About contrastive unsupervised representation learning for classification and its convergence

Ibrahim Merad, Yiyang Yu, Emmanuel Bacry, Stéphane Gaïffas

About contrastive unsupervised representation learning for classification and its convergence

Ibrahim Merad* Yiyang Yu† Emmanuel Bacry‡ Stéphane Gaïffas§¶

December 3, 2020

Abstract

Contrastive representation learning has been recently proved to be very efficient for self-supervised training. These methods have been successfully used to train encoders which perform comparably to supervised training on downstream classification tasks. A few works have started to build a theoretical framework around contrastive learning in which guarantees for its performance can be proven. We provide extensions of these results to training with multiple negative samples and for multiway classification. Furthermore, we provide convergence guarantees for the minimization of the contrastive training error with gradient descent of an overparametrized deep neural encoder, and provide some numerical experiments that complement our theoretical findings.

Keywords. Unsupervised Learning · Contrastive Learning · Deep Neural Networks · Theoretical guarantees

1 Introduction

The aim of this work is to provide additional theoretical guarantees for *contrastive learning* (van den Oord et al., 2018), which corresponds to methods allowing to learn useful data representations in an *unsupervised* setting. Unsupervised representation learning was initially approached with a fair amount of success by training through the minimization of losses coming from "pretext" tasks, a technique known as *self-supervision* (Doersch and Zisserman, 2017), where labels can be automatically constructed. Notable examples of pretext tasks in computer vision include colorization (Zhang et al., 2016), transformation prediction (Gidaris et al., 2018; Dosovitskiy et al., 2014) or predicting patch relative positions (Doersch et al., 2015). Some theoretical guarantees (Lee et al., 2020) were recently proposed to support training on pretext tasks.

Contrastive learning is also known to be very effective for pretraining supervised methods (Chen et al., 2020a,b; Grill et al., 2020; Caron et al., 2020), where we can observe that, quite surprisingly, the gap between unsupervised and supervised performance has been closed for tasks such as image classification: the use of a pretrained image encoder on top of simple classification layers, that are trained on a fraction of the labels available, allows to achieve an accuracy comparable to that of a fully supervised end-to-end training (Hénaff et al., 2019; Grill et al., 2020). Contrastive methods show also strong success in natural language processing (Logeswaran and Lee, 2018; Mikolov

*LPSM, UMR 8001, Université de Paris, Paris, France, imerad@lpsm.paris
†LPSM, UMR 8001, Université de Paris, Paris, France, yyu@lpsm.paris
‡CEREMADE, Université Paris-Dauphine, bacry@ceremade.dauphine.fr
§LPSM, UMR 8001, Université de Paris, Paris, France, gaiffas@lpsm.paris
¶DMA, CNRS UMR 8553, Ecole normale supérieure, Paris, France

et al., 2013; Devlin et al., 2018; van den Oord et al., 2018), video classification (Sun et al., 2019), reinforcement learning (Srinivas et al., 2020) and time-series (Franceschi et al., 2019).

Although the papers cited above introduce methods with considerable variations, they mostly agree on the following basic pretraining approach: provided a dataset, an encoder is trained using a contrastive loss whose minimization allows to learn embeddings that are *similar* for pairs of samples (called the *positives*) that are close to each other (such as pairs of random data augmentations of the same image, see He et al. (2020); Chen et al. (2020a)), while such embeddings are *contrasted* for dissimilar pairs (called the *negatives*).

However, despite growing efforts (Saunshi et al., 2019; Wang and Isola, 2020), as of today, few theoretical results have been obtained. For instance, there is still no clear theoretical explanation of how a supervised task could benefit from an upstream unsupervised pretraining phase, or of what could be the theoretical guarantees for the convergence of the minimization procedure of the contrastive loss during this pretraining phase. Getting some answers to these questions would undoubtedly be a step towards a better theoretical understanding of contrastive representation learning.

Our contributions in this paper are twofold. In Section 3, we provide new theoretical guarantees for the classification performance of contrastively trained models in the case of multiway classification tasks, using *multiple* negative samples. We extend results from Saunshi et al. (2019) to show that unsupervised training performance reflects on a subsequent classification task in the case of multiple tasks and when a high number of negative samples is used. In Section 4, we prove a convergence result for an *explicit* algorithm (gradient descent), when training overparametrized deep neural network for unsupervised contrastive representation learning. We explain how results from Allen-Zhu et al. (2019) about training convergence of overparametrized deep neural networks can be applied to a contrastive learning objective. The results and major assumptions of both Sections 3 and 4 are illustrated in Section 5 through experiments on a few simple datasets.

2 Related work

A growing literature attempts to build a theoretical framework around contrastive learning and to provide justifications for its success beyond intuitive ideas. In Saunshi et al. (2019) a formalism is proposed together with results on classification performance based on unsupervisedly learned representation. However, these results do not explain the performance gain that is observed empirically (Chen et al., 2020a; He et al., 2020) when a high number of negative samples are used, while the results proposed in Section 3 below hold for an arbitrary large number of negatives (and decoupled from the number of classification tasks). A more recent work (Wang and Isola, 2020) emphasizes the two tendencies encouraged by the contrastive loss: the encoder's outputs are incentivized to spread evenly on the unit hypersphere, and encodings of same-class samples are driven close to each other while those of different classes are driven apart. Interestingly, this work also shows how the tradeoff between these two aspects can be controlled, by introducing weight factors in the loss leading to improved performance. Chuang et al. (2020) considers the same setting as Saunshi et al. (2019) and addresses the bias problem that comes from collisions between positive and negative sampling in the unsupervised constrastive loss. They propose to simulate unbiased negative sampling by assuming, among other things, extra access to positive sampling. However, one has to keep in mind that an excessive access to positive sampling gets the setting closer to that of supervised learning.

In a direction that is closer to the result proposed in Section 4 below, Wen (2020) provides a theoretical guarantee on the training convergence of gradient descent for an overparametrized model that is trained with an unsupervised contrastive loss, using earlier works by Allen-Zhu et al. (2019). However, two separate encoders are considered instead of a single one: one for the query,

which corresponds to a sample from the dataset, and one for the (positive and negative) samples to compare the query to. In this setting, it is rather unclear how the two resulting encoders are to be used for downstream classification. In Section 4 below, we explain how the results from Allen-Zhu et al. (2019) can be used for the more realistic setting of a single encoder, by introducing a reasonable assumption on the encoder outputs.

3 Unsupervised training improves supervised performance

In this section, we provide new results in the setting previously considered in Saunshi et al. (2019). We assume that data are distributed according to a finite set $\mathcal{C}$ of latent classes, and denote $N_{\mathcal{C}} = \mathrm{card}(\mathcal{C})$ its cardinality. Let ρ be a discrete distribution over $\mathcal{C}$ that is such that

$$\sum_{c \in \mathcal{C}} \rho(c) = 1 \quad \text{and} \quad \rho(c) > 0$$

for all $c \in \mathcal{C}$. We denote $\mathcal{D}_c$ a distribution over the feature space $\mathcal{X}$ from a class $c \in \mathcal{C}$. In order to perform unsupervised contrastive training, on the one hand we assume that we can sample *positive* pairs (x, x^+) from the distribution

$$\mathcal{D}_{\mathrm{sim}}(x, x^+) = \sum_{c \in \mathcal{C}} \rho(c) \mathcal{D}_c(x) \mathcal{D}_c(x^+), \tag{1}$$

namely, (x, x^+) is sampled as a mixture of independent pairs conditionally to a shared latent class, sampled according to ρ. On the other hand, we assume that we can sample *negative* samples x^- from the distribution

$$\mathcal{D}_{\mathrm{neg}}(x^-) = \sum_{c \in \mathcal{C}} \rho(c) \mathcal{D}_c(x^-). \tag{2}$$

Given $k \leq N_{\mathcal{C}} - 1$, a $(k+1)$-way classification task is a subset $\mathcal{T} \subseteq \mathcal{C}$ of cardinality $|\mathcal{T}| = k+1$, which induces the conditional distribution

$$\mathcal{D}_{\mathcal{T}}(c) = \rho(c \mid c \in \mathcal{T})$$

for $c \in \mathcal{C}$ and we define

$$\mathcal{D}_{\mathcal{T}}(x, c) = \mathcal{D}_{\mathcal{T}}(c) \mathcal{D}_c(x).$$

In particular, we denote as $\mathcal{C}$, whenever there is no ambiguity, the $N_{\mathcal{C}}$-way classification task where the labels are sampled from ρ, namely $\mathcal{D}_{\mathcal{C}}(x, c) = \rho(c) \mathcal{D}_c(x)$.

Supervised loss and mean classifier. For an encoder function $f : \mathcal{X} \to \mathbb{R}^d$, we define a supervised loss (cross-entropy with the best possible linear classifier on top of the representation) over task $\mathcal{T}$ as

$$L_{\mathrm{sup}}(f, \mathcal{T}) = \inf_{W \in \mathbb{R}^{|\mathcal{T}| \times d}} \mathbb{E}_{(x,c) \sim \mathcal{D}_{\mathcal{T}}} \left[-\log \left(\frac{\exp\left(W f(x)\right)_c}{\sum_{c' \in \mathcal{T}} \exp\left(W f(x)\right)_{c'}} \right) \right]. \tag{3}$$

Then, it is natural to consider the *mean* or *discriminant* classifier with weights W^μ which stacks, for $c \in \mathcal{T}$, the vectors

$$W^\mu_{c,:} = \mathbb{E}_{x \sim \mathcal{D}_c} \left[f(x) \right] \tag{4}$$

and whose corresponding (supervised) loss is given by

$$L^\mu_{\mathrm{sup}}(f, \mathcal{T}) = \mathbb{E}_{(x,c) \sim \mathcal{D}_{\mathcal{T}}} \left[-\log \left(\frac{\exp\left(W^\mu f(x)\right)_c}{\sum_{c' \in \mathcal{T}} \exp\left(W^\mu f(x)\right)_{c'}} \right) \right]. \tag{5}$$

Note that, obviously, one has $L_{\mathrm{sup}}(f, \mathcal{T}) \leq L^\mu_{\mathrm{sup}}(f, \mathcal{T})$.

Unsupervised contrastive loss. We consider the unsupervised contrastive loss *with N negative samples* given by

$$L_{\text{un}}^{N}(f) = \mathbb{E}_{\substack{(x,x^+)\sim\mathcal{D}_{\text{sim}} \\ X^-\sim\mathcal{D}_{\text{neg}}^{\otimes N}}} \left[-\log \left(\frac{\exp\left(f(x)^T f(x^+)\right)}{\exp\left(f(x)^T f(x^+)\right) + \sum_{x^-\in X^-} \exp\left(f(x)^T f(x^-)\right)} \right) \right], \quad (6)$$

where $\mathcal{D}_{\text{sim}}$ is given by Equation (1) and where $\mathcal{D}_{\text{neg}}^{\otimes N}$ stands for the N tensor product of the $\mathcal{D}_{\text{neg}}$ distribution given by Equation (2). When a single negative sample is used ($N = 1$), we will use the notation $L_{\text{un}}(f) = L_{\text{un}}^{1}(f)$. In the rest of the paper, N will stand for the number of negatives used in the unsupervised loss (6).

3.1 Inequalities for unsupervised training with multiple classes

The following Lemma states that the unsupervised objective with a single negative sample can be related to the supervised loss for which the target task is classification over the whole set of latent classes $\mathcal{C}$.

Lemma 1. *For any encoder $f : \mathcal{X} \to \mathbb{R}^d$, one has*

$$L_{\text{sup}}(f, \mathcal{C}) \leq L_{\text{sup}}^{\mu}(f, \mathcal{C}) \leq \frac{1}{p_{\text{min}}^{\rho}} L_{\text{un}}(f) + \log N_{\mathcal{C}}, \quad (7)$$

where $p_{\text{min}}^{\rho} = \min_c \rho(c)$.

The proof of Lemma 1 is given in the appendix, and uses a trick from Lemma 4.3 in Saunshi et al. (2019) relying on Jensen's inequality. This Lemma relates the unsupervised and the supervised losses, a shortcoming being the introduction of p_{min}^{ρ}, which is small for a large $N_{\mathcal{C}}$ since obviously $p_{\text{min}}^{\rho} \leq 1/N_{\mathcal{C}}$.

The analysis becomes more difficult with a larger number of negative samples. Indeed, in this case, one needs to carefully keep track of how many distinct classes will be represented by each draw. This is handled by Theorem B.1 of Saunshi et al. (2019), but the bound given therein only estimates an expectation of the supervised loss w.r.t. the random subset of classes considered (so called tasks). For multiple negative samples, the approach adopted in the proof of Lemma 1 above further degrades, since p_{min}^{ρ} would be replaced by the minimum probability among tuple draws, an even much smaller quantity.

We propose the following Lemma, which assumes that the number of negative samples is large enough compared to the number of latent classes.

Lemma 2. *Consider the unsupervised objective with N negative samples as defined in Equation (6) and assume that N satisfies $N = \Omega(N_{\mathcal{C}} \log N_{\mathcal{C}})$. Then, we have*

$$L_{\text{sup}}(f, \mathcal{C}) \leq L_{\text{sup}}^{\mu}(f, \mathcal{C}) \leq \frac{1}{p_{cc}^{\rho}(N)} L_{\text{un}}^{N}(f), \quad (8)$$

where $p_{cc}^{\rho}(N)$ is the probability to have all coupons after N draws in an $N_{\mathcal{C}}$-coupon collector problem with draws from ρ.

The proof of Lemma 2 is given in the appendix. In this result, $p_{cc}^{\rho}(N)$ is related to the following coupon collector problem. Assume that ρ is the uniform distribution over $\mathcal{C}$ and let T be the random number of necessary draws until each $c \in \mathcal{C}$ is drawn at least once. It is known (see for instance Motwani and Raghavan (1995)) that the expectation and variance of T are respectively

given by $N_{\mathcal{C}} H_{N_{\mathcal{C}}}$ and $(N_{\mathcal{C}}\pi)^2/6$, where H_n is the n-th harmonic number $H_n = \sum_{i=1}^{n} 1/i$. This entails using Chebyshev's inequality that

$$\mathbb{P}\left(|T - N_{\mathcal{C}} H_{N_{\mathcal{C}}}| \geq \beta N_{\mathcal{C}}\right) \leq \frac{\pi^2}{6\beta^2}$$

for any $\beta > 0$, so that whenever ρ is sufficiently close to a uniform distribution and $N = \Omega(N_{\mathcal{C}} \log N_{\mathcal{C}})$, the probability p_{cc}^{ρ} is reasonably high. Due to the randomness of the classes sampled during training, it is difficult to obtain a better inequality than Lemma 2 if we want to upper bound $L_{\text{un}}^{N}(f)$ by the supervised $L_{\text{sup}}(f, \mathcal{C})$ on all classes. However, the result can be improved by considering the average loss over tasks $L_{\text{sup},k}(f)$, as explained in the next Section.

3.2 Guarantees on the average supervised loss

In this Section, we bound the average of the supervised classification loss on tasks that are subsets of $\mathcal{C}$. Towards this end, we need to assume (only in this Section) that ρ is uniform. We consider supervised tasks consisting in distinguishing one latent class from k other classes, given that they are distinct and uniformly sampled from $\mathcal{C}$. We define the average supervised loss of f for $(k+1)$-way classification as

$$L_{\text{sup},k}(f) = \mathbb{E}_{\mathcal{T}\sim\mathcal{D}^{k+1}}\left[L_{\text{sup}}\left(f, \mathcal{T}\right)\right], \tag{9}$$

where $\mathcal{D}^{k+1}$ is the uniform distribution over $(k+1)$-way tasks, which means uniform sampling of $\{c_1, \cdots, c_{k+1}\}$ *distinct* classes in $\mathcal{C}$. We define also the average supervised loss of the mean classifier

$$L_{\text{sup},k}^{\mu}(f) = \mathbb{E}_{\mathcal{T}\sim\mathcal{D}^{k+1}}\left[L_{\text{sup}}^{\mu}\left(f, \mathcal{T}\right)\right], \tag{10}$$

where we recall that $L_{\text{sup}}^{\mu}\left(f, \mathcal{T}\right)$ is given by (5). The next Proposition is a generalization to arbitrary values of k and N of Lemma 4.3 from Saunshi et al. (2019), where it is assumed $k = 1$ and $N = 1$.

Proposition 1. *Consider the unsupervised loss $L_{\text{un}}^{N}(f)$ from Equation (6) with N negative samples. Assume that ρ is uniform over $\mathcal{C}$ and that $2 \leq k+1 \leq N_{\mathcal{C}}$. Then, any encoder function $f : \mathcal{X} \to \mathbb{R}^d$ satisfies*

$$L_{\text{sup},k}(f) \leq L_{\text{sup},k}^{\mu}(f) \leq \frac{k}{1-\tau_N^+}\left(L_{\text{un}}^{N}(f) - \tau_N^+ \log(N+1)\right)$$

with $\tau_N^+ = \mathbb{P}\left[c_i = c, \forall i \mid (c, c_1, \cdots, c_N) \sim \rho^{\otimes N+1}\right]$.

The proof of Proposition 1 is given in the appendix. This Proposition states that, in a setting similar to that of Saunshi et al. (2019), on average, the $(k+1)$-way supervised classification loss is upper-bounded by the unsupervised loss (both with $N = 1$ negative or $N > 1$ negatives), that contrastive learning algorithms actually minimize. Therefore, these results give hints for the performances of the learned representation for downstream tasks.

Also, while Saunshi et al. (2019) only considers an unsupervised loss with $N = k$ negatives along with $(k+1)$-way tasks for evaluation, the quantities N and k are decoupled in Proposition 1. Furthermore, whenever ρ is uniform, one has $\tau_N^+ = \sum_{c\in\mathcal{C}} \rho(c)^{N+1} = N_{\mathcal{C}}^{-N}$, which decreases to 0 as $N \to +\infty$, so that a larger number of negatives N makes $k/(1-\tau_N^+)$ smaller and closer to k. This provides a step towards a better understanding of what is actually done in practice with unsupervised contrastive learning. For instance, $N = 65536$ negatives are used in He et al. (2020).

While we considered a generic encoder f and a generic setting in this Section, the next Section 4 considers a more realistic setting of an unsupervised objective with a fixed available dataset, and the study of an *explicit* algorithm for the training of f.

4 Convergence of gradient descent for contrastive unsupervised learning

This section leverages results from Allen-Zhu et al. (2019) to provide convergence guarantees for gradient-descent based minimization of the contrastive training error, where the unsupervisedly trained encoder is an overparametrized deep neural network.

Deep neural network encoder. We consider a family of encoders f defined as a deep feed-forward neural network following Allen-Zhu et al. (2019). We quickly restate its structure here for the sake of completeness. A deep neural encoder f is parametrized by matrices $A \in \mathbb{R}^{m \times d_x}$, $B \in \mathbb{R}^{d \times m}$ and $W_1, \dots, W_L \in \mathbb{R}^{m \times m}$ for some depth L. For an input $x \in \mathbb{R}^{d_x}$, the feed-forward output $y \in \mathbb{R}^d$ is given by

$$g_0 = Ax, \quad h_0 = \phi(g_0), \quad g_l = W_l h_{l-1}, \quad h_l = \phi(g_l) \quad \text{for} \quad l = 1, \dots, L,$$
$$y = Bh_L,$$

where ϕ is the ReLU activation function. Note that the architecture can also include residual connections and convolutions, as explained in Allen-Zhu et al. (2019).

We know from Allen-Zhu et al. (2019) that, provided a δ-separation condition on the dataset (x_i, y_i) for $i = 1, \dots, n$ with $\delta > 0$ and sufficient overparametrization of the model ($m = \Omega\left(\text{poly}(n, L, \delta^{-1}) \cdot d\right)$), the optimisation of the least-squares error $\frac{1}{2}\sum_{i=1}^n \|\widehat{y}_i - y_i\|_2^2$ using gradient descent provably converges to an arbitrarily low value $\epsilon > 0$, where $\widehat{y}_i = f(x_i)$ are the network outputs. Moreover, the convergence is linear i.e. the number of required epochs is $T = O(\log(1/\epsilon))$, although involving a constant of order $\text{poly}(n, L, \delta^{-1})$. Although this result does not directly apply to contrastive unsupervised learning, we explain below how it can be adapted provided a few additional assumptions.

Ideally, we would like to prove a convergence result on the unsupervised objective defined in Equation (6). However, we need to define an objective through an explicitly given dataset so that it falls within the scope of Allen-Zhu et al. (2019). Regarding this issue, we assume in what follows that we dispose of a set of fixed triplets $(x, x^+, x^-) \in (\mathbb{R}^{d_x})^3$ we train on.

Objective function. Let us denote this fixed training set $\{(x_i, x_i^+, x_i^-)\}_{i=1}^n$. Each element leads to an output $z_i = (f(x_i), f(x_i^+), f(x_i^-))$ by the encoder and we optimize the empirical objective

$$\widehat{L}_{\text{un}}(f) = \sum_{i=1}^n \zeta(f(x_i)^T(f(x_i^-) - f(x_i^+))) = \sum_{i=1}^n \ell(z_i), \tag{11}$$

where we introduced the loss function $\ell(z_i) = \ell(z_{i,1}, z_{i,2}, z_{i,3}) = \zeta(z_{i,1}^T(z_{i,3} - z_{i,2}))$ with $\zeta(x) = \log(1 + e^x)$. Note that $\widehat{L}_{\text{un}}(f)/n$ is the empirical counterpart of the unsupervised loss (6). Our management of the set of training triplets can be compared to that of Wen (2020) who similarly fixes them in advance but uses multiple negatives and the same x_i as a positive. However, two distinct encoders are trained therein, one for the reference sample x_i and another for the rest. We consider here the more realistic case where a single encoder is trained. Our approach also applies to multiple negatives, but we only use a single one here for simplicity. We need the following data separation assumption from Allen-Zhu et al. (2019).

Assumption 1. *We assume that all the samples $x \in \mathcal{X}_{\text{data}} = \bigcup_{i=1}^n \{x_i, x_i^+, x_i^-\}$ are normalized $\|x\| = 1$ and that there exists $\delta > 0$ such that $\|x - x'\|_2 \geq \delta$ for any $x, x' \in \mathcal{X}_{\text{data}}$.*

Note that sampling the positives and negatives x_i^+, x_i^- need not to be made through simple draws from the dataset. A common practice in contrastive learning (Chen et al., 2020a) is to use data augmentations, where we replace $x_i^\pm$ by $\psi(x_i^\pm)$ for an augmentation function ψ also drawn at random. Such an augmentation can include, whenever inputs are color images, Gaussian noise, cropping, resizing, color distortion, rotation or a combination thereof, with parameters sampled at random in prescribed intervals. The setting considered here allows the case where $x_i^\pm$ are actually augmentations (we won't write $\psi(x_i^\pm)$ but simply $x_i^\pm$ to simplify notations), provided that Assumption 1 is satisfied and that such augmentations are performed and fixed before training. Note that, in practice, the augmentations are themselves randomly sampled at each training iteration (Chen et al., 2020a). Unfortunately, this would make the objective intractable and the convergence result we are about to derive does not apply in that case.

In order to apply the convergence result from Allen-Zhu et al. (2019), we need to prove that the following gradient-Lipschitz condition

$$\ell(z+z') \leq \ell(z) + \langle \nabla \ell(z), z' \rangle + \frac{L_{\text{smooth}}}{2} \|z'\|^2 \tag{12}$$

holds for any $z, z' \in \mathbb{R}^{3d}$, for some constant $L_{\text{smooth}} > 0$, where ℓ is the loss given by (11). However, as defined previously, ℓ does not satisfy (12) without extra assumptions. We propose to bypass this problem by making the following additional assumption on the norms of the outputs of the encoder.

Assumption 2. *For each element $x \in \mathcal{X}_{\text{data}}$, the output $z = f(x) \in \mathbb{R}^d$ satisfies*

$$\eta < \|z\| < C$$

during and at the end of the training of the encoder f, for some constants $0 < \eta < C < +\infty$.

In Section 5, we check experimentally on three datasets (see Figure 3 herein) that this assumption is rather realistic. The lower bound $\eta > 0$ is necessary and used in Lemma 4 below, while the upper bound C is used in the next Lemma 3, which establishes the gradient-Lipschitz smoothness of the unsupervised loss ℓ and provides an estimation of L_{smooth}.

Lemma 3. *Consider the unsupervised loss ℓ given by (11), grant Assumption 2 and define the set*

$$B^3 = \left\{ z = (z_1, z_2, z_3) \in (\mathbb{R}^d)^3 \ : \ \max_{j=1,2,3} \|z_j\|_2^2 \leq C^2 \right\}$$

where $C > 0$ is defined in Assumption 2. Then, the restriction of ℓ to B^3 satisfies (12) with a constant $L_{\text{smooth}} \leq 2 + 8C^2$.

The proof of Lemma 3 is given in the appendix. Now, we can state the main result of this Section.

Theorem 1. *Grant both Assumptions 1 and 2, let $\epsilon > 0$ and let $\widehat{L}_{\text{un}}(f)$ be the loss given by (11). Then, assuming that*

$$m \geq \Omega\Big(\frac{\text{poly}(n, L, \delta^{-1}) \cdot d}{\epsilon} \Big),$$

the gradient descent algorithm with a learning rate ν and a number of steps T such that

$$\nu = \Theta\Big(\frac{d\delta}{\text{poly}(n, L) \cdot m} \Big) \quad \text{and} \quad T = O\Big(\frac{\text{poly}(n, L)}{\delta^2 \epsilon^2} \Big),$$

finds a parametrization of the encoder f satisfying

$$\widehat{L}_{\text{un}}(f) \leq \epsilon.$$

The proof of Theorem 1 is given in the appendix. Although it uses Theorem 6 from Allen-Zhu et al. (2019), it is actually *not* an immediate consequence of it. Indeed, in our case, the Theorem 6 therein only allows us to conclude that $\|\nabla \widehat{L}_{\mathrm{un}}(f)\| \leq \epsilon$, where the gradient is taken w.r.t. the outputs of f. The convergence of the objective itself is obtained thanks to the following Lemma whose proof is given in the appendix.

Lemma 4. *Grant Assumption 2 and assume that the parameters of the encoder f are optimized so that $\|\nabla \widehat{L}_{\mathrm{un}}(f)\| \leq \epsilon$ with $\epsilon < \eta/2$, where η is defined in Assumption 2. Then, for any $i = 1, \dots, n$, we have $\ell(z_i) \leq 2\epsilon/\eta$ where $z_i = (f(x_i), f(x_i^+), f(x_i^-))$.*

This Lemma is crucial for proving Theorem 1 as it allows to show, in this setting, that the reached critical point is in fact a global minimum.

A natural idea would be then to combine Theorem 1 with Proposition 1 in order to prove that gradient descent training of the encoder using the unsupervised contrastive loss helps to minimize the supervised loss. This paper makes a step towards such a result, but let us stress that it requires much more work, to be considered in future papers, the technical problems to be addressed being as follows. Firstly, the result of Theorem 1 applies to $\widehat{L}_{\mathrm{un}}(f)$ and cannot be directly extrapolated on $L_{\mathrm{un}}(f)$. Doing so would require a sharp control of the generalization error, while Theorem 1 is about the training error only. Secondly, Assumption 1 requires that all samples are separated and, in particular, distinct. This cannot hold when the objective is defined through an expectation as we did in Section 3. Indeed, it would be invalidated simply by reusing a sample in two different triples.

5 Experiments

In this section, we report experiments that illustrate our theoretical findings.

Datasets and Experiments. We use a small convolutional network as encoder on MNIST (LeCun and Cortes, 2010) and FashionMNIST (Xiao et al., 2017), and VGG-16 (Simonyan and Zisserman, 2015) on CIFAR-10 (Krizhevsky et al., 2009). Experiments are performed with PyTorch (Paszke et al., 2019).

Results. Figure 1 provides an illustration of Lemma 1, where we display the values of L_{un} (i.e., L_{un}^N with $N = 1$) and $L_{\mathrm{sup}}^{\mu}(f, \mathcal{C})$ along training iterations over 5 separate runs (and their average). We observe that Inequality (7) is satisfied on these experiments, even when the $\log N_{\mathcal{C}}$ term is discarded. Moreover, both losses follow a similar trend. Figure 2 illustrates Lemma 2 for several values of N. Once again, we observe that both losses behave similarly, and that Inequality (8) seems to hold even without the $1/p_{cc}^{\rho}$ term (removed for these displays).

Finally, Figure 3 displays the minimum and maximum Euclidean norms of the outputs of the encoder along training. On these examples, we observe that one can indeed assume these norms to be lower and upper bounded by constants, as stated in Assumption 2.

6 Conclusion

This work provides extensions to previous results on contrastive unsupervised learning, in order to somewhat improve the theoretical understanding of the performance that is empirically observed with pre-trained encoders used for subsequent supervised task. The main hindrance to tighter bounds in Section 3 is the blind randomness of negative sampling, which is unavoidable

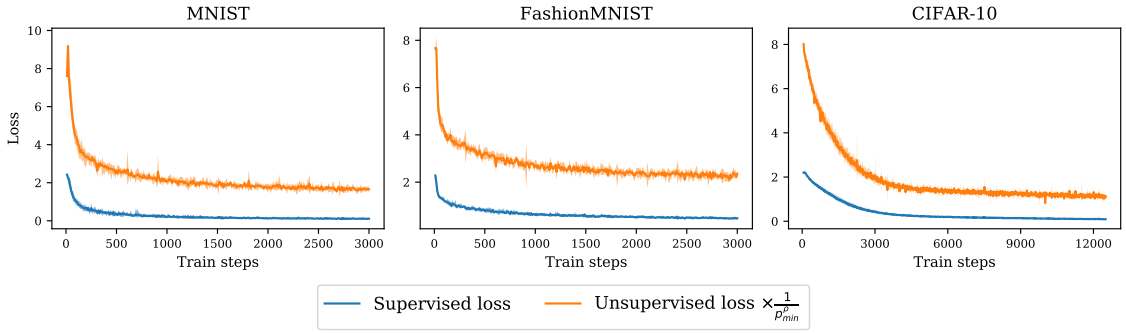

Figure 1: Illustration of Lemma 1: we observe that Inequality (7) is satisfied on these examples, even without the $\log N_{\mathcal{C}}$ term, and that both losses behave similarly (5 runs are displayed together with their average).

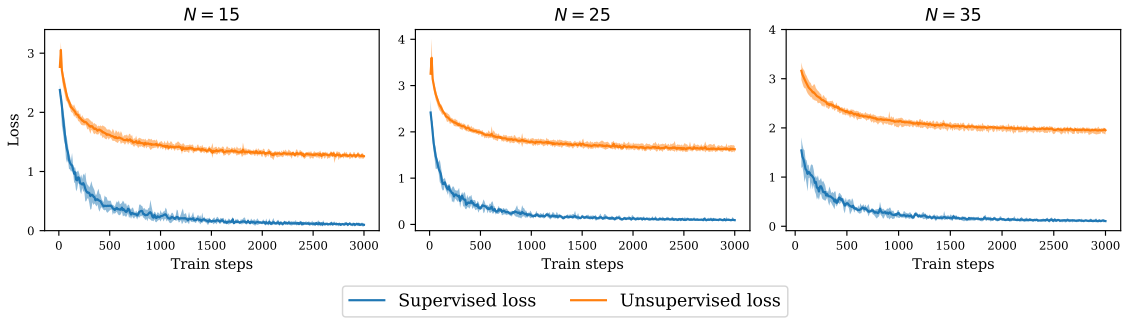

Figure 2: Illustration of Lemma 2 with $N = 15, 25, 35$ on MNIST. We observe again that both the unsupervised and supervised losses behave similarly and that Inequality (8) is satisfied in these experiments, even without the $1/p_{cc}^{\rho}$ factor (5 runs are displayed together with their average).

in the unsupervised setting. Section 4 explains how recent theoretical results about gradient descent training of overparametrized deep neural networks can be used for unsupervised contrastive learning, and concludes with an explanation of why combining the results from Sections 3 and 4 requires many extra technicalities to be considered in future works. Let us conclude by stressing, once again, our motivations for doing this: unsupervised learning theory is much less developed than supervised learning theory, and recent empirical results (see Section 1) indicate that some forms of contrastive learning enable the learning of powerful representations without supervision. In many fields of application, labels are too difficult, too expensive or too invasive to obtain (in medical applications, see for instance Ching et al. (2018)). We believe that a better understanding of unsupervised learning is therefore of utmost importance.

Acknowledgements. This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). Yiyang Yu was supported by grants from Région Ile-de-France.

7 Proofs for Section 3

Apart from the similarity between the unsupervised and supervised loss, the proof of Lemma 1 uses properties of log-sum-exp.

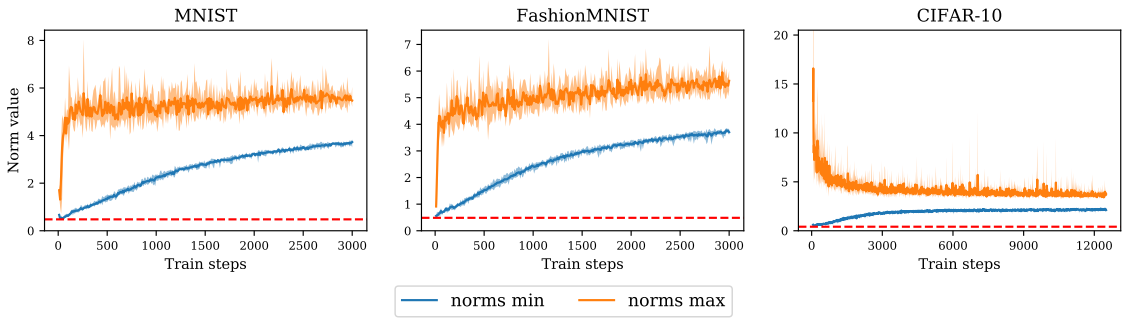

Figure 3: Minimum and maximum Euclidean norms of the outputs of the encoder along contrastive unsupervised training. We observe that Assumption 2 is satisfied on these examples (5 runs are displayed together with their average), the dashed line shows that the minimum norms are away from 0 even in the early iterations.

Proof of Lemma 1. We first rewrite the unsupervised loss as:

$$L_{\text{un}}(f) = \mathbb{E}_{(x,x^+)\sim\mathcal{D}_{\text{sim}},x^-\sim\mathcal{D}_{\text{neg}}} \log\left(1 + \exp\left(f(x)^T(f(x^-) - f(x^+))\right)\right)$$

where we recognize the ζ function $\zeta(x) = \log(1 + e^x)$. We start by using Jensen's inequality

$$\begin{aligned}
L_{\text{un}}(f) &= \mathbb{E}_{\substack{(x,x^+)\sim\mathcal{D}_{\text{sim}} \\ x^-\sim\mathcal{D}_{\text{neg}}}} \left[\zeta\left(f(x)^T(f(x^-) - f(x^+))\right)\right] \\
&\geq \mathbb{E}_{c,c^-\sim\rho,x\sim\mathcal{D}_c} \left[\zeta\left(f(x)^T(\mu_{c^-} - \mu_c)\right)\right] \\
&\geq p_{\min}^{\rho} \mathbb{E}_{c\sim\rho,x\sim\mathcal{D}_c} \left[\max_{c^-} \zeta\left(f(x)^T(\mu_{c^-} - \mu_c)\right)\right] \\
&= p_{\min}^{\rho} \mathbb{E}_{c\sim\rho,x\sim\mathcal{D}_c} \left[\max_{c^-} \text{LSE}\left(0, f(x)^T(\mu_{c^-} - \mu_c)\right)\right] \\
&\geq p_{\min}^{\rho} \left(\mathbb{E}_{c\sim\rho,x\sim\mathcal{D}_c} \left[\text{LSE}\left(f(x)^T(\mu_{c_1} - \mu_c), \ldots, f(x)^T(\mu_{c_{N_\mathcal{C}}} - \mu_c)\right)\right] - \log N_\mathcal{C}\right) \\
&= p_{\min}^{\rho} \left(L_{\text{sup}}^{\mu}(f, \mathcal{C}) - \log N_\mathcal{C}\right)
\end{aligned}$$

where we have used properties of the log-sum-exp function

$$\max(x_1, \ldots, x_n) \leq \text{LSE}(x_1, \ldots, x_n) \leq \max(x_1, \ldots, x_n) + \log n,$$

the fact that LSE is non-negative whenever one of its arguments is, and for $x \in \mathbb{R}^{2n}$ we have

$$\text{LSE}(x) = \text{LSE}(\text{LSE}(x_1, x_2), \ldots, \text{LSE}(x_{2n-1}, x_{2n})) \leq \max_{j=1,\ldots,n} \text{LSE}(x_{2j-1}, x_{2j}) + \log n.$$

□

The proof of Lemma 2 considers the sample draws where all classes are represented.

Proof of Lemma 2. Let $I \in [N_\mathcal{C}]^N$ the random vector of classes for each negative sample ($I \sim \rho^{\otimes N}$) and let J be the set of represented classes i.e. $J = \{I_j \mid j \in [N]\}$. We have, again with Jensen's inequality

$$\begin{aligned}
L_{\text{un}}^N(f) &= \mathbb{E}_{x,x^+,x_1^-,\ldots,x_N^-} \left[\text{LSE}\left(0, f(x)^T(f(x_1^-) - f(x^+)), \ldots, f(x)^T(f(x_N^-) - f(x^+))\right)\right] \\
&\geq \mathbb{E}_{c\sim\rho,I\sim\rho^{\otimes N},x\sim\mathcal{D}_c} \left[\text{LSE}\left(0, f(x)^T(\mu_{I_1} - \mu_c), \ldots, f(x)^T(\mu_{I_N} - \mu_c)\right)\right] \\
&\geq \mathbb{P}\left(|J| = N_\mathcal{C}\right) \mathbb{E}_{\substack{c\sim\rho \\ I\sim\rho^{\otimes N} \\ x\sim\mathcal{D}_c}} \left[\text{LSE}\left(0, f(x)^T(\mu_{I_1} - \mu_c), \ldots, f(x)^T(\mu_{I_N} - \mu_c)\right) \mid |J| = N_\mathcal{C}\right] \\
&\geq \mathbb{P}\left(|J| = N_\mathcal{C}\right) L_{\text{sup}}^{\mu}(f, \mathcal{C}),
\end{aligned}$$

where we used that for $\mathcal{S} \subset [n]$ and $x \in \mathbb{R}^n$ we have $\mathrm{LSE}(x_{\mathcal{S}}) \leq \mathrm{LSE}(x)$ with $x_{\mathcal{S}}$ the restriction of x to the indices in $\mathcal{S}$. Finally, we have $\mathbb{P}\left(|J| = N_{\mathcal{C}}\right) = p_{cc}^{\rho}(N)$. □

We restate Proposition 1 for cases $N = 1$ and $N > 1$. The proof uses Jensen's inequality and the uniformity of ρ.

Proposition 2 (3.3 (restated))**.** *Consider the unsupervised loss $L_{\mathrm{un}}^{N}(f)$ from Equation (6) with N negative samples. Assume that ρ is uniform over $\mathcal{C}$ and that $2 \leq k+1 \leq N_{\mathcal{C}}$. Then,*

(1) any encoder function $f : \mathcal{X} \to \mathbb{R}^d$ satisfies

$$L_{\mathrm{sup},k}(f) \leq L_{\mathrm{sup},k}^{\mu}(f) \leq \frac{k}{1-\tau^+}\left(L_{\mathrm{un}}(f) - \tau^+\right)$$

with $\tau^+ = \mathbb{P}_{c,c' \sim \rho^2}\left(c = c'\right)$, where $L_{\mathrm{un}}(f)$ is the unsupervised loss from Equation (6) with $N = 1$ negative sample;

(2) more generally,

$$L_{\mathrm{sup},k}(f) \leq L_{\mathrm{sup},k}^{\mu}(f) \leq \frac{k}{1-\tau_N^+}\left(L_{\mathrm{un}}^{N}(f) - \tau_N^+ \log(N+1)\right)$$

with $\tau_N^+ = \mathbb{P}(c_i = c, \forall i \mid c \sim \rho, (c_1, \cdots, c_N) \sim \rho^N)$, and where $L_{\mathrm{un}}^{N}(f)$ is the unsupervised loss from Equation (6).

Proof of Proposition 1. Let's start with (1). By Jensen's inequality, then use $\log = \log_2$ without loss of generality, and split the expectation into cases $c^- \neq c$ and $c^- = c$,

$$\begin{aligned}
L_{\mathrm{un}}(f) &= \mathbb{E}_{(c,c^-)\sim\rho^2}\mathbb{E}_{x,x^+\sim\mathcal{D}_c,x^-\sim\mathcal{D}_{c^-}}\left[\log\left(1+\exp\left(f(x)^T\left(f(x^-)-f(x^+)\right)\right)\right)\right] \\
&\geq \mathbb{E}_{(c,c^-)\sim\rho^2,x\sim\mathcal{D}_c}\left[\log\left(1+\exp\left(f(x)^T\left(\mu_{c^-}-\mu_c\right)\right)\right)\right] \\
&= (1-\tau^+)\mathbb{E}_{c\sim\rho,x\sim\mathcal{D}_c}\mathbb{E}_{c^-\sim\rho}\left[\log\left(1+\exp\left(f(x)^T\left(\mu_{c^-}-\mu_c\right)\right)\right)\middle|c^- \neq c\right] + \tau^+.
\end{aligned}$$

Let us write explicitly the uniform distribution ρ on $\mathcal{C}$. On the one hand,

$$\begin{aligned}
&\mathbb{E}_{c^-\sim\rho}\left[\log\left(1+\exp\left(f(x)^T\left(\mu_{c^-}-\mu_c\right)\right)\right)\middle|c^- \neq c\right] \\
&= \sum_{c^-\in\mathcal{C}\setminus\{c\}} \frac{1}{N_{\mathcal{C}}-1}\log\left(1+\exp\left(f(x)^T\left(\mu_{c^-}-\mu_c\right)\right)\right),
\end{aligned}$$

on the other hand, And this is for every $c^- \in \mathcal{C}\setminus\{c\}$. We rearrange the double sum according to c^-

Hence, using the uniformity of ρ,

$$\begin{aligned}
&\mathbb{E}_{c^-\sim\rho}\left[\log\left(1+\exp\left(f(x)^T\left(\mu_{c^-}-\mu_c\right)\right)\right)\middle|c^- \neq c\right] \\
=&\frac{1}{k}\mathbb{E}_{c_1,\ldots,c_k\sim\rho^{\otimes k}}\left[\sum_{i=1}^{k}\log\left(1+\exp\left(f(x)^T\left(\mu_{c_i}-\mu_c\right)\right)\right)\middle|\{c,c_1,\ldots,c_k\}\text{ distinct}\right] \\
\geq&\frac{1}{k}\mathbb{E}_{c_1,\ldots,c_k\sim\rho^{\otimes k}}\left[\log\left(1+\sum_{i=1}^{k}\exp\left(f(x)^T\left(\mu_{c_i}-\mu_c\right)\right)\right)\middle|\{c,c_1,\ldots,c_k\}\text{ distinct}\right].
\end{aligned}$$

That means we have

$$
\begin{aligned}
L_{\mathrm{un}}(f) &\geq \frac{1-\tau^+}{k} \mathop{\mathbb{E}}_{\substack{c\sim\rho, x\sim\mathcal{D}_c \\ c_1,\ldots,c_k\sim\rho^{\otimes k}}} \left[\log\left(1+\sum_{i=1}^{k} \exp\left(f(x)^T\left(\mu_{c_i}-\mu_c\right)\right)\right) \middle| \{c,c_1,\ldots,c_k\} \text{ distinct} \right] + \tau^+ \\
&= \frac{1-\tau^+}{k} \mathbb{E}_{\mathcal{T}\sim\mathcal{D}^{k+1}} \mathbb{E}_{(x,c)\sim\mathcal{D}_{\mathcal{T}}} \left[-\log\left(\frac{\exp(f(x)^T\mu_c)}{\exp(f(x)^T\mu_c) + \sum_{\substack{c^-\in\mathcal{T} \\ c^-\neq c}} \exp\left(f(x)^T\mu_{c^-}\right)} \right) \right] + \tau^+ \\
&= \frac{1-\tau^+}{k} L^{\mu}_{\mathrm{sup},k}(f) + \tau^+.
\end{aligned}
$$

As for (2), again by Jensen's inequality, and split the expectation into cases $c_i^- = c, \forall i$ and $\exists c_i^- \neq c$,

$$
\begin{aligned}
L^N_{\mathrm{un}}(f) &= \mathbb{E}_{(c,c_i^-)\sim\rho^{N+1}} \mathbb{E}_{x,x^+\sim\mathcal{D}_c, x_i^-\sim\mathcal{D}_{c_i^-}} \left[\log\left(1+\sum_{i=1}^{N} \exp\left(f(x)^T\left(f(x_i^-)-f(x^+)\right)\right)\right)\right] \\
&\geq \mathbb{E}_{(c,c_i^-)\sim\rho^{N+1}, x\sim\mathcal{D}_c} \left[\log\left(1+\sum_{i=1}^{N} \exp\left(f(x)^T\left(\mu_{c_i^-}-\mu_c\right)\right)\right)\right] \\
&= (1-\tau_N^+) \mathop{\mathbb{E}}_{\substack{c\sim\rho \\ x\sim\mathcal{D}_c}} \mathbb{E}_{c_i^-\sim\rho^N} \left[\log\left(1+\sum_{i=1}^{N} \exp\left(f(x)^T\left(\mu_{c^-}-\mu_c\right)\right)\right) \middle| \exists c_i^- \neq c \right] + \tau_N^+ \log(N+1)
\end{aligned}
$$

with

$$
\tau_N^+ = \mathbb{P}(c_i = c, \forall i \mid c\sim\rho, c_i\sim\rho^N) = \sum_{c\in\mathcal{C}} \rho(c)^{N+1} = N_{\mathcal{C}}^{-N}.
$$

Considering the fact that

$$
\begin{aligned}
&\mathbb{E}_{c_i^-\sim\rho^N} \left[\log\left(1+\sum_{i=1}^{N} \exp\left(f(x)^T\left(\mu_{c^-}-\mu_c\right)\right)\right) \middle| \exists c_i^- \neq c \right] \geq \\
&\qquad \mathbb{E}_{c^-\sim\rho} \left[\log\left(1+\exp\left(f(x)^T\left(\mu_{c^-}-\mu_c\right)\right)\right) \middle| c^- \neq c \right],
\end{aligned}
$$

then by similar computations as in (1), we have

$$
L^N_{\mathrm{un}}(f) \geq \frac{1-\tau_N^+}{k} L^{\mu}_{\mathrm{sup},k}(f) + \tau_N^+ \log(N+1).
$$

□

8 Proofs for Section 4

Let us first prove that under Assumption 2, the objective is gradient-Lipschitz w.r.t. the network outputs.

Lemma 5 (Lemma 4.1)**.** *Consider the unsupervised loss ℓ given by (11), grant Assumption 2 and define the set*

$$
B^3 = \left\{ z=(z_1,z_2,z_3)\in(\mathbb{R}^d)^3 \ : \ \max_{j=1,2,3} \|z_j\|_2^2 \leq C^2 \right\}
$$

where $C > 0$ is defined in Assumption 2. Then, the restriction of ℓ to B^3 satisfies (12) with a constant $L_{\mathrm{smooth}} \leq 2 + 8C^2$.

Proof. We will prove this result by bounding the norm of the Hessian matrix.

Let us write the gradient of $\ell(z)$ with respect to z first. We have $z \in \mathbb{R}^{3d}$. For ease of writing, we define the matrices $A_1, A_2, A_3 \in \mathbb{R}^{3d \times d}$ as

$$A_1 = \begin{pmatrix} I_d \\ 0_d \\ 0_d \end{pmatrix} \quad A_2 = \begin{pmatrix} 0_d \\ I_d \\ 0_d \end{pmatrix} \quad A_3 = \begin{pmatrix} 0_d \\ 0_d \\ I_d \end{pmatrix}$$

where $I_d, 0_d \in \mathbb{R}^{d \times d}$ are the identity and zero matrix respectively. With this notation, we have $z_i = A_i^T z$ for $i = 1, 2, 3$ the three contiguous thirds of z's coordinates.

Our purpose is to compute

$$\frac{\partial}{\partial z}\ell(z) = \frac{\partial}{\partial z}\left[-\log\left(\frac{\exp\left(z_1^T z_2\right)}{\exp\left(z_1^T z_2\right) + \exp\left(z_1^T z_3\right)}\right)\right].$$

Denote $\cos_{i,j} = z_i^T z_j$, we can now compute for $i, j \in \{1, 2, 3\}$ $(i \neq j)$

$$\frac{\partial}{\partial z}\cos_{i,j} = \left(A_i A_j^T + A_j A_i^T\right) z =: \partial \cos_{i,j} \in \mathbb{R}^{3d}.$$

Now, denote $v = \operatorname{softmax}\left(\cos_{1,2}, \cos_{1,3}\right) \in \mathbb{R}^2$, we can write

$$\frac{\partial}{\partial z}\ell(z) = (v_1 - 1)\partial \cos_{1,2} + v_2 \partial \cos_{1,3}\,.$$

We proceed with the following computation

$$\frac{\partial^2}{\partial z^2}\cos_{i,j} = A_i A_j^T + A_j A_i^T,$$

which we will denote simply as $\partial^2 \cos_{i,j}$. Before we get the Hessian of loss, we still need to compute

$$\partial v := \frac{\partial v}{\partial z} = \left(\operatorname{diag}(v) - vv^T\right)\begin{pmatrix} \partial \cos_{1,2}^T \\ \partial \cos_{1,3}^T \end{pmatrix} \in \mathbb{R}^{2 \times 3d}.$$

Now we can write

$$\frac{\partial^2}{\partial z^2}\ell(z) = (v_1 - 1)\partial^2 \cos_{1,2} + v_2 \partial^2 \cos_{1,3} + \begin{pmatrix} \partial \cos_{1,2} & \partial \cos_{1,3} \end{pmatrix} \partial v.$$

We can now estimate the norm of this matrix which will provide an estimation for the Lipschitz constant.

We find that

$$\left\|\partial \cos_{i,j}\right\| \leq 2 \max\left(\|z_i\|, \|z_j\|\right),$$

keeping in mind that the matrix $\operatorname{diag}(v) - vv^T$ has norm at most $1/2$, this leads to

$$\left\|\begin{pmatrix} \partial \cos_{1,2} & \partial \cos_{1,3} \end{pmatrix} \partial v\right\| = 8 \max_{i,j}\left(\|z_i\| \|z_j\|\right).$$

We have also that $\left\|\partial^2 \cos_{i,j}\right\| = 1$.

All in all, we have $\left\|\frac{\partial^2}{\partial z^2}\ell(z)\right\| = 2 + 8 \max_{i,j}\left(\|z_i\| \|z_j\|\right)$. Recalling that we restricted $\mathbb{R}^{3d}$ so that we have $\max_i \|z_i\| \leq C$ the result follows. □

Theorem 1 is actually obtained in two steps. First, Theorem 6 from Allen-Zhu et al. (2019) allows us to obtain that the gradient of the objective $\nabla \widehat{L}_{\text{un}}(f)$ with respect to the network outputs reaches arbitrarily low values. Then, combining this with Assumption 2, this result can be extended into the objective itself.

Following appendix A of Allen-Zhu et al. (2019), we need to define the loss vectors for our model. These are originally defined as $\text{loss}_i = y_i - y_i^*$ (y_i and y_i^* are respectively the output and label corresponding to an input x_i from the dataset) for the ℓ^2 loss. More generally, for a network output z_i, they are defined as

$$\text{loss}_i = \nabla_z \ell(z_i).$$

Following the unsupervised training protocol, samples are fed into the network three at a time x, x^+ and x^-. Let us denote θ the parameters of the network f, for a triplet (x_i, x_i^+, x_i^-), the trick is to write:

$$\frac{\partial}{\partial \theta} \ell(z_i) = \frac{\partial z}{\partial \theta} \underbrace{\frac{\partial}{\partial z} \ell(z_i)}_{\text{loss}}$$

with z_i the concatenation of $f(x_i), f(x_i^+), f(x_i^-)$.

By denoting $(x_1, x_2, x_3) = (x_i, x_i^+, x_i^-)$, the previous writing is equivalent to

$$\sum_{j=1}^{3} \frac{\partial f(x_j)}{\partial \theta} A_j^T \frac{\partial}{\partial z} \ell(z_i)$$

and by letting $\text{loss}_{i,j} = A_j^T \frac{\partial}{\partial z} \ell(z_i)$, we obtain a triplet of loss vectors for each data triple (matrices A_j defined in the previous proof).

Lemma 6. *Grant Assumption 1 and let $\widehat{L}_{\text{un}}(f)$ be the loss incurred by f:*

$$\widehat{L}_{\text{un}}(f) = \sum_{i=1}^{n} \ell(f(x_i), f(x_i^+), f(x_i^-))$$

and let $\epsilon > 0$ be the desired precision. Then, assuming $m \geq \Omega\left(\text{poly}(n, L, \delta^{-1}) \cdot d\epsilon^{-1}\right)$, the gradient descent with learning rate $\nu = \Theta\left(\frac{d\delta}{\text{poly}(n,L) \cdot m}\right)$ finds parameters such that

$$\|\nabla \widehat{L}_{\text{un}}(f)\| \leq \epsilon$$

after a number of steps $T = O\left(\frac{\text{poly}(n,L)}{\delta^2 \epsilon^2}\right)$.

Proof. This result follows from Allen-Zhu et al. (2019) (see Theorem 6 and appendix A). It corresponds to the case of a non-convex bounded loss function. We only need to check the used loss function ℓ is bounded and gradient-Lipschitz smooth. The latter condition is verified due to Lemma 5 and Assumption 2.

As for the boundedness, it is also a consequence of Assumption 2 and the fact that the softplus function satisfies

$$\zeta(x) \sim^{x \to +\infty} x \quad \text{and} \quad \lim_{x \to -\infty} \zeta(x) = 0.$$

□

From here, we can derive a result for the objective itself (Theorem 1) thanks to the following Lemma.

Lemma 7 (Lemma 4.2). *Grant Assumption 2 and assume that the parameters of the encoder f are optimized so that $\|\nabla \widehat{L}_{\text{un}}(f)\| \leq \epsilon$ with $\epsilon < \eta/2$, where η is defined in Assumption 2. Then, for any $i = 1, \ldots, n$, we have $\ell(z_i) \leq 2\epsilon/\eta$ where $z_i = (f(x_i), f(x_i^+), f(x_i^-))$.*

Proof. Since we assume $\|\nabla \widehat{L}_{\text{un}}(f)\| \leq \epsilon$, this also implies that $\max_{i,j} \|\text{loss}_{i,j}\| \leq \epsilon$ (see Theorem 3 of Allen-Zhu et al. (2019) and its variant in appendix A).

We can write the norms $\|\text{loss}_{i,j}\|$ as:

$$\begin{aligned}
\|\text{loss}_{i,1}\| &= \|(v_1 - 1)z_{i,2} + v_2 z_{i,3}\| \\
\|\text{loss}_{i,2}\| &= |v_1 - 1| \, \|z_{i,1}\| \\
\|\text{loss}_{i,3}\| &= v_2 \, \|z_{i,1}\|
\end{aligned}$$

where we defined $v = \text{softmax}(z_1^T z_2, z_1^T z_3)$.

Thanks to Assumption 2, we can argue that $\|z_{i,j}\| \geq \eta$. These quantities can be small for $v_1 \to 1$ and $v_2 \to 0$. Since we have $\max_{i,j} \|\text{loss}_{i,j}\| \leq \epsilon$, this implies in particular that for all i we get $\|\text{loss}_{i,3}\| \leq \epsilon$ which means $v_2 \leq \epsilon/\eta$, and we have $v_2 = \sigma(z_1^T(z_3 - z_2))$. So for an instance $i \in [n]$ the loss term in the objective is:

$$\begin{aligned}
\zeta(z_{i,1}^T(z_{i,3} - z_{i,2})) &= \log(1 + \exp(z_{i,1}^T(z_{i,3} - z_{i,2}))) \\
&= -\log(\sigma(-z_{i,1}^T(z_{i,3} - z_{i,2}))) \\
&= -\log(1 - \sigma(z_{i,1}^T(z_{i,3} - z_{i,2}))) \\
&= -\log(1 - v_2) \leq \frac{v_2}{1 - v_2} \leq 2v_2 \leq 2\epsilon/\eta,
\end{aligned}$$

where we used the inequality $-\log(1 - x) \leq \frac{x}{1-x}$ for $0 \leq x < 1$, and the assumption that $\epsilon < \eta/2$. □

Lemma 7 allows us to deduce that the objective is well optimized (we treated the loss term for a single triplet here but the same methods can be applied to the whole objective with a number of gradient steps which is still polynomial).

Proof of Theorem 1. Theorem 1 is the consequence of combining Lemma 6 applied using $\frac{\epsilon\eta}{2n}$ instead of ϵ and Lemma 7 (the $1/n$ factor can be absorbed by the $\text{poly}(n, L)$ factors in the bounds of Lemma 6). □

References

Z. Allen-Zhu, Y. Li, and Z. Song. A Convergence Theory for Deep Learning via Over-Parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments, 2020.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations, 2020a.

T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton. Big self-supervised models are strong semi-supervised learners, 2020b.

T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. M. Hoffman, et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141):20170387, 2018.

C.-Y. Chuang, J. Robinson, L. Yen-Chen, A. Torralba, and S. Jegelka. Debiased contrastive learning, 2020.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

C. Doersch and A. Zisserman. Multi-task self-supervised visual learning. *CoRR*, abs/1708.07860, 2017. URL http://arxiv.org/abs/1708.07860.

C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. *CoRR*, abs/1505.05192, 2015. URL http://arxiv.org/abs/1505.05192.

A. Dosovitskiy, J. T. Springenberg, M. A. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with convolutional neural networks. *CoRR*, abs/1406.6909, 2014. URL http://arxiv.org/abs/1406.6909.

J.-Y. Franceschi, A. Dieuleveut, and M. Jaggi. Unsupervised scalable representation learning for multivariate time series. In *Advances in Neural Information Processing Systems*, pages 4650–4661, 2019.

S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. *CoRR*, abs/1803.07728, 2018. URL http://arxiv.org/abs/1803.07728.

J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020.

K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

O. J. Hénaff, A. Srinivas, J. D. Fauw, A. Razavi, C. Doersch, S. M. A. Eslami, and A. van den Oord. Data-efficient image recognition with contrastive predictive coding. *CoRR*, abs/1905.09272, 2019. URL http://arxiv.org/abs/1905.09272.

A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Y. LeCun and C. Cortes. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

J. D. Lee, Q. Lei, N. Saunshi, and J. Zhuo. Predicting what you already know helps: Provable self-supervised learning, 2020.

L. Logeswaran and H. Lee. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*, 2018.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995. doi: 10.1017/CBO9780511814075.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

N. Saunshi, O. Plevrakis, S. Arora, M. Khodak, and H. Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637, 2019.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

A. Srinivas, M. Laskin, and P. Abbeel. CURL: Contrastive unsupervised representations for reinforcement learning, 2020.

C. Sun, F. Baradel, K. Murphy, and C. Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019.

A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748.

T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere, 2020.

Z. Wen. Convergence of end-to-end training in deep unsupervised contrasitive learning, 2020.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms, 2017.

R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. *CoRR*, abs/1603.08511, 2016. URL http://arxiv.org/abs/1603.08511.