{\zeta_y\}_y$ form an

origin-centered **regular simplex**

inscribed in the sphere $\mathbb{S}^{h-1}_{\rho_{\mathcal{Z}}}$ of radius $\rho_{\mathcal{Z}}$



2. The $\{\zeta_y\}_y$ form an

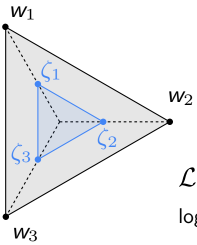
   origin-centered **regular simplex**

   inscribed in the sphere $\mathbb{S}^{h-1}_{\rho_{\mathcal{Z}}}$ of radius $\rho_{\mathcal{Z}}$

3. $\exists r_{\mathcal{W}} > 0$ s.t. the **weights** form an

   origin-centered **regular simplex**

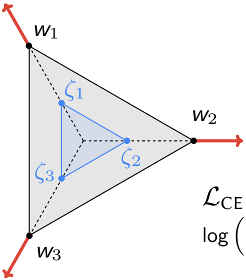
   inscribed in the sphere $\mathbb{S}^{h-1}_{r_{\mathcal{W}}}$ and aligned to the former.



$$\mathcal{L}_{\mathrm{CE}}(Z, W;\ Y) \geq \log\left(1 + (K-1)\exp\left(-\rho_{\mathcal{Z}}\frac{\sqrt{K}}{K-1}\|\boldsymbol{W}\|_{\boldsymbol{F}}\right)\right)$$

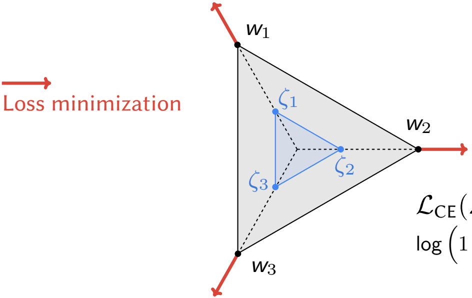
Theory

Loss minimization



$$\mathcal{L}_{\mathrm{CE}}(Z, W;\ Y) \geq \log\left(1 + (K-1)\exp\left(-\rho_{\mathcal{Z}}\frac{\sqrt{K}}{K-1}\underbrace{\|\boldsymbol{W}\|_F}_{=\sqrt{\sum_{i=1}^{K}\|\boldsymbol{w_i}\|^2}}\right)\right)$$

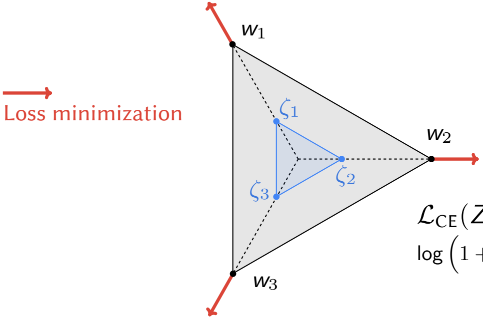
The lower bound **decreases** with $\|w_i\|$



$$\mathcal{L}_{\mathrm{CE}}(Z, W;\ Y) \geq \log\left(1 + (K-1)\exp\left(-\rho_{\mathcal{Z}} \frac{\sqrt{K}}{K-1} \|\boldsymbol{W}\|_{F}\right)\right)$$

The lower bound **decreases** with $\|w_i\|$

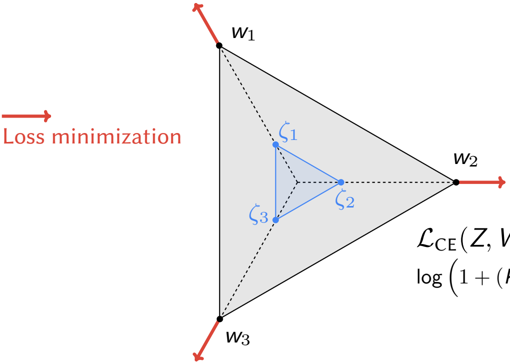
Theory



$$\mathcal{L}_{\mathrm{CE}}(Z, W;\ Y) \geq \log\left(1 + (K-1)\exp\left(-\rho_{\mathcal{Z}}\frac{\sqrt{K}}{K-1}\|\boldsymbol{W}\|_F\right)\right)$$

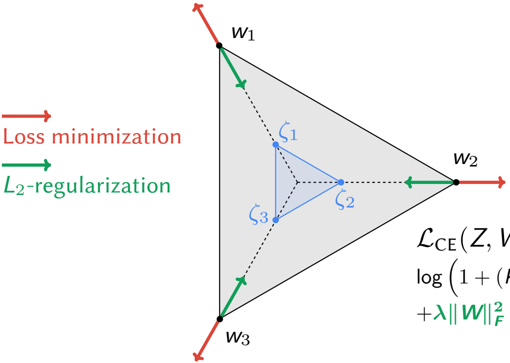
The lower bound **decreases** with $\|w_i\|$

Theory



$$\mathcal{L}_{\mathrm{CE}}(Z, W;\; Y) \geq \log\left(1 + (K-1)\exp\left(-\rho_{\mathcal{Z}}\frac{\sqrt{K}}{K-1}\|\boldsymbol{W}\|_F\right)\right)$$

The lower bound **decreases** with $\|w_i\|$

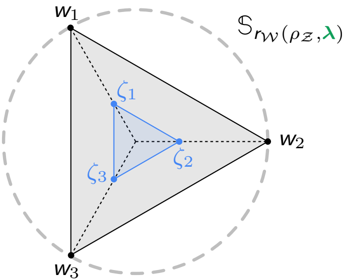


$$\mathcal{L}_{\text{CE}}(Z, W; Y) + \boldsymbol{\lambda} \|\boldsymbol{W}\|_F^2 \geq \log\left(1 + (K-1)\exp\left(-\rho_{\mathcal{Z}} \frac{\sqrt{K}}{K-1} \|\boldsymbol{W}\|_F\right)\right) + \boldsymbol{\lambda}\|\boldsymbol{W}\|_F^2$$

The lower bound **decreases** with $\|w_i\|$

Adding **L_2-regularization** with strength $\boldsymbol{\lambda}$ determines the w_i.

Theory



The lower bound **decreases** with $\|w_i\|$

Adding L_2**-regularization** with strength $\boldsymbol{\lambda}$ determines the w_i.

Corollary *L_2-Regularized Cross-Entropy*

The ***L_2*-regularized** cross-entropy loss $\mathcal{L}_{\mathrm{CE}}(Z, W;\ Y) + \lambda \| W \|_F^2$, is **minimal** if and only if $\exists\ \zeta_1, \ldots, \zeta_K \in \mathbb{R}^h$ such that:

Corollary *L_2-Regularized Cross-Entropy*

The ***L_2*-regularized** cross-entropy loss $\mathcal{L}_{\mathrm{CE}}(Z, W;\ Y) + \lambda \|W\|_F^2$, is **minimal** if and only if $\exists\ \zeta_1, \ldots, \zeta_K \in \mathbb{R}^h$ such that:

1. The **classes collapse**: $\forall n \in [N] : z_n = \zeta_{y_n}$

Corollary *L_2-Regularized Cross-Entropy*

The **L_2-regularized** cross-entropy loss $\mathcal{L}_{\mathrm{CE}}(Z, W;\ Y) + \lambda \|W\|_F^2$, is **minimal** if and only if $\exists\ \zeta_1, \ldots, \zeta_K \in \mathbb{R}^h$ such that:

1. The **classes collapse**: $\forall n \in [N] : z_n = \zeta_{y_n}$
2. The $\{\zeta_y\}_y$ form an

   origin-centered **regular simplex**

   inscribed in the sphere, $\mathbb{S}^{h-1}_{\rho_{\mathcal{Z}}}$ of radius $\rho_{\mathcal{Z}}$

Corollary *L_2-Regularized Cross-Entropy*

The **L_2-regularized** cross-entropy loss $\mathcal{L}_{\mathrm{CE}}(Z, W;\ Y) + \lambda \| W \|_F^2$, is **minimal** if and only if $\exists\ \zeta_1, \ldots, \zeta_K \in \mathbb{R}^h$ such that:

1. The **classes collapse**: $\forall n \in [N] : z_n = \zeta_{y_n}$
2. The $\{\zeta_y\}_y$ form an

   origin-centered **regular simplex**

   inscribed in the sphere, $\mathbb{S}^{h-1}_{\rho_{\mathcal{Z}}}$ of radius $\rho_{\mathcal{Z}}$
3. The **weights** form an

   origin-centered **regular simplex**

   inscribed in the sphere of radius $r_{\mathcal{W}}(\rho_{\mathcal{Z}}, \lambda)$ and aligned to the former[3].

[3] $r_{\mathcal{W}}(\rho_{\mathcal{Z}}, \lambda)$ is the solution of $2\lambda r_{\mathcal{W}}(e^{\frac{K}{K-1}\rho_{\mathcal{Z}} r_{\mathcal{W}}} + K - 1) - \rho_{\mathcal{Z}} = 0$

Numerical Simulation

Lets verify the theory on a **toy example** . . .

Numerical Simulation

Lets verify the theory on a **toy example** . . .

1. Sample $z_1, \ldots, z_{400}$ (with $K = 4$ classes) on the sphere $\mathbb{S}^2$

Numerical Simulation

Lets verify the theory on a **toy example** ...

1. Sample $z_1, \ldots, z_{400}$ (with $K = 4$ classes) on the sphere $\mathbb{S}^2$
2. Minimize (L_2-regularized) **CE** and **SC**, respectively

Numerical Simulation

Lets verify the theory on a **toy example** ...

1. Sample $z_1, \ldots, z_{400}$ (with $K = 4$ classes) on the sphere $\mathbb{S}^2$
2. Minimize (L_2-regularized) **CE** and **SC**, respectively
3. Ensure that boundary conditions are fulfilled

Experiments

Question

How good is the **simplex arrangement** achieved in practice?

Experiments

Question

How good is the **simplex arrangement** achieved in practice?

Model / Dataset: ResNet-18 / CIFAR100

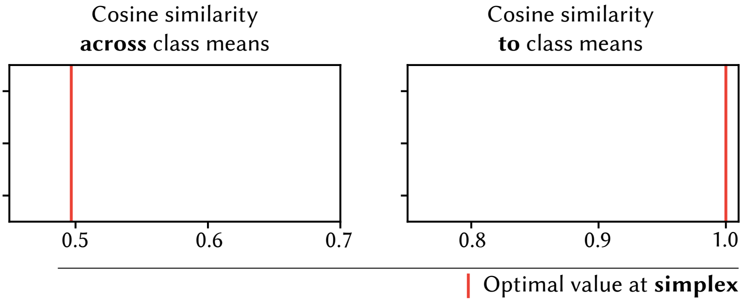
Statistics: Cosine similarities from **training** representations

Question

How good is the **simplex arrangement** achieved in practice?

Model / Dataset: ResNet-18 / CIFAR100

Statistics: Cosine similarities from **training** representations



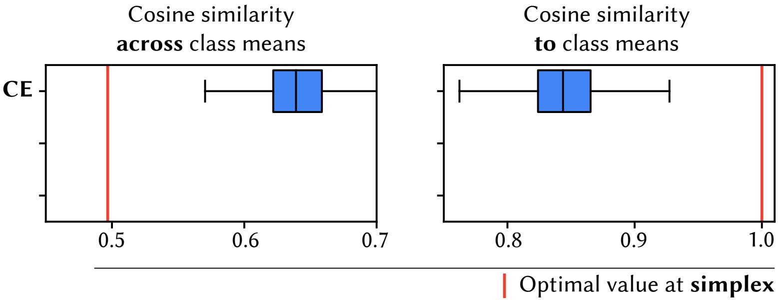
Experiments

Question

How good is the **simplex arrangement** achieved in practice?

Model / Dataset: ResNet-18 / CIFAR100

Statistics: Cosine similarities from **training** representations

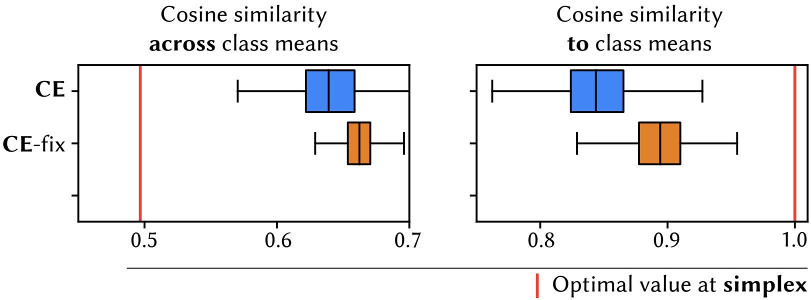


Question

How good is the **simplex arrangement** achieved in practice?

Model / Dataset: ResNet-18 / CIFAR100

Statistics: Cosine similarities from **training** representations



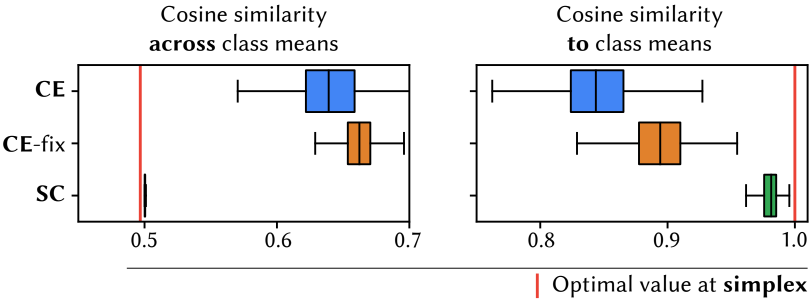
CE-fix [Mettes et al., NeurIPS '19] (*weights at simplex by construction*)

Question

How good is the **simplex arrangement** achieved in practice?

Model / Dataset: ResNet-18 / CIFAR100

Statistics: Cosine similarities from **training** representations



CE-fix [Mettes et al., NeurIPS '19] (*weights at simplex by construction*)

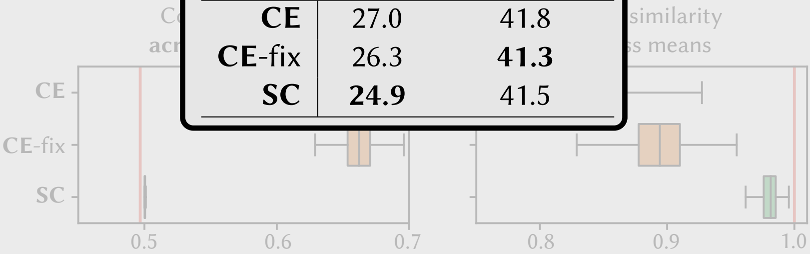
Question

How good is the **simplex arrangement** achieved in practice?

Model / Dataset: ResNet-18 / CIFAR100

Statistics: Co … resentations

Loss	**CIFAR100** err. [%] w aug.	w/o aug.
CE	27.0	41.8
CE-fix	26.3	**41.3**
SC	**24.9**	41.5



| Optimal value at **simplex**

CE-fix [Mettes et al., NeurIPS '19] (*weights at simplex by construction*)

Experiments

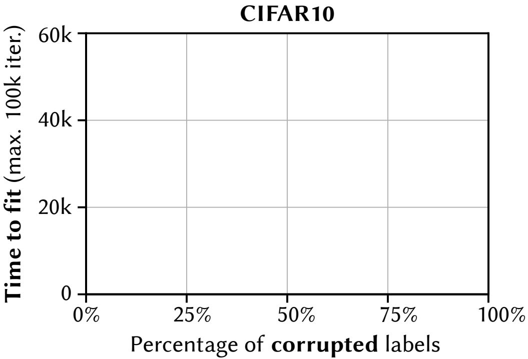
- Both losses lead to a **close-to-simplex** solution
- **SC** reaches this loss-optimal state more closely
- Results indicate **close-to-simplex** $\Rightarrow$ **lower error**

An interesting, final observation

Repeat the **random label** exp. of [Zhang et al., ICLR ’17]

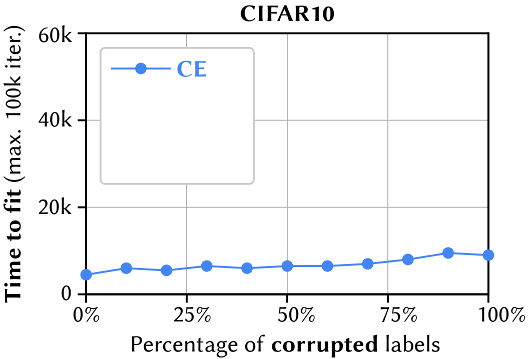
An interesting, final observation

Repeat the **random label** exp. of [Zhang et al., ICLR '17]



An interesting, final observation

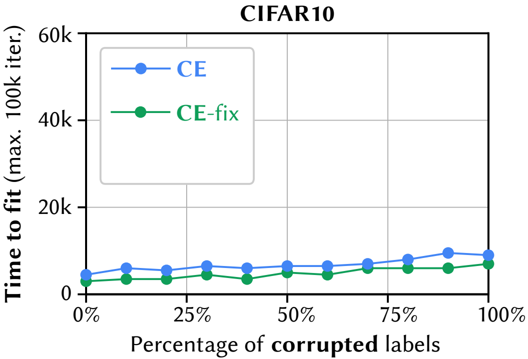
Repeat the **random label** exp. of [Zhang et al., ICLR '17]



- For **CE**, roughly **linear** scaling (as reported previously)

An interesting, final observation

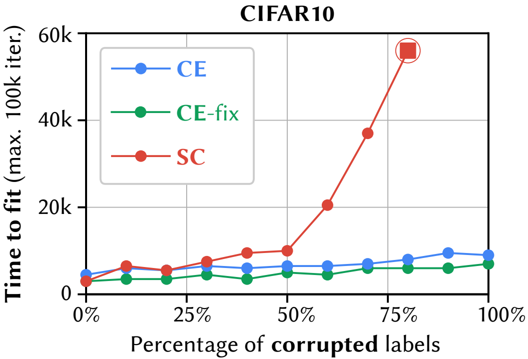
Repeat the **random label** exp. of [Zhang et al., ICLR '17]



- For **CE**, roughly **linear** scaling (as reported previously)

An interesting, final observation

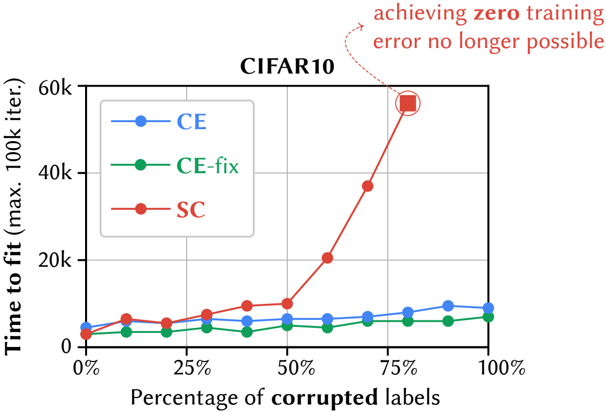
Repeat the **random label** exp. of [Zhang et al., ICLR '17]



- For **CE**, roughly **linear** scaling (as reported previously)
- For **SC**, we observe **superlinear** scaling behavior

An interesting, final observation

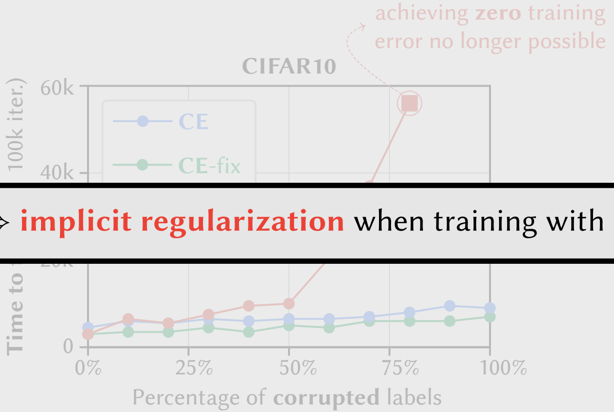
Repeat the **random label** exp. of [Zhang et al., ICLR '17]



- For **CE**, roughly **linear** scaling (as reported previously)
- For **SC**, we observe **superlinear** scaling behavior

An interesting, final observation

Repeat the **random label** exp. of [Zhang et al., ICLR '17]



⇒ **implicit regularization** when training with SC

- For **CE**, roughly **linear** scaling (as reported previously)
- For **SC**, we observe **superlinear** scaling behavior

Summary

Summary

- **Theory** shows,

  training models with **CE** and **SC** strives
  for the **same** arrangement of representations

Summary

- **Theory** shows,

  training models with **CE** and **SC** strives
  for the **same** arrangement of representations

- **Empirically**,

  models trained with **CE** and **SC** behave **differently**

Summary

- **Theory** shows,

  training models with **CE** and **SC** strives
  for the **same** arrangement of representations

- **Empirically**,

  models trained with **CE** and **SC** behave **differently**

$\Rightarrow$ This is caused by **differing optimization** dynamics

Summary

- **Theory** shows,

  training models with **CE** and **SC** strives
  for the **same** arrangement of representations

- **Empirically**,

  models trained with **CE** and **SC** behave **differently**

$\Rightarrow$ This is caused by **differing optimization** dynamics

Probably, rooted in the **interaction terms** among representations in the **SC** loss function

Open Questions

Open Questions

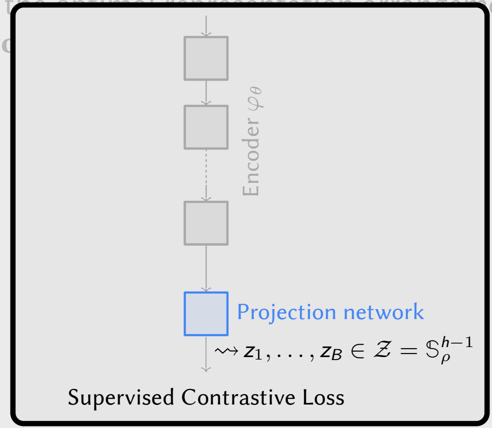
- Is the **powerful enough encoder** assumption justified?

Open Questions

- Is the **powerful enough encoder** assumption justified?
- How is the optimal representation arrangement affected by a **projection network**?

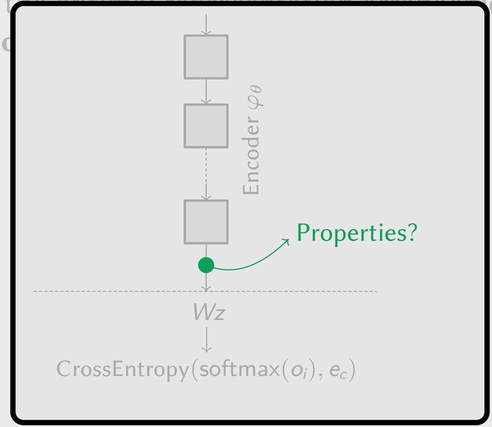
Open Questions

- Is the **powerful enough encoder** assumption justified?
- How is t ent affected by a **pro**



Open Questions

- Is the **powerful enough encoder** assumption justified?
- How is t ent affected by a **pro**



Open Questions

- Is the **powerful enough encoder** assumption justified?
- How is the optimal representation arrangement affected by a **projection network**?
- Why does **SC** “prevent” to easily fit to **random labels**?

Thank You!

Source code available here: