Tight PAC-Bayesian Risk Certificates for Contrastive Learning

Anna van Elst, Debarghya Ghoshdastidar

Technical University of Munich
School of Computation, Information and Technology

[ ]

Abstract

Recent advances in the theory of supervised learning have focused on deriving risk certificates for neural networks. However, despite its empirical success, contrastive representation learning has received limited theoretical attention. Existing works either result in vacuous risk certificates (bounds based on Rademacher complexity or f-divergence) or they require impractical assumptions (independence of augmented samples). In this paper, we improve upon previous work by deriving non-vacuous risk certificates for the SimCLR framework, a widely used contrastive learning approach. Specifically, our contributions are threefold: (1) using a PAC-Bayesian approach, we establish two new risk certificates for the SimCLR loss without making strong assumptions and show that our certificates are non-vacuous and significantly tighter than previous ones, (2) we refine existing bounds on the downstream classification loss by incorporating SimCLR-specific factors, including data augmentation and temperature scaling, and (3) we derive risk certificates for the contrastive zero-one risk. We empirically support our results through experiments on MNIST and CIFAR-10.

Introduction

Contrastive representation learning has emerged as a powerful paradigm in self-supervised learning, enhancing data representations by bringing semantically similar pairs closer together in the representation space than randomly drawn negative samples (Wu et al. 2018; He et al. 2020; Le-Khac, Healy, and Smeaton 2020). This technique gained widespread attention with the introduction of the SimCLR framework, which proposed a simple formulation of contrastive learning (Chen et al. 2020). The SimCLR framework employs a carefully designed contrastive loss to maximize the similarity between the representations of augmented views of the same sample while minimizing the similarity between the representations of augmented views from different samples (Sohn 2016; Chen et al. 2020). This approach significantly outperformed previous self-supervised and semi-supervised learning methods on the ImageNet benchmark. Key components contributing to its success include: (1) composition of multiple data augmentation operations, (2) a projection head, (3) normalized embeddings, and temperature scaling in the contrastive loss, and (4) a large batch size, which determines the number of negative samples K (Chen et al. 2020). However, despite its empirical success, a comprehensive understanding of SimCLR's performance and a rigorous certification of its generalization abilities remain challenging (Bao, Nagano, and Nozawa 2022; Nozawa, Germain, and Guedj 2020).

Arora et al. introduced a theoretical framework for contrastive unsupervised representation learning and derived the first generalisation bounds via Rademacher complexity (Arora et al. 2019). This motivated further study of contrastive unsupervised learning in a PAC-Bayesian setting, diverging from Rademacher complexities (Nozawa, Germain, and Guedj 2020). By analyzing the mean classifier, Arora et al. additionally provided the first generalization bound on downstream classification, later improved by (Bao, Nagano, and Nozawa 2022; Wang et al. 2022). However, these advancements largely rely on a theoretical framework assuming independent and identically distributed (i.i.d.) data: each sample is a tuple $(x, x^+, x_1^-, \dots, x_k^-)$ where x and x^+ come from the same distribution, while $x_1^-, \dots, x_k^-$ are k i.i.d. samples presumably unrelated to x (Arora et al. 2019; Nozawa, Germain, and Guedj 2020). In contrast, the SimCLR loss uses a batch B of size m as a sample, applies data augmentation to generate m positive pairs (x, x^+), treats the remaining augmented samples as negative, and averages the sub-losses over all positive pairs. As a result, the individual loss terms of the SimCLR loss are non-i.i.d, as each term considers a positive pair along with its negative samples. This practical setting has been shown to enhance performance and provide efficient computation, yet a thorough theoretical analysis of this approach remains to be conducted. A key issue is that PAC-Bayes bounds no longer apply in this non-i.i.d. setting. One approach is to compute the PAC-Bayes bound over a dataset of i.i.d. batches, though this weakens the risk certificate by replacing the dataset size n with the typically smaller number of batches. Alternatively, Nozawa, Germain, and Guedj suggest using f-divergence to address the non-i.i.d. setting, but their proposed bounds are vacuous. In this paper, we show that current PAC-Bayes bounds can be improved by leveraging Hoeffding's and McDiarmid's inequalities (Hoeffding 1994; McDiarmid et al. 1989). Moreover, most work relies on the conditional independence assumption (Arora et al. 2019; Nozawa, Germain, and Guedj 2020; Bao, Nagano, and Nozawa 2022), which assumes that x and x^+ are independent. However, this assumption does not hold in the SimCLR framework, where x and x^+ are augmented views of the same sample. In contrast, Wang et al. propose a bound that does not rely on conditional independence (Wang et al. 2022). Additionally, most existing work does not consider practical settings that are known to enhance performance in the SimCLR framework, such as temperature scaling, projection head, and large batch sizes (Bao, Nagano, and Nozawa 2022; Wang et al. 2022). A comparative analysis of these works is summarized in Table 1.

Contrastive Loss	**non-vacuous**	**non-i.i.d.**	**SimCLR**
Arora et al.	×	×	×
Nozawa et al.	✓	✓	×
Th. 1 & 2 (ours)	✓	✓	✓
Downstream	**large K**	**data aug.**	**τ-scaling**
Arora et al.	×	×	×
Bao et al.	✓	×	×
Wang et al.	✓	✓	×
Th. 3 (ours)	✓	✓	✓

Table 1: Comparative analysis of previous works. In this table, a checkmark (✓) indicates that the method satisfies the specified condition, and a cross (×) denotes that it does not. The table is divided into two parts: the first part compares generalization bounds for contrastive learning, assessing whether they are non-vacuous, applicable in non-iid settings, and directly applicable to the SimCLR loss. Although Nozawa et al. do not provide a SimCLR-specific bound, their f-divergence bound can be extended to SimCLR, as detailed in Appendix A. The second part compares existing bounds on downstream classification loss, focusing on their incorporation of a large number of negative samples (K), data augmentation, and temperature (τ) scaling in the contrastive loss.

This paper aims to develop practical risk certificates for contrastive learning applicable to the widely used SimCLR framework. Using a PAC-Bayesian approach, we establish two new risk certificates for the SimCLR loss considering its non-i.i.d. characteristics and build on recent advances in risk certificates for neural networks (Perez-Ortiz et al. 2021b,a). We show that our certificates are non-vacuous and significantly tighter than previous ones (Nozawa, Germain, and Guedj 2020). Additionally, we refine existing bounds on downstream classification loss (Wang et al. 2022; Bao, Nagano, and Nozawa 2022) by incorporating SimCLR-specific factors, including data augmentation and temperature scaling. Finally, we extend our analysis to the contrastive zero-one risk and derive corresponding risk certificates. We evaluate our results by conducting experiments on MNIST and CIFAR-10.

Preliminaries

In this section, we first explain the main components of the SimCLR framework, detailing notation and underlying assumptions. Next, we introduce essential concepts from PAC-Bayes theory.

SimCLR Framework

Several works have proposed mathematical formulations for analyzing contrastive representation learning (Arora et al. 2019; Wang and Isola 2020; HaoChen et al. 2021). Here, we present the mathematical formulations and assumptions specific to the SimCLR framework (Chen et al. 2020).

Data. Let $\overline{\mathcal{X}}$ denote the input space, such as $\overline{\mathcal{X}} \subset \mathbb{R}^{\text{channel}\times\text{width}\times\text{height}}$ for color images. We have access only to an unlabeled training dataset. We use $\mathcal{A}(\cdot \mid \bar{x})$ to denote the distribution of augmented samples generated from a given sample $\bar{x} \in \overline{\mathcal{X}}$. Let $\bar{X}$ represent a batch of m i.i.d. samples drawn from $\overline{\mathcal{X}}$. For each data point $\bar{x} \in \bar{X}$, we generate two augmented samples $x, x^{+} \sim \mathcal{A}(\cdot \mid \bar{x})$. We refer to (x, x^{+}) as the positive pair for $\bar{x} \in \bar{X}$. We denote X^{-} as the set of negative samples for $\bar{x}$ and consider two losses: (1) the SimCLR loss where the remaining $2(m-1)$ augmented samples within the minibatch serve as negative examples for $\bar{x}$, and (2) the simplified SimCLR loss where we consider $m-1$ augmented samples generated from different anchor samples, ensuring the negative samples are i.i.d. We denote B as the batch of m i.i.d. positive pairs.
We define the data distribution $\mathcal{S}$ as the process that generates a positive pair (x, x^{+}) according to the following scheme:

1. Draw a sample $\bar{x} \sim \mathcal{D}_{\overline{\mathcal{X}}}$;
2. Draw two augmented samples $x, x^{+} \sim \mathcal{A}(\cdot \mid \bar{x})$.

Representation function. SimCLR learns a representation function $f : \mathcal{X} \to \mathbb{S}^{d-1}$ that maps inputs to unit vectors in d-dimensional space. For convenience, we consider the set of representation functions $\mathcal{F}$ (i.e., neural networks) to be parameterized by the weight space $\mathcal{W} \subset \mathbb{R}^{p}$, where p denotes the number of parameters of the neural network. Thus, each representation function $f \in \mathcal{F}$ is determined by its weight vector $w \in \mathcal{W}$. When incorporating a projection head, we denote the backbone features as $f(x) \in \mathbb{R}^d$ and the projection head output as $f_1(x) \in \mathbb{S}^{k-1}$, where typically $d \geq k$. Post contrastive training, the projection head is removed, and only the backbone features are used for downstream tasks.

Contrastive loss. Since the feature representations are normalized, we define the cosine similarity between two representations $u, v \in \mathbb{S}^{d-1}$ as $\mathrm{sim}(u, v) := u^\top v$. Let $\tau \in \mathbb{R}_+$ be a temperature parameter. For a sample $\bar{x} \in \bar{X}$, the contrastive loss $\ell_{\text{cont}}(x^+, x, X^-)$ is defined as:

$$-\log \frac{\exp\left(\frac{f(x)^\top f(x^+)}{\tau}\right)}{\exp\left(\frac{f(x)^\top f(x^+)}{\tau}\right) + \sum_{x' \in X^-} \exp\left(\frac{f(x)^\top f(x')}{\tau}\right)}.$$

Remark: The contrastive loss used in SimCLR is slightly different and defined as:

$$\ell_{SimCLR} = \frac{\ell_{\text{cont}}(x^+, x, X^-) + \ell_{\text{cont}}(x, x^+, X^-)}{2}.$$

However, due to symmetry, we will use the contrastive loss ℓ_{cont} without any loss of generality.
For a batch size m, the SimCLR population loss is defined as:

$$L(f) = \mathop{\mathbb{E}}_{B \sim \mathcal{S}^m} \left[\frac{1}{m} \sum_{i=1}^{m} \ell_{cont}((x_i, x_i^+), X_i^-) \right]$$

where $\ell_{cont}((x_i, x_i^+), X_i^-)$ denotes the contrastive loss for the positive pair (x_i, x_i^+) with respect to its set of negative samples X_i^-. We define the SimCLR empirical loss over the dataset $S \sim \mathcal{S}^n$ as follows:

$$\widehat{L}_S(f) = \frac{1}{n} \sum_{i=1}^{n} \ell_{cont}((x_i, x_i^+), X_i^-).$$

Evaluation of representations. In SimCLR, representations are evaluated using a linear classifier to assess their quality. The multi-class classifier $g : \mathcal{X} \to \mathbb{R}^C$ incorporates the learned representation $f : \mathcal{X} \to \mathbb{R}^d$ (which remains frozen) and linear parameters $W \in \mathbb{R}^{C \times d}$, defined by $g(\cdot) := Wf(\cdot)$, where $d \in \mathbb{N}$ denotes the dimensionality of the representation. The linear classifier is learned by minimizing the supervised loss (i.e., cross-entropy loss) of the multi-class classifier g expressed as:

$$L_{\text{CE}}(f, W) = \mathbb{E}_{(x,y)} \left[-\log \frac{\exp\left(f(x)^\top w_y\right)}{\sum_{i=1}^{C} \exp\left(f(x)^\top w_i\right)} \right]$$

where $W := [w_1 \cdots w_C]^\top$, $\mathcal{D}$ denotes the data distribution, x represents the input sample, and y is its corresponding label from the set $\{1, \ldots, C\}$. The top1 accuracy of the linear classifier can be evaluated using $\texttt{top1} = 1 - R_{\text{top1}}(f, W)$, where

$$R_{\text{top1}}(f, W) = \mathbb{E}_{(x,y)} \left[\mathbb{I}_{\{f(x)^\top w_y < \max_{c \neq y} f(x)^\top w_c\}} \right].$$

PAC-Bayes Theory

PAC-Bayes theory, initially developed for simple classifiers (Seeger 2002; Catoni 2007; Germain et al. 2009), has been extended to neural network classifiers in recent years (Dziugaite and Roy 2017; Perez-Ortiz et al. 2021b), to contrastive learning (Nozawa, Germain, and Guedj 2020), and variational autoencoders (Chérief-Abdellatif et al. 2022). Here, we present the essential notions of PAC-Bayes theory and we outline the most common PAC-Bayes generalization bounds.

Notation. Let P denote a prior and Q a posterior over the weight space $\mathcal{W}$. In PAC-Bayes theory, the distance between the prior and posterior distributions is often quantified using the Kullback-Leibler (KL) divergence, defined as follows:

$$\mathrm{KL}\left(Q \| Q'\right) = \int_{\mathcal{W}} \log\left(\frac{dQ}{dQ'}\right) dQ.$$

For $q, q' \in [0, 1]$, we define the binary KL divergence as:

$$\mathrm{kl}\left(q \| q'\right) = q \log\left(\frac{q}{q'}\right) + (1 - q) \log\left(\frac{1 - q}{1 - q'}\right).$$

This measure quantifies the divergence of the Bernoulli distribution with parameter q from the Bernoulli distribution with parameter q'. Let $\mathcal{Z}$ denote an example space, $\mathcal{D}_{\mathcal{Z}}$ the distribution over $\mathcal{Z}$, and $\ell_w : \mathcal{Z} \to [0, 1]$ a loss function parameterized by $w \in \mathcal{W}$. The risk $L : \mathcal{W} \to [0, 1]$ is defined as:

$$L(w) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}_{\mathcal{Z}}} [\ell_w(z)].$$

Here, $L(w)$ represents the expected value of $\ell_w(z)$ under the distribution $\mathcal{D}_{\mathcal{Z}}$. Let n be an integer, and the empirical risk for a dataset $S = (z_1, \ldots, z_n) \in \mathcal{Z}^n$ is defined as:

$$\widehat{L}_S(w) = \frac{1}{n} \sum_{i=1}^{n} \ell_w(z_i).$$

Here, $\widehat{L}_S(w)$ computes the average loss $\ell_w(z_i)$ over the dataset S.

PAC-Bayes bounds. We extend the previously defined losses for a given weight w to losses for a given distribution Q over weights. Accordingly, the population loss of Q is defined as:

$$L(Q) = \int_{\mathcal{W}} L(w)\, Q(dw).$$

Similarly, the empirical loss of Q over a dataset S is given by:

$$\widehat{L}_S(Q) = \int_{\mathcal{W}} \widehat{L}_S(w)\, Q(dw).$$

The PAC-Bayes bounds relate the population loss $L(Q)$ to the empirical loss $\widehat{L}_S(Q)$ and other quantities through inequalities that hold with high probability. One of the fundamental results in PAC-Bayes theory is the PAC-Bayes-kl inequality, originally known as the PAC-Bayes relative entropy bound, from which various other PAC-Bayes bounds can be derived.

Theorem (PAC-Bayes-kl). For any data-free distribution P over $\mathcal{W}$ (i.e., prior), and for any $\delta \in (0, 1)$, with a probability of at least $1 - \delta$ over size-n i.i.d. samples S, simultaneously for all distributions Q over $\mathcal{W}$ (i.e., posterior), the following inequality holds:

$$\mathrm{kl}\left(\widehat{L}_S(Q) \| L(Q)\right) \leq \frac{\mathrm{KL}\left(Q \| P\right) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{n}.$$

The PAC-Bayes-kl bound can be employed to derive the classic PAC-Bayes bound using Pinsker's inequality $\mathrm{kl}(\hat{p} \| p) \geq 2(p - \hat{p})^2$.

PAC-Bayes classic bound. For any prior P over $\mathcal{W}$, and any $\delta \in (0, 1)$, with a probability of at least $1 - \delta$ over size-n i.i.d. random samples S, simultaneously for all posterior distributions Q over $\mathcal{W}$, the following inequality holds:

$$L(Q) \leq \widehat{L}_S(Q) + \sqrt{\frac{\mathrm{KL}\left(Q \| P\right) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}}.$$

Risk Certificates for Contrastive Learning

In this section, we present our main contributions: (1) using a PAC-Bayesian approach, we establish non-vacuous risk certificates for the SimCLR loss, considering its non-i.i.d. characteristics, (2) we derive a bound on the supervised loss tailored to SimCLR training, which incorporates data augmentation and temperature scaling, and (3) we extend our analysis to the contrastive zero-one risk and derive risk certificates.

Risk Certificates for Contrastive Loss

This section presents our first result: two PAC-Bayesian risk certificates for the SimCLR loss.

Notation. The SimCLR population loss under distribution Q is defined as:

$$L(Q) := \int_{\mathcal{W}} L(f_w)\, Q(dw)$$

where $f_w : \mathcal{X} \to \mathbb{S}^{d-1}$ represents the representation function parametrized by weights $w \in \mathcal{W}$. Let n be an integer and $S \sim \mathcal{S}^n$. Similarly, we define the SimCLR empirical loss of Q for the dataset S as

$$\widehat{L}_S(Q) := \int_{\mathcal{W}} \widehat{L}_S(f_w)\, Q(dw).$$

Here, $\widehat{L}_S(f_w)$ represents the empirical risk associated with the representation function f_w evaluated on the dataset S.
Let m denote the batch size and B a scaling constant defined as

$$B := \frac{1}{\tau} + \log\left((K+1)\exp\left(\frac{1}{\tau}\right) + \varepsilon\right)$$

where $K = |X^-|$ corresponds to the number of negative samples. Let $L'_S(Q)$ denote the ε-modified SimCLR empirical loss, where ℓ_{cont} is replaced by ℓ'_{cont} defined as follows:

$$-\log \frac{\text{sim}^+(x)}{\text{sim}^+(x) + \sum_{x' \in X^-} \text{sim}'(x) + 2\varepsilon}.$$

Here, $\text{sim}^+(x) = \exp\left(\frac{f(x)^\top f(x^+)}{\tau}\right)$ denotes the positive similarity term, and $\text{sim}'(x) = \exp\left(\frac{f(x)^\top f(x')}{\tau}\right)$ denotes the negative similarity term. We define the expression for ε corresponding to: (1) the SimCLR loss: $\varepsilon = c\sqrt{\frac{2(m-1)\log\left(\frac{2}{\delta}\right)}{\alpha}}$ and (2) the simplified SimCLR loss: $\varepsilon = c\sqrt{\frac{(m-1)\log\left(\frac{2}{\delta}\right)}{2\alpha}}$ where $c = e^{\frac{1}{\tau}} - e^{-\frac{1}{\tau}}$.

We introduce a novel PAC-Baye bound for the SimCLR population loss, extending the kl-PAC-Bayes bound through concentration inequalities (Hoeffding 1994).

Theorem 1 (non-i.i.d. SimCLR kl-PAC-Bayes bound). Let $\delta \in (0,1)$ be a confidence parameter. With probability at least $1-\delta$ over dataset S, for all Q:

$$\frac{1}{B}L(Q) \le \inf_{\alpha \in (0,1)} \left\{ h(L'_S(Q), C_n) + \left(\frac{\delta}{2}\right)^{\frac{1}{\alpha}} \right\},$$

where $h(L'_S(Q), C_n) = \text{kl}^{-1}\left(\frac{1}{B}L'_S(Q) + \left(\frac{\delta}{2}\right)^{\frac{1-\alpha}{\alpha}}, C_n\right)$, with $L'_S(Q)$ being the ε-modified empirical loss and $C_n = \frac{\text{KL}(Q\|P) + \log\left(\frac{\sqrt{n}}{\delta}\right)}{n}$ a complexity term.

Proof sketch. To address the non-i.i.d. characteristics of the SimCLR loss, we derive an intermediate loss satisfying the i.i.d. assumption by applying a concentration bound on the negative samples. We then apply the kl-PAC-Bayes bound to this intermediate loss and since the intermediate empirical loss is not directly computable, we derive an ε-modified empirical loss using a concentration bound. The detailed proof is provided in Appendix C.

Additionally, we can prove that the SimCLR loss satisfies the bounded difference assumption for $c_i = \frac{C}{n}$ where $C = 4 + (m-1)\log\frac{(m-1)+e^2}{m}$. This enables us to derive an additional PAC-Bayes bound using McDiarmid's inequality (McDiarmid et al. 1989).

Theorem 2 (non-i.i.d. McAllester's PAC-Bayes bound). Let $\delta \in (0,1)$ be a confidence parameter. With probability at least $1-\delta$ over dataset S, for all Q:

$$\frac{1}{C}L(Q) \le \frac{1}{C}\widehat{L}_S(Q) + \sqrt{\frac{\text{KL}(Q \parallel P) + \log\frac{2n}{\delta}}{2(n-1)}}.$$

Proof. We adapt the proofs from McAllester's PAC-Bayes bound by incorporating the bounded difference assumption (McAllester 2003a,b). The detailed proof is provided in Appendix D.

Bound on the Downstream Classification Loss

Next, we establish a bound on the supervised loss for the downstream classification task.

Notation. We use the cross-entropy loss for normalized features $f : \mathcal{X} \to \mathbb{S}^{d-1}$ as follows:

$$L_{\text{CE}}(f, W) = \mathbb{E}_{(x,y)}\left[-\log \frac{\exp\left(f(x)^\top w_y\right)}{\sum_{i=1}^{C} \exp\left(f(x)^\top w_i\right)}\right].$$

We now establish a lower bound on the SimCLR loss, refining the bound from Bao et al. (Bao, Nagano, and Nozawa 2022). This bound does not rely on the conditional independence assumption and is applicable across different temperature and batch size parameters.

Theorem 3 (upper-bound on the linear classifier loss). For all $f : \mathcal{X} \to \mathbb{S}^{d-1}$, the following inequality holds:

$$\min_{W \in \mathbb{R}^{C \times d}} L_{\text{CE}}(f, W) \le \min \left\{ \begin{array}{l} \beta(f, \sigma) \\ \tau\beta(f, \sigma) + \alpha \end{array} \right.$$

with $\beta(f,\sigma) = \frac{\sigma}{\tau} + L(f) + \Delta$.
Here, $\sigma = \mathbb{E}_{(x,y)}\left[\|f(x) - \mu_y\|_2\right]$ represents the intra-class feature deviation, while for K the number of negative samples, $\Delta = \log\left(\frac{C}{K}\cosh^2\left(\frac{1}{\tau}\right)\right)$ denotes a constant. The term α is given by:

$$\alpha = \log(C) + \min\{0, \log(\cosh^2(1)) - \tau\Delta\}.$$

Remark 1. If we wish to eliminate the dependency on σ, given that the features are normalized, we observe $\sigma \leq 2$.

Remark 2. This theorem can be extended to include a simple projection head of the form $(I_k, 0)$ by considering the contrastive loss on the projected features $L(f_1)$ instead of $L(f)$ and a modified version of the supervised loss $L_{\mathrm{CE}}(f, W)$:

$$\mathbb{E}_{x,y\sim\mathcal{D}}\left[-\log\frac{\exp\left(f_1(x)^\top w_y^{(1)} + f_2(x)^\top w_y^{(2)}\right)}{\sum_{i=1}^{C}\exp\left(f_1(x)^\top w_i^{(1)} + f_2(x)^\top w_i^{(2)}\right)}\right]$$

where $f_2(x) \in \mathbb{R}^{d-k}$ is defined such that $f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix}$.

Proof sketch. The first part of the theorem is derived by directly combining the techniques from Bao et al. (Bao, Nagano, and Nozawa 2022) and Wang et al. (Wang et al. 2022) and the second part is obtained by refining the first part and eliminating the temperature scaling in the log-sum-exp term. The detailed proof is provided in Appendix E.

Risk Certificate for Contrastive Zero-One Risk

In the previous sections, we demonstrated that non-vacuous risk certificates can be obtained for the SimCLR loss and that the SimCLR loss acts as a surrogate loss for downstream classification. Similarly to Nozawa, Germain, and Guedj, we extend our analysis to the contrastive zero-one risk defined as follows:

$$r((x, x^+), X^-) = \frac{1}{|X^-|}\sum_{x' \in X^-} \mathbb{I}_{\{f(x)^\top f(x^+) < f(x)^\top f(x')\}}.$$

Remark. The SimCLR loss can be seen as a surrogate loss for the contrastive zero-one risk.

We denote $R(f)$ as the population contrastive zero-one risk and $\widehat{R}_S(f)$ as its empirical counterpart. We extend Theorems 1 and 2 to the contrastive zero-one risk. First, we present Theorem 4 corresponding to the extension of Theorem 1:

Theorem 4 (non-i.i.d. zero-one kl-PAC-Bayes bound). Let $\delta \in (0,1)$ be a confidence parameter. With probability at least $1-\delta$ over dataset S, for all Q:

$$R(Q) \leq \inf_{\alpha\in(0,1)}\left\{h\left(\widehat{R}_S(Q), C_n\right) + \gamma + \left(\frac{\delta}{2}\right)^{\frac{1}{\alpha}}\right\},$$

with

$$\begin{cases} \gamma = \sqrt{\dfrac{\log\left(\frac{2}{\delta}\right)}{2(m-1)\alpha}} \\ C_n = \dfrac{\mathrm{KL}\left(Q\|P\right) + \log\left(\frac{\sqrt{n}}{\delta}\right)}{n} \\ h\left(\widehat{R}_S(Q), C_n\right) = \mathrm{kl}^{-1}\left(\widehat{R}_S(Q) + \gamma + \left(\dfrac{\delta}{2}\right)^{\frac{1-\alpha}{\alpha}}, C_n\right) \end{cases}$$

Proof sketch. The proof uses the same approach as Theorem 1 and is provided in Appendix F.

Next, we present Theorem 5 corresponding to the extension of Theorem 2:

Theorem 5 (non-i.i.d. McAllester's PAC-Bayes bound). Let $\delta \in (0,1)$ be a confidence parameter. With probability at least $1-\delta$ over dataset S, for all Q:

$$R(Q) \leq \widehat{R}_S(Q) + 2\sqrt{\frac{\mathrm{KL}(Q \parallel P) + \log\frac{2n}{\delta}}{2(n-1)}}.$$

Proof. The proof uses the same approach as Theorem 2 with $C = 2$ and is provided in Appendix G.

Experiments

In this section, we describe the experimental setup and empirically demonstrate that our risk certificates improved upon previous risk ones through experiments on the CIFAR-10 dataset. The code for our experiments is available in PyTorch (Paszke et al. 2017). Experimental results on the MNIST dataset and additional experimental details are available in Appendix H.

Datasets and models. We use two popular benchmarks: (1) CIFAR-10, which consists of 50,000 training images and 10,000 test images (Krizhevsky and Hinton 2009), and (2) MNIST, which consists of 60,000 training images and 10,000 test images (LeCun, Cortes, and Burges 2010), as provided in torchvision (Marcel and Rodriguez 2010). The images are preprocessed by normalizing all pixels per channel based on the training data. Data augmentation includes random cropping, resizing (with random flipping), and color distortions, as detailed in Appendix H (Chen et al. 2020). We employ a 7-layer convolutional neural network (CNN) with max-pooling every two layers for CIFAR-10 experiments and a 3-layer CNN for MNIST experiments. We use a 2-layer MLP projection head to project to a 128-dimensional latent space, with a feature dimensionality of 2048 for CIFAR-10 and 512 for MNIST. ReLU activations are used in each hidden layer. The mean parameters μ_0 of the prior are initialized randomly from a truncated centered Gaussian distribution with a standard deviation of $1/\sqrt{n_{\text{in}}}$, where n_{in} is the dimension of the inputs to a particular layer, truncating at ± 2 standard deviations (Perez-Ortiz et al. 2021a). The prior distribution scale hyperparameter (standard deviation σ_0) is selected from $\{0.01, 0.05, 0.1\}$.

PAC-Bayes Learning. The learning and certification strategy involves three steps: (1) choose or learn a prior from a subset of the dataset; (2) learn a posterior on the entire training dataset; (3) evaluate the risk certificate for the posterior on a subset of the dataset independent of the prior. We experiment with two types of priors: informed and random. The informed prior is learned using a subset of the training dataset via empirical risk minimization or PAC-Bayes objective minimization. The posterior is initialized to the prior and learned using the entire training dataset by PAC-Bayes objective minimization. We use the *PAC-Bayes with Backprop* (PBB) procedure (Perez-Ortiz et al. 2021b) and the following f_{classic} objective :

$$f_{\text{classic}}(Q) = \frac{1}{B}\widehat{L}_S(Q) + \sqrt{\frac{\eta \mathrm{KL}\left(Q \| P\right) + \log\left(\frac{\sqrt{n}}{\delta}\right)}{2n}}$$

where η in $[0, 1]$ is coefficient introduced to control the influence of the KL term in the training objective, referred to as the KL penalty. We use a KL penalty term of 10^{-6} for learning prior and no penalty term for learning the posterior. We use SGD with momentum as optimizer and we perform a grid search for momentum values in $\{0.8, 0.85, 0.90, 0.95\}$ and learning rates in $\{0.1, 0.5, 1.0, 1.5\}$. Training was conducted for 100 epochs and we select the hyperparameters that give the best risk certificates. We experiment with different temperatures selected $\{0.2, 0.5, 0.7, 1\}$. Unless otherwise specified, experiments are run using a probabilistic prior with the simplified SimCLR loss, a batch size of $m = 250$, and 80% of the data for training the prior.

Numerical Risk Certificates. Since $\widehat{L}_S(Q)$ is intractable, the final risk certificates are computed using Monte Carlo weight sampling. Specifically, we approximate $\widehat{L}_S(Q)$ using $\widehat{Q}_p = \sum_{j=1}^{p} \delta_{W_j}$, where $W_1, \ldots, W_p \sim Q$ are i.i.d. samples. We compute all risk certificates with $\delta = 0.04$, and $p = 100$ Monte Carlo model samples. We report the risk certificates for both the contrastive loss and the contrastive zero-one risk using Theorems 1 and 2 for the contrastive loss, and Theorems 4 and 5 for the zero-one risk. To find the best value of α for Theorems 1 and 4, we perform a grid search over $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. We observe that $\alpha = 0.4$ provides the tightest risk certificates. These risk certificates are compared with existing ones, including the kl-PAC-Bayes bound over i.i.d. batches, Catoni's PAC-Bayes bound over i.i.d. batches ((Nozawa, Germain, and Guedj 2020)), the classic PAC-Bayes bound over i.i.d. batches, and the f-divergence PAC-Bayes bound. For additional details, see Appendix A and Appendix B.

Linear Evaluation. We assess the quality of the learned representations through linear evaluation (Chen et al. 2020): we report the cross-entropy loss and top-1 accuracy of linear classifiers trained on features either before or after the projection head. The classifiers are trained for 20 epochs on the standard image classification task ($C = 10$) using the Adam optimizer with a learning rate of 0.01. Finally, we report the bounds on the downstream classification loss (cross-entropy loss of a trained linear classifier), derived from Theorem 3 and compare it with the existing bound (Bao, Nagano, and Nozawa 2022).

Results. Table 2 demonstrates that our proposed risk certificates for the SimCLR loss are non-vacuous and significantly outperform existing certificates on CIFAR-10, closely aligning with the corresponding test losses. Interestingly, Theorem 1 yields tighter certificates for $\tau \leq 0.5$, while Theorem 2 is more effective for $\tau > 0.5$. We also observe that Catoni's bound is significantly tighter than the kl-PAC-Bayes bound, which is consistent with its known advantage when KL/n is large (Zhou et al. 2018). Unsurprisingly, the classic PAC-Bayes bound is looser than both of these bounds. Additionally, Table 3 shows that Theorem 3 improves upon the bound from Bao et al. which results in exponential growth when $\tau \leq 0.5$ (Bao, Nagano, and Nozawa 2022). Moreover, models trained using *PAC-Bayes by Backprop* achieve competitive top1 accuracy. Table 4 further illustrates

		SimCLR Loss			
		$\tau = 1$	$\tau = 0.7$	$\tau = 0.5$	$\tau = 0.2$
	Test Loss	4.945	4.640	4.257	2.7076
Risk Certificate	kl bound (iid)	7.164	7.674	8.246	9.954
	Catoni's bound (iid)	7.095	7.556	7.959	9.446
	Classic bound (iid)	8.475	8.698	8.91	10.166
	Nozawa et al.	27.03	30.138	33.27	48.472
	Th. 1 (ours)	5.537	5.491	5.492	6.223
	Th. 2 (ours)	5.203	5.328	6.269	43.779
	KL/n	0.0013	0.0014	0.0014	0.0013

Table 2: Comparison of risk certificates for the SimCLR loss using different PAC-Bayes bounds for varying temperature values on CIFAR-10. *kl bound (iid)* refers to the standard kl-PAC-Bayes bound computed over i.i.d. batches, *Catoni's bound (iid)* refers to Catoni's PAC-Bayes bound computed over i.i.d. batches (Nozawa, Germain, and Guedj 2020), *Classic bound (iid)* refers to the classic PAC-Bayes bound computed over i.i.d. batches, and *Nozawa et al.* refers to the PAC-Bayes bound based on f-divergence. Although Nozawa et al.'s bound uses χ^2 divergence, we use KL divergence, which already results in vacuous bounds and would not improve with χ^2, since $\mathrm{KL}(P\|Q) \leq \chi^2(P\|Q)$. *Th. 1* and *Th. 2* are the PAC-Bayes risk certificates computed using our Theorem 1 and Theorem 2, respectively. We report test losses and observe that our bounds are remarkably tight. We also report the complexity term, KL/n, where KL represents the Kullback-Leibler divergence between the prior and posterior distributions, and n is the dataset size used to compute the risk certificate.

		$\tau = 1$	$\tau = 0.7$	$\tau = 0.5$	$\tau = 0.2$
	Bao et al.	3.2720	4.0274	5.3015	12.588
	Th. 3 (ours)	3.2720	4.0274	4.9533	4.6079
Proj.	Sup. Loss	1.903	1.837	1.7971	1.7547
	top-1	0.4868	0.5710	0.6205	0.6677
	Sup. Loss	1.765	1.718	1.705	1.699
	top-1	0.6350	0.6939	0.7102	0.7278

Table 3: Comparison of upper bounds on downstream classification loss on CIFAR-10. We compare the original bound from Bao et al. with our refined bound (Theorem 3). The supervised loss of the linear classifier trained on the projected features is reported, as it is directly related to the theoretical upper bound. Additionally, we report the supervised loss of the linear classifier trained on the features after removing the projection head. We empirically observe that the supervised loss of a linear classifier trained on the full features (without projection) is consistently lower than the loss of a linear classifier trained on the projected features, aligning with previous findings (Chen et al. 2020). For reference, we also include top-1 accuracy.

that our risk certificates for contrastive zero-one risk are notably tight, surpassing existing certificates. Additionally, we observe that theorem 5 consistently outperforms Theorem 4. Overall, our risk certificates are competitive, even for low temperatures—a known challenge. In PAC-Bayes learning, we observe that a randomly initialized prior leads to poor convergence and sub-optimal risk certificates, while an informed prior results in better performance and tighter certificates. A probabilistic prior seems to provide tighter certificates than a deterministic one. However, PAC-Bayes learning in models with a large number of parameters remains challenging and warrants further investigation (most studies focus on 2 or 3 hidden layers, while our largest model has 7).

Discussion and Future Work

In summary, we have presented novel PAC-Bayesian risk certificates tailored for the SimCLR framework, addressing inherent challenges such as non-i.i.d. characteristics of the SimCLR loss and the integration of data augmentation and temperature scaling. Our experiments on CIFAR-10 and MNIST show that our bounds yield non-vacuous risk certificates and significantly outperform previous ones across various temperature settings.

Bounding techniques. Theorems 1 and 4 rely on concentration bounds to apply the kl-PAC-Bayes bound in an i.i.d. setting. Although the kl-PAC-Bayes bound was selected for its tightness, any bound that respects our proof's assumptions could be applicable. For instance, Catoni's bound, which is often tighter when KL/n is large, could be applied since the function $1 - \exp(-x)$ is non-decreasing, similar to the kl function when fixing the second argument. An overview of Theorems 1 and 4 using Catoni's bound is provided in Appendix J. Moreover, theorems 2 and 5 were obtained using McDiarmid's inequality,

		Contrastive 0-1 Risk			
		$\tau = 1$	$\tau = 0.7$	$\tau = 0.5$	$\tau = 0.2$
	Test Loss	0.0601	0.0433	0.0324	0.0199
Risk Certificate	kl bound (iid)	0.497	0.488	0.47	0.432
	Catoni's bound (iid)	0.469	0.466	0.435	0.417
	Classic bound (iid)	0.542	0.54	0.53	0.505
	Nozawa et al.	3.009	3.099	3.139	2.973
	Th. 4 (ours)	0.367	0.353	0.342	0.329
	Th. 5 (ours)	0.129	0.117	0.107	0.093
	KL $/n$	0.0013	0.0014	0.0014	0.0013

Table 4: Comparison of risk certificates for the contrastive zero-one risk with various PAC-Bayes bounds and temperature values on CIFAR-10. For additional details, see Table 1. *Th. 4* and *Th. 5* refer to the PAC-Bayes risk certificates derived from Theorems 4 and 5, respectively. We report corresponding test losses and observe that our risk certificates are notably tight, particularly those from Theorem 5 and seem promising for selecting the temperature parameter τ.

integrated into McAllester's PAC-Bayes bound. Since this variant is known to be less tight than the kl bound or Catoni's bound, it would be interesting to explore whether McDiarmid's inequality can be incorporated into these tighter bounds, potentially improving Theorems 2 and 5.

Extension to other non-i.i.d. losses. While this paper primarily focuses on contrastive learning and specifically on the SimCLR framework, the approaches we propose to address the non-i.i.d. characteristics of the SimCLR loss can be readily applied to other loss presenting similar non-i.i.d. characteristics, such as ranking losses (Chen et al. 2009) and Barlow Twins losses (Zbontar et al. 2021). On the one hand, our non-i.i.d. McAllester PAC-Bayes bound offers a clear and straightforward formulation, requiring only a bounded difference assumption. On the other hand, our non-i.i.d. kl-PAC-Bayes bounds, while more opaque, require only a Hoeffding's assumption on the dependent terms.

Impact of temperature scaling. Temperature scaling in the SimCLR loss remains challenging as the scaling constant loosens the PAC-Bayes bounds, suggesting the need for more adapted PAC-Bayes bounds for this type of loss. Regarding the bound on the downstream classification loss, our approach better handles smaller temperatures than previous bounds, though it still struggles to perfectly align with downstream classification losses at low temperatures.

Projection head. Empirically, we observe that classification loss is lower when features are used without a projection head compared to with one, yet the theoretical role of the projection head can not fully be understood with our downstream classification bound. In this work, we proposed an approach to integrate a simple projection head into our bound, and it would be interesting to link this with the work of Jing et al., which suggests that a fixed low-rank diagonal projector might suffice instead of a trainable projection head (Jing et al. 2021).

PAC-Bayes learning. Additionally, our models trained with *PAC-Bayes by Backprop* achieve competitive accuracy despite using a much smaller model (7-layer CNN vs. 50-layer CNN) and training for fewer epochs (100 vs. 500) compared to Chen et al. (Chen et al. 2020). Although this paper focuses on deriving better risk certificates rather than improving the PAC-Bayes learning algorithm, there is a need to extend the PAC-Bayes paradigm to large-scale neural networks, such as ResNet50.

Summary of PAC-Bayes Bounds for the SimCLR Loss

Previous PAC-Bayes Bounds

Below is the list of previous PAC-Bayes bounds applicable to the SimCLR loss, where n is the dataset size and m is the batch size:

- **Catoni's PAC-Bayes Bound (over i.i.d. batches):**

$$\frac{1}{B_\ell} L(Q) \leq \inf_{\lambda > 0} \left\{ \frac{1 - \exp\left(-\frac{\lambda}{B_\ell} \widehat{L}_S(Q) - m^{\frac{\mathrm{KL}(Q\|\mathcal{P}) + \log\frac{1}{\delta}}{n}} \right)}{1 - \exp(-\lambda)} \right\}$$

- **kl-PAC-Bayes Bound (over i.i.d. batches):**

$$\frac{1}{B_\ell}L(Q) \leq \mathrm{kl}^{-1}\left(\frac{1}{B_\ell}\widehat{L}_S(Q), m\frac{\mathrm{KL}(Q \parallel \mathcal{P}) + \log\frac{2\sqrt{n}}{\delta\sqrt{m}}}{n}\right)$$

- **Classic PAC-Bayes Bound (over i.i.d. batches):**

$$\frac{1}{B_\ell}L(Q) \leq \frac{1}{B_\ell}\widehat{L}_S(Q) + \sqrt{m\frac{\mathrm{KL}(Q \parallel \mathcal{P}) + \log\frac{2\sqrt{n}}{\delta\sqrt{m}}}{2n}}$$

- **f-divergence Bound (Nozawa et al.'s bound):**

$$\frac{1}{B_\ell}L(Q) \leq \frac{1}{B_\ell}\widehat{L}_S(Q) + \sqrt{\frac{m-1}{n\delta}\left(\chi^2(Q \parallel P) + 1\right)}$$

Proof for the f-divergence Bound. We adapt the PAC-Bayes f-divergence presented by Nozawa, Germain, and Guedj to the SimCLR loss (Nozawa, Germain, and Guedj 2020). We have:

$$L(Q) \leq L_S(Q) + \sqrt{\frac{\mathcal{M}_2}{\delta}\left(\chi^2(Q \parallel P) + 1\right)},$$

where $\mathcal{M}_2 = \mathbb{E}_{\mathbf{f}\sim\mathcal{P}}\mathbb{E}_{S\sim\mathcal{S}^m}\left(\left|L(\mathbf{f}) - \widehat{L}_S(\mathbf{f})\right|^2\right)$.
$\mathcal{M}_2$ can be upper-bounded using the following covariance:

$$\mathrm{Cov}\left(\ell\left(\mathbf{z}_i\right), \ell\left(\mathbf{z}_j\right)\right)\begin{cases}\leq B_\ell^2 & \text{if } i, j \text{ are in the same batch}\\ = 0 & \text{otherwise}\end{cases}$$

where $\ell\left(\mathbf{z}_i\right) = \ell_{cont}((x_i, x_i^+), X_i^-)$. Indeed, we have (Alquier and Guedj 2018):

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\ell\left(\mathbf{z}_i\right) - \mathbb{E}_{X_i^-}\left[\ell\left(\mathbf{z}_i\right)\right]\right)^2\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}\left[\ell\left(\mathbf{z}_i\right), \ell\left(\mathbf{z}_j\right)\right]$$

which implies for the simplified SimCLR loss:

$$\mathcal{M}_2 \leq \frac{1}{n^2}\sum_{i=1}^{n}(m-1)B_\ell^2 = \frac{m-1}{n}B_\ell^2.$$

Our novel PAC-Bayes Bounds

Below is the list of our novel PAC-Bayes bounds applicable to the SimCLR loss:
- **Theorem 1 (ours):**

$$\frac{1}{B}L(Q) \leq \inf_{\alpha\in(0,1)}\left\{\mathrm{kl}^{-1}\left(\frac{1}{B}L'_S(Q) + \left(\frac{\delta}{2}\right)^{\frac{1-\alpha}{\alpha}}, \frac{\mathrm{KL}(Q\|P) + \log\left(\frac{\sqrt{n}}{\delta}\right)}{n}\right) + \left(\frac{\delta}{2}\right)^{\frac{1}{\alpha}}\right\},$$

where $L'_S(Q)$ is the ε-modified empirical loss with $\varepsilon = c\sqrt{\frac{(m-1)\log\left(\frac{2}{\delta}\right)}{2\alpha}}$ for the simplified SimCLR loss, and $\varepsilon = c\sqrt{\frac{2(m-1)\log\left(\frac{2}{\delta}\right)}{\alpha}}$ for SimCLR loss.
- **Theorem 2 (ours):**

$$\frac{1}{C}L(Q) \leq \frac{1}{C}\widehat{L}_S(Q) + \sqrt{\frac{\mathrm{KL}(Q \parallel \mathcal{P}) + \log\frac{2n}{\delta}}{2(n-1)}}$$

Summary of PAC-Bayes Bounds for the Contrastive Zero-One Risk

Below is the list of PAC-Bayes bounds applicable to the contrastive zero-one risk, where n is the dataset size and m is the batch size. Since the kl-PAC-Bayes bound is tighter than both the classic and Catoni's bounds, we present only the kl-PAC-Bayes bound.

- kl-PAC-Bayes Bound (over i.i.d. batches):

$$R(Q) \leq \mathrm{kl}^{-1}\left(\widehat{R}_S(Q), m\frac{\mathrm{KL}(Q \parallel \mathcal{P}) + \log\frac{2\sqrt{n}}{\delta\sqrt{m}}}{n}\right)$$

- Catoni's PAC-Bayes Bound (over i.i.d. batches):

$$R(Q) \leq \inf_{\lambda > 0}\left\{\frac{1 - \exp\left(-\lambda \widehat{R}_S(Q) - m\frac{\mathrm{KL}(Q\|\mathcal{P}) + \log\frac{1}{\delta}}{n}\right)}{1 - \exp(-\lambda)}\right\}$$

- Classic PAC-Bayes Bound (over i.i.d. batches):

$$R(Q) \leq \widehat{R}_S(Q) + \sqrt{m\frac{\mathrm{KL}(Q \parallel \mathcal{P}) + \log\frac{2\sqrt{n}}{\delta\sqrt{m}}}{2n}}$$

- f-divergence PAC-Bayes Bound (Nozawa's bound):

$$R(Q) \leq \widehat{R}_S(Q) + \sqrt{\frac{m-1}{n\delta}\left(\chi^2(Q \parallel P) + 1\right)}$$

- Theorem 4 (ours):

$$R(Q) \leq \inf_{\alpha \in (0,1)}\left\{\mathrm{kl}^{-1}\left(\widehat{R}_S(Q) + \gamma + \left(\frac{\delta}{2}\right)^{\frac{1-\alpha}{\alpha}}, \frac{\mathrm{KL}\left(Q\|P\right) + \log\left(\frac{\sqrt{n}}{\delta}\right)}{n}\right) + \gamma + \left(\frac{\delta}{2}\right)^{\frac{1}{\alpha}}\right\},$$

with $\gamma = \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2(m-1)\alpha}}$.

- Theorem 5 (ours):

$$R(Q) \leq \widehat{R}_S(Q) + 2\sqrt{\frac{\mathrm{KL}(Q \parallel P) + \log\frac{2n}{\delta}}{2(n-1)}}$$

Proof of Theorem 1

In this section, we provide a detailed proof of Theorem 1.

Theorem 1. Let $\delta \in (0,1)$ be a confidence parameter. With probability at least $1-\delta$ over dataset S, for all Q:

$$\frac{1}{B}L(Q) \leq \inf_{\alpha \in (0,1)}\left\{\mathrm{kl}^{-1}\left(\frac{1}{B}L'_S(Q) + \left(\frac{\delta}{2}\right)^{\frac{1-\alpha}{\alpha}}, \frac{\mathrm{KL}(Q\|P) + \log\left(\frac{\sqrt{n}}{\delta}\right)}{n}\right) + \left(\frac{\delta}{2}\right)^{\frac{1}{\alpha}}\right\},$$

where $L'_S(Q)$ is the ε-modified empirical loss with $\varepsilon = c\sqrt{\frac{(m-1)\log\left(\frac{2}{\delta}\right)}{2\alpha}}$ corresponding to the simplified SimCLR loss, $\varepsilon = c\sqrt{\frac{2(m-1)\log\left(\frac{2}{\delta}\right)}{\alpha}}$ corresponding to the SimCLR loss. The scaling constant B is defined as $B := \frac{1}{\tau} + \log\left((K+1)\exp\left(\frac{1}{\tau}\right) + \varepsilon\right)$ for $K = |X^-|$ the number of negative samples.

Proof sketch. The SimCLR loss exhibits non-i.i.d. characteristics, violating the PAC-Bayes bounds assumption due to the negative samples in each positive pair's subloss. To address this, we derive an intermediate loss satisfying the i.i.d. assumption by applying a concentration bound on the negative samples and show that the SimCLR loss can be upper-bounded by this intermediate loss (lemmas 1 and 2). We then apply the PAC-Bayes bound to this intermediate loss (lemma 3). Since the intermediate empirical loss is not directly computable, we derive an ε-modified empirical loss using a concentration bound and show that the intermediate empirical loss can be upper-bounded by the ε-modified empirical loss with high probability over the dataset S (lemmas 4 and 5). Finally, we combine all results with a union bound (lemma 6) and take the infimum over α to complete the proof.

To begin, we apply a Hoeffding's inequality to the term involving the negative samples, specifically:

$$S_{m-1}(x) := \sum_{x' \in X^-} \exp\left(\frac{f(x)^\top f(x')}{\tau}\right).$$

The result is explicitly stated in the following lemma.

Lemma 1 (concentration bound on the negative samples). Let $c = e^{\frac{1}{\tau}} - e^{-\frac{1}{\tau}}$. For all $\delta_c \in (0, 1)$,

$$\mathbb{P}\left(S_{m-1}(x) - \mathbb{E}[S_{m-1}(x)] \geq \varepsilon \mid x\right) \leq \delta_c,$$

with $\varepsilon = c\sqrt{\frac{(m-1)\log\left(\frac{1}{\delta_c}\right)}{2}}$ corresponding to the simplified SimCLR loss and $\varepsilon = c\sqrt{2(m-1)\log\left(\frac{1}{\delta_c}\right)}$ for the SimCLR loss.

Next, using lemma 1, we upper-bound the SimCLR loss defined as

$$L(f) := \mathbb{E}_{S \sim \mathcal{S}^m}\left[\frac{1}{m}\sum_{i=1}^{m} \ell_{\text{cont}}((x_i, x_i^+), X_i^-)\right].$$

by the following intermediate loss

$$\widetilde{L}(f) := \mathbb{E}_{(x,x^+)\sim\mathcal{S}}\left[\tilde{\ell}(x, x^+)\right],$$

where

$$\tilde{\ell}(x, x^+) := -\log \frac{\exp\left(\frac{f(x)^\top f(x^+)}{\tau}\right)}{\exp\left(\frac{f(x)^\top f(x^+)}{\tau}\right) + \mu(x) + \varepsilon}.$$

Lemma 2 (upper-bound on the SimCLR population loss by an intermediate loss). For all $f \sim Q$,

$$L(f) \leq \widetilde{L}(f) + B_\ell \delta_c.$$

Now, we extend the definition of the previously defined losses and obtain the intermediate population loss over Q:

$$\widetilde{L}(Q) = \mathbb{E}_{f\sim\mathcal{Q}}\left[\widetilde{L}(f)\right],$$

and the empirical counterpart:

$$\widetilde{L}_S(Q) = \mathbb{E}_{f\sim\mathcal{Q}}\left[\widetilde{L}_S(f)\right].$$

We obtain:

$$L(Q) \leq \widetilde{L}(Q) + B_\ell \delta_c.$$

We will now derive a PAC-Bayes-kl bound for the intermediate population loss.

Lemma 3 (PAC-Bayes-kl bound for the intermediate loss). Given a prior P over $\mathcal{F}$ and $\delta_{\text{pb}} \in (0, 1)$, with probability at least $1 - \delta_{\text{pb}}$ over training i.i.d. samples $S \sim \mathcal{S}^n$, for all Q over $\mathcal{F}$, we have:

$$\frac{1}{B}\widetilde{L}(Q) \leq \text{kl}^{-1}\left(\frac{1}{B}\widetilde{L}_S(Q), \frac{\text{KL}\left(Q\|P\right) + \log\left(\frac{2\sqrt{n}}{\delta_{\text{pb}}}\right)}{n}\right),$$

where $B := \frac{1}{\tau} + \log\left((K+1)\exp\left(\frac{1}{\tau}\right) + \varepsilon\right)$ and $K = |X^-|$ corresponds to the number of negative samples.

Next, given the property of kl^{-1} as a monotonically increasing function of its first argument when fixing the second argument (Perez-Ortiz et al. 2021b), we can upper bound the intermediate empirical loss $\widetilde{L}_S(Q)$ by a term similar to the SimCLR empirical loss called the ε-modified loss and denoted as $L'_S(Q)$ and where each individual loss term is defined as:

$$\ell'(x, x^+, X^-) := -\log \frac{\text{sim}^+(x)}{\text{sim}^+(x) + S_{m-1}(x) + 2\varepsilon}.$$

Lemma 4 (upper-bound on the intermediate loss). Let $\alpha \in (0,1)$. With probability at least $1-\delta_c^\alpha$ over dataset S:

$$\widetilde{L}_S(Q) \leq B\delta_c^{1-\alpha} + L'_S(Q).$$

To prove lemma 4, we show the following useful lemma:

Lemma 5. Let $\alpha \in (0,1)$. If

$$\mathbb{P}_{X^-}\left[S_{m-1}(x) - \mu(x) \leq -\varepsilon \mid w, x\right] \leq \delta_c$$

then, with probability at least $1-\delta_c^\alpha$ over dataset S,

$$\frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{f\sim Q}\left[\mathbb{I}_{\{S_{m-1}(x_i)+\varepsilon \leq \mu(x_i)\}} \mid x_i, X_i^-\right] \leq \delta_c^{1-\alpha}.$$

Finally, we combine the previous results with a union bound.

Lemma 6. Let $\alpha \in (0,1)$. Given a prior P over $\mathcal{F}$, with probability at least $1-\delta$ over dataset S, for all Q over $\mathcal{F}$:

$$\frac{1}{B}L(Q) \leq \mathrm{kl}^{-1}\left(\frac{1}{B}L'_S(Q) + \left(\frac{\delta}{2}\right)^{\frac{1-\alpha}{\alpha}}, \frac{\mathrm{KL}\left(Q\|P\right) + \log\left(\frac{\sqrt{n}}{\delta}\right)}{n}\right) + \left(\frac{\delta}{2}\right)^{\frac{1}{\alpha}},$$

where the ε corresponding to the simplified SimCLR loss is defined as:

$$\varepsilon = c\sqrt{\frac{(m-1)\log\left(\frac{2}{\delta}\right)}{2\alpha}},$$

and for the SimCLR loss:

$$\varepsilon = c\sqrt{\frac{2(m-1)\log\left(\frac{2}{\delta}\right)}{\alpha}}.$$

Proof of lemma 1. Conditioned on x, we have a sum of bounded and independent variables for the simplified SimCLR loss: each variable is lower-bounded by $a = e^{-\frac{1}{\tau}}$ and upper-bounded by $b = e^{\frac{1}{\tau}}$ since

$$-\frac{1}{\tau} \leq \frac{f(x)^\top f(x')}{\tau} \leq \frac{1}{\tau}.$$

Thus, we define $c := e^{\frac{1}{\tau}} - e^{-\frac{1}{\tau}}$ and use the Hoeffding inequality: for all $\varepsilon > 0$,

$$\mathbb{P}\left(S_{m-1}(x) - \mathbb{E}[S_{m-1}(x)] \geq \varepsilon \mid x\right) \leq \exp\left(-\frac{2\varepsilon^2}{(m-1)c^2}\right).$$

For convenience, we set $\delta_c := \exp\left(-\frac{2\varepsilon^2}{(m-1)c^2}\right)$ and derive an expression for ε:

$$\varepsilon = c\sqrt{\frac{(m-1)\log\left(\frac{1}{\delta_c}\right)}{2}}.$$

We derive a similar concentration bound for the SimCLR loss by grouping terms involving augmented views of the same sample into one variable:

$$\exp\left(\frac{f(x)^\top f(x')}{\tau}\right) + \exp\left(\frac{f(x)^\top f(x'^+)}{\tau}\right).$$

Here, the constant c becomes $2c$, and ε is given by

$$\varepsilon = c\sqrt{2(m-1)\log\left(\frac{1}{\delta_c}\right)}.$$

Proof of lemma 2. Let $f \sim Q$. We define $\mathrm{sim}^+(x) := \exp\left(\frac{f(x)^\top f(x^+)}{\tau}\right)$ and $\mu(x) := \mathbb{E}\left[S_{m-1}(x)\right]$. For an augmented sample x, we can upper-bound $\ell_{\text{cont}}((x, x^+), X^-)$ by the following term:

$$-\log\left(\frac{\mathrm{sim}^+(x)}{\mathrm{sim}^+(x)+\mu(x)+\varepsilon}\right)\mathbb{I}_{\{S_{m-1}(x)<\mu(x)+\varepsilon\}}-\log\left(\frac{\mathrm{sim}^+(x)}{\mathrm{sim}^+(x)+S_{m-1}(x)}\right)\mathbb{I}_{\{S_{m-1}(x)\geq\mu(x)+\varepsilon\}}.$$

Let B_ℓ be defined as $B_\ell := \frac{2}{\tau}+\log(m)$ for the simplified SimCLR loss and $B_\ell := \frac{2}{\tau}+\log(2m-1)$ for the SimCLR loss. We can show that

$$\ell_{\mathrm{cont}}((x,x^+),X^-)\leq B_\ell.$$

Using the previously established Hoeffding's inequality,

$$\mathbb{P}\left(S_{m-1}(x)\geq\mu(x)+\varepsilon\mid x\right)\leq\delta_c,$$

and noting that $\mathbb{I}_{\{S_{m-1}(x)<\mu(x)+\varepsilon\}}\leq 1$, we obtain

$$\ell_{\mathrm{cont}}((x,x^+),X^-)\leq-\log\frac{\mathrm{sim}^+(x)}{\mathrm{sim}^+(x)+\mu(x)+\varepsilon}+B_\ell\mathbb{I}_{\{S_{m-1}(x)\geq\mu(x)+\varepsilon\}}.$$

Plugging this into the SimCLR population loss, we have

$$\begin{aligned}L(f)&\leq\mathbb{E}_{S\sim\mathcal{S}^m}\left[\frac{1}{m}\sum_{i=1}^{m}-\log\frac{\mathrm{sim}^+(x_i)}{\mathrm{sim}^+(x_i)+\mu(x_i)+\varepsilon}\right]+B_\ell\delta_c\\&\leq\mathbb{E}_{(x,x^+)\sim\mathcal{S}}\left[-\log\frac{\exp\left(\frac{f(x)^\top f(x^+)}{\tau}\right)}{\exp\left(\frac{f(x)^\top f(x^+)}{\tau}\right)+\mu(x)+\varepsilon}\right]+B_\ell\delta_c.\end{aligned}$$

We denote $\tilde{\ell}$ as the intermediate loss of a positive pair (x,x^+):

$$\tilde{\ell}(x,x^+):=-\log\frac{\exp\left(\frac{f(x)^\top f(x^+)}{\tau}\right)}{\exp\left(\frac{f(x)^\top f(x^+)}{\tau}\right)+\mu(x)+\varepsilon}.$$

Consequently, we define the intermediate population loss:

$$\widetilde{L}(f):=\mathbb{E}_{(x,x^+)\sim\mathcal{S}}\left[\tilde{\ell}(x,x^+)\right],$$

and the intermediate empirical loss for a dataset $S\sim\mathcal{S}^n$:

$$\widetilde{L}_S(f):=\frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}(x_i,x_i^+).$$

All in all, we have shown that the SimCLR population loss can be upper-bounded by an intermediate loss that maintains the i.i.d. assumption:

$$L(f)\leq\widetilde{L}(f)+B_\ell\delta_c.$$

Proof of lemma 3. By rescaling the intermediate loss function to the interval $[0,1]$, we can directly apply the PAC-Bayes-kl bound.

Proof of lemma 4. Let $\alpha\in(0,1)$. We apply the following Hoeffding's inequality for an augmented sample x:

$$\mathbb{P}\left(S_{m-1}(x)-\mu(x)\leq-\varepsilon\mid x\right)\leq\delta_c.$$

For an augmented sample x, we can upper-bound $\tilde{\ell}(x,x^+)$ by the following term:

$$-\log\left(\frac{\mathrm{sim}^+(x)}{\mathrm{sim}^+(x)+\mu(x)+\varepsilon}\right)\mathbb{I}_{\{S_{m-1}(x)+\varepsilon\leq\mu(x)\}}-\log\left(\frac{\mathrm{sim}^+(x)}{\mathrm{sim}^+(x)+S_{m-1}(x)+2\varepsilon}\right)\mathbb{I}_{\{S_{m-1}(x)+\varepsilon>\mu(x)\}}.$$

Noting that $\mathbb{I}_{\{S_{m-1}(x)+\varepsilon>\mu(x)\}}\leq 1$ and that $\tilde{\ell}(x,x^+)\leq B$, we obtain

$$\tilde{\ell}(x,x^+)\leq\mathbb{I}_{\{S_{m-1}(x)+\varepsilon\leq\mu(x)\}}B+\ell'(x,x^+,X^-)$$

where ℓ' is the ε-modified loss:

$$\ell'(x, x^+, X^-) := -\log \frac{\mathrm{sim}^+(x)}{\mathrm{sim}^+(x) + S_{m-1}(x) + 2\varepsilon}.$$

Plugging this into the intermediate empirical loss, we have:

$$\widetilde{L}_S(Q) \leq B\frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{f\sim Q}\left[\mathbb{I}_{\{S_{m-1}(x_i)+\varepsilon \leq \mu(x_i)\}} \mid x_i, X_i^-\right] + L'_S(Q).$$

Using lemma 5, we have: with probability at least $1 - \delta_c^\alpha$ over dataset S:

$$\widetilde{L}_S(Q) \leq B\delta_c^{1-\alpha} + L'_S(Q).$$

Proof of lemma 5. We define the event $A(f, x, X^-) := \{S_{m-1}(x) - \mu(x) \leq -\varepsilon_\delta\}$. Assume

$$\mathbb{P}_{X^-}\left[A(f, x, X^-) \mid w, x\right] \leq \delta.$$

We need to show that

$$\mathbb{P}_S\left[\frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{f\sim Q}\left[\mathbb{I}_{\{A(f,x,X^-)\}} \mid x_i, X_i^-\right] > \delta^{1-\alpha}\right] \leq \delta^\alpha.$$

Using Markov's inequality, we derive:

$$\begin{aligned}
\mathbb{P}_S\left[\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathop{\mathbb{E}}_{f\sim Q}\left[\mathbb{I}_{\{A(f,x,X^-)\}} \mid x_i, X_i^-\right] > \delta^{1-\alpha}\right] &\leq \tfrac{1}{\delta^{1-\alpha}}\mathbb{E}_S\left[\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathop{\mathbb{E}}_{f\sim Q}\left[\mathbb{I}_{\{A(f,x,X^-)\}} \mid x_i, X_i^-\right]\right] \\
&\leq \tfrac{1}{\delta^{1-\alpha}}\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathbb{E}_{x_i, X_i^-}\left[\mathbb{E}_f\left[\mathbb{I}_{\{A(f,x_i,X_i^-)\}} \mid x_i, X_i^-\right]\right] \\
&= \tfrac{1}{\delta^{1-\alpha}}\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathbb{E}_{f,x_i,X_i^-}\left[\mathbb{I}_{\{A(f,x_i,X_i^-)\}}\right] \\
&= \tfrac{1}{\delta^{1-\alpha}}\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathbb{E}_{f,x_i}\left[\mathbb{E}_{X_i^-}\left[\mathbb{I}_{\{A(f,x_i,X_i^-)\}} \mid f, x_i\right]\right] \\
&\leq \tfrac{1}{\delta^{1-\alpha}}\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathbb{E}_{f,x_i}\left[\delta\right] \\
&\leq \delta^\alpha.
\end{aligned}$$

Proof of lemma 6. Let $\alpha \in (0, 1)$. Using lemma 2 and given that $B_\ell \leq B$, we have:

$$L(Q) \leq \widetilde{L}(Q) + B\delta_c.$$

Next, using lemma 3, we obtain: with probability at least $1 - \delta_{\text{pb}}$ over training i.i.d. samples $S \sim \mathcal{S}^n$, for all Q over $\mathcal{F}$, we have:

$$\frac{1}{B}\widetilde{L}(Q) \leq \mathrm{kl}^{-1}\left(\frac{1}{B}\widetilde{L}_S(Q), \frac{\mathrm{KL}\left(Q\|P\right) + \log\left(\frac{2\sqrt{n}}{\delta_{\text{pb}}}\right)}{n}\right) + \delta_c.$$

Using the union bound and lemma 4, we obtain: with probability at least $1 - \delta_{\text{pb}} - \delta_c^\alpha$ over dataset S, for all Q over $\mathcal{F}$,

$$\frac{1}{B}L(Q) \leq \mathrm{kl}^{-1}\left(\frac{1}{B}L'_S(Q) + \delta_c^{1-\alpha}, \frac{\mathrm{KL}\left(Q\|P\right) + \log\left(\frac{2\sqrt{n}}{\delta_{\text{pb}}}\right)}{n}\right) + \delta_c.$$

Let $\delta \in (0, 1)$ be a confidence parameter and $\alpha \in (0, 1)$. Setting $\delta_c^\alpha = \frac{\delta}{2}$ and $\delta_{\text{pb}} = \frac{\delta}{2}$ finishes the proof.

Proof of Theorem 2

Proof of Bounded Difference Assumption

The SimCLR empirical loss over the dataset S of size n is defined as:

$$\widehat{L}_S(f) = \frac{1}{n}\sum_{i=1}^{n} \ell_{cont}((x_i, x_i^+), X_i^-)$$

Definition 1 (bounded difference assumption) Let A be some set and $\phi : A^n \to R$. We say ϕ satisfies the bounded difference assumption if $\exists c_1, \ldots, c_n \geq 0$ s.t. $\forall i, 1 \leq i \leq n$

$$\sup_{x_1,\ldots,x_n,x_i'\in A} \left|\phi\left(x_1, \ldots, x_i, \ldots, x_n\right) - \phi\left(x_1, \ldots, x_i', \ldots, x_n\right)\right| \leq c_i$$

That is, if we subsitute x_i to x'_i, while keeping other x_j fixed, ϕ changes by at most c_i.

First, we can prove that the SimCLR empirical loss satisfies the bounded difference assumption with $c_i = \frac{C}{n}$. Indeed, we can prove the following lemma:

Lemma 1. We define $L_S(f) = \phi\left(x_1, \ldots, x_i, \ldots, x_n\right)$ and $C = \frac{4}{\tau} + (m-1)\log\frac{(m-1)+e^{\frac{2}{\tau}}}{m}$, we have

$$\sup_{x_1,\ldots,x_n,x'_i \in X} \left|\phi\left(x_1, \ldots, x_i, \ldots, x_n\right) - \phi\left(x_1, \ldots, x'_i, \ldots, x_n\right)\right| \le \frac{C}{n}.$$

Remark. For $m \in [50, 1000]$ and $\tau = 1$, C ranges from 9.89 to 10.36.

Proof. Let i be in $[1, n]$. We aim to bound the following quantity:

$$\Delta\phi(x_i) = \phi\left(x_1, \ldots, x_i, \ldots, x_n\right) - \phi\left(x_1, \ldots, x'_i, \ldots, x_n\right)$$

Due to symmetry, we can assume that $i \in [1, m]$, meaning we are studying a sample from the first batch. This quantity can be decomposed into 2 non-null terms:

$$\Delta\phi(x_i) = \frac{1}{n}\left[\delta_i(x_i) + \sum_{\substack{j=1\\ j\neq i}}^{m} \delta_j(x_i)\right]$$

where

$$\delta_i(x_i) := \ell_{cont}((x_i, x_i^+), X_i^-) - \ell_{cont}((x'_i, x_i'^+), X_i^-)$$

and

$$\delta_j(x_i) := \ell_{cont}((x_j, x_j^+), X_j^-) - \ell_{cont}((x_j, x_j^+), \widetilde{X}_j^-)$$

with $\widetilde{X}_j^-$ being the set of negative samples perturbed with x'_i. Using $\mathrm{sim}^+(x) := \exp\left(\frac{f(x)^\top f(x^+)}{\tau}\right)$ and $S_{m-1}(x) := \sum_{x' \in X^-} \exp\left(\frac{f(x)^\top f(x')}{\tau}\right)$, we can rewrite the first term as

$$\delta_i(x_i) = \log\frac{\mathrm{sim}^+(x'_i)}{\mathrm{sim}^+(x_i)} + \log\frac{\mathrm{sim}^+(x_i) + S_{m-1}(x_i)}{\mathrm{sim}^+(x'_i) + S_{m-1}(x'_i)}.$$

We can show that $\delta_i(x_i) \le \frac{4}{\tau}$ due to the fact that $\forall x, y$,

$$-\frac{1}{\tau} \le f(x)^\top f(y) \le \frac{1}{\tau}.$$

We can rewrite the second term as

$$\delta_j(x_i) = \log\frac{K + \mathrm{sim}(x_j, x_i)}{K + \mathrm{sim}(x_j, x'_i)}$$

where $(m-1)e^{-\frac{1}{\tau}} \le K \le (m-1)e^{\frac{1}{\tau}}$. We can show that $\delta_j(x_i) \le \log\frac{(m-1)e^{-\frac{1}{\tau}} + e^{\frac{1}{\tau}}}{(m-1)e^{-\frac{1}{\tau}} + e^{-\frac{1}{\tau}}} = \log\frac{(m-1)+e^{\frac{2}{\tau}}}{m}$. Combining the previous results, we obtain

$$\Delta\phi(x_i) \le \frac{1}{n}\left(\frac{4}{\tau} + (m-1)\log\frac{(m-1)+e^{\frac{2}{\tau}}}{m}\right).$$

Setting $C = \frac{4}{\tau} + (m-1)\log\frac{(m-1)+e^{\frac{2}{\tau}}}{m}$, we have proved that

$$\sup_{x_1,\ldots,x_n,x'_i \in X} \left|\Delta\phi(x_i)\right| \le \frac{C}{n}.$$

Proof of McAllester's non-i.i.d. Bound

We can prove the following PAC-Bayesian bound tailored to the SimCLR loss.

Theorem 2. With probability at least $1 - \delta$ over an i.i.d. sample S,

$$\forall Q, \quad L(Q) \le L_S(Q) + C\sqrt{\frac{\mathrm{KL}(Q \parallel P) + \log\frac{2n}{\delta}}{2(n-1)}}.$$

Proof. We adapt the proofs from McAllester's ”Simplified PAC-Bayesian Margin Bounds” and ”PAC-Bayesian Stochastic Model Selection” by incorporating the bounded difference assumption (McAllester 2003b,a). McAllester's proofs rely on the following two lemmas.

Lemma 1. If for $x > 0$,

$$P(|X| \geq x) \leq 2e^{-nx^2},$$

then

$$\mathbb{E}\left[e^{(n-1)X^2}\right] \leq 2n.$$

Lemma 2. Let h be a non-negative and convex function. If for a fixed $f \sim P$,

$$\mathop{\mathbb{E}}_{S\sim\mathcal{S}^n}\left[e^{(n-1)h(L_S(f)-L(f))}\right] \leq 2n,$$

then with probability at least $1-\delta$ over i.i.d. dataset S:

$$\forall Q, \quad h(L_S(Q) - L(Q)) \leq \frac{\mathrm{KL}(Q\|P) + \log\frac{2n}{\delta}}{n-1}.$$

From McDiarmid's inequality, for $\varepsilon > 0$,

$$\mathrm{P}\left(|L(f) - L_S(f)| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right),$$

we get for $X = L_S(f) - L(f)$,

$$\mathrm{P}\left(|X| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2n\varepsilon^2}{C^2}\right).$$

Rewriting this for $x > 0$,

$$\mathrm{P}_S\left(\frac{\sqrt{2}}{C}|X| \geq x\right) \leq 2\exp\left(-2x^2\right).$$

Using Lemma 1, we derive:

$$\mathbb{E}_S\left[e^{(n-1)h(X)}\right] \leq 2n,$$

where $h : x \mapsto \frac{2x^2}{C^2}$. This implies with probability at least $1-\delta$ over i.i.d. dataset S:

$$\forall Q, \quad h(L_S(Q) - L(Q)) \leq \frac{\mathrm{KL}(Q\|P) + \log\frac{2n}{\delta}}{n-1},$$

which can be rewritten as

$$\forall Q, \quad L(Q) \leq L_S(Q) + C\sqrt{\frac{\mathrm{KL}(Q\|P) + \log\frac{2n}{\delta}}{2(n-1)}}.$$

Proof of Lemma 1. The proof is from ”PAC-Bayesian Stochastic Model Selection” (McAllester 2003b). Assume for $x > 0$,

$$P(|X| \geq x) \leq 2e^{-nx^2}.$$

Then the continuous density that maximizes $\mathbb{E}\left[e^{(n-1)|X|^2}\right]$ and satisfies the previous inequality is such that

$$\int_x^\infty f_{|X|}(u)\mathrm{d}u = 2e^{-nx^2},$$

which gives $f_{|X|}(u) = 4nue^{-nu^2}$. We derive

$$\mathbb{E}\left[e^{(n-1)|X|^2}\right] \leq \int_0^\infty e^{(n-1)u^2} f_{|X|}(u)\mathrm{d}u,$$

which gives

$$\mathbb{E}\left[e^{(n-1)|X|^2}\right] \leq 2n \int_0^\infty 2ue^{-u^2} \mathrm{d}u = 2n.$$

Proof of Lemma 2. The proof is from "Simplified PAC-Bayesian Margin Bounds" (McAllester 2003a). Let h be a non-negative and convex function and P be any prior distribution. Assume for a fixed $f \sim P$,

$$\mathop{\mathbb{E}}_{S \sim \mathcal{S}^n}\left[e^{(n-1)h(L_S(f)-L(f))}\right] \leq 2n.$$

This implies

$$\mathbb{E}_S\left[\mathbb{E}_{f \sim P}\left[e^{(n-1)h(L_S(f)-L(f))}\right]\right] \leq 2n.$$

Applying Markov's inequality, we obtain

$$\mathrm{P}_S\left[\mathbb{E}_{f \sim P}\left[e^{(n-1)h(L_S(f)-L(f))}\right] \leq \frac{2n}{\delta}\right] \geq 1-\delta.$$

Next, we use a shift of measure:

$$\mathbb{E}_{f \sim Q}\left[(n-1)h\left(L_S(f)-L(f)\right)\right] \leq \mathrm{KL}(Q\|P) + \log \mathbb{E}_{f \sim P}\left[e^{(n-1)h(L_S(f)-L(f))}\right].$$

Combining the previous results, with probability at least $1-\delta$ over i.i.d. dataset S:

$$\mathbb{E}_{f \sim Q}\left[(n-1)h\left(L_S(f)-L(f)\right)\right] \leq \mathrm{KL}(Q\|P) + \log \frac{2n}{\delta}.$$

Since h is convex, applying Jensen's inequality finishes the proof:

$$(n-1)h\left(L_S(Q)-L(Q)\right) \leq \mathrm{KL}(Q\|P) + \log \frac{2n}{\delta}.$$

Proof of Theorem 3

In this section, we prove Theorem 3. We begin by demonstrating the first part of the theorem:

Bound A. *For all $f : \mathcal{X} \to \mathbb{S}^{d-1}$, the following inequality holds:*

$$\min_{W \in \mathbb{R}^{C \times d}} L_{\mathrm{CE}}(f, W) \leq \frac{\sigma}{\tau} + L(f) + \log\left(\frac{C}{K}\cosh^2\left(\frac{1}{\tau}\right)\right),$$

where $\sigma = \mathbb{E}_{(x,y)}\left[\|f(x)-\mu_y\|_2\right]$ represents the intra-class feature deviation.

Proof sketch. Bound A is derived by directly combining the techniques from Bao et al. (Bao, Nagano, and Nozawa 2022) and Wang et al. (Wang et al. 2022)
Next, we establish the second part:

Bound B. *For all $f : \mathcal{X} \to \mathbb{S}^{d-1}$, the following inequality holds:*

$$\min_{W \in \mathbb{R}^{C \times d}} L_{\mathrm{CE}}(f, W) \leq \sigma + \tau L(f) + \Delta_2,$$

where $\Delta_2 = \log(C) + \min\left\{\log(\cosh^2(1)), \tau \log\left(\frac{C}{K}\cosh^2\left(\frac{1}{\tau}\right)\right)\right\}$.
Proof sketch. Bound B is obtained by refining Bound A and eliminating the temperature scaling in the log-sum-exp term, leveraging the property that the log-sum-exp function acts as a smooth maximum.

Theorem 3: Bound A

Lower-bound on the contrastive population loss. Given that $\exp\left(f(x)^\top f(x^+)\right) \geq 0$ for any (x, x^+), we can derive a lower-bound on the contrastive loss:

$$-\mathbb{E}_{S \sim \mathcal{S}^m}\left[\frac{1}{m}\sum_{i=1}^m \frac{f(x_i)^\top f(x_i{}^+)}{\tau}\right] + \mathbb{E}_{S \sim \mathcal{S}^m}\left[\frac{1}{m}\sum_{i=1}^m \log\left(\sum_{x' \in X_i^-} \exp\left(\frac{f(x_i)^\top f(x')}{\tau}\right)\right)\right] \leq L(f).$$

Upper Bound on the Similarity of a Positive Pair. Firstly, it is evident that

$$\mathbb{E}_{S\sim\mathcal{S}^m}\left[\frac{1}{m}\sum_{i=1}^{m}\frac{f(x_i)^\top f({x_i}^+)}{\tau}\right]=\mathbb{E}_{(x,x^+)\sim\mathcal{S}}\left[\frac{f(x)^\top f(x^+)}{\tau}\right].$$

Using the Cauchy-Schwarz inequality, we can show for any μ_y:

$$\left|f(x)^\top\frac{f(x^+)-\mu_y}{\|f(x^+)-\mu_y\|}\right|\leq 1,$$

which implies

$$\left|f(x)^\top\left(f(x^+)-\mu_y\right)\right|\leq\left\|f(x^+)-\mu_y\right\|.$$

Assuming y is the label of the positive pair (x,x^+), we have

$$\mathbb{E}_{(x,x^+)\sim\mathcal{S}}\left[f(x)^\top f(x^+)\right]=\mathbb{E}_{(x,x^+)\sim\mathcal{S}}\left[f(x)^\top\left(f(x^+)-\mu_y\right)+f(x)^\top\mu_y\right].$$

Thus, we obtain:

$$\mathbb{E}_{(x,x^+)\sim\mathcal{S}}\left[\frac{f(x)^\top f(x^+)}{\tau}\right]\leq\frac{\sigma}{\tau}+\mathbb{E}_{(x,y)}\left[\frac{f(x)^\top\mu_y}{\tau}\right],$$

where $\sigma:=\mathbb{E}_{(x,y)}\left[\|f(x)-\mu_y\|_2\right]$ denotes the intra-class feature deviation.

Lower Bound on the Log-sum-exp Term. Our goal is to establish a lower bound on the log-sum-exp term. Let $\mathrm{LSE}(\mathbf{z}):=\ln\left(\sum_j\exp\left(z_j\right)\right)$ and $\mathbf{z}:=\left\{\frac{f(x)^\top f(x')}{\tau}\right\}_{x'\in X^-}$. Firstly, we observe that

$$\mathop{\mathbb{E}}_{S\sim\mathcal{S}^m}\left[\frac{1}{m}\sum_{i=1}^{m}\log\left(\sum_{x'\in X_i^-}\exp\left(\frac{f(x)^\top f(x')}{\tau}\right)\right)\right]=\frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{S\sim\mathcal{S}^m}\left[\mathrm{LSE}(\mathbf{z})\right].$$

Thus, our objective reduces to bounding $\mathop{\mathbb{E}}_{S\sim\mathcal{S}^m}\left[\mathrm{LSE}(\mathbf{z})\right]$. This expectation can be expressed as:

$$\mathop{\mathbb{E}}_{S\sim\mathcal{S}^m}\left[\mathrm{LSE}(\mathbf{z})\right]=\mathbb{E}_x\mathbb{E}_{X^-}\left[\mathrm{LSE}(\mathbf{z})\right].$$

Following (Bao, Nagano, and Nozawa 2022), our objective is to establish a lower bound on

$$\mathbb{E}_{X^-}\left[\mathrm{LSE}(\mathbf{z})\right]=\mathbb{E}_{\{x',y'\}}\left[\mathrm{LSE}(\mathbf{z})\right],$$

where y' denotes the label of x'. We define

$$\mathbf{z}^\mu:=\left\{f(x)^\top\mu_{y'}\right\}_{y'}.$$

Using Jensen's inequality and the convexity of the LSE function, we derive

$$\mathbb{E}_{\{y'\}}\mathbb{E}_{\{x'\}}\left[\mathrm{LSE}(\mathbf{z})\mid y'\right]\geq\mathbb{E}_{\{y'\}}\left[\mathrm{LSE}\left(\mathbb{E}_{\{x'\}}\left[\mathbf{z}\mid y'\right]\right)\right]$$

and thus,

$$\mathbb{E}_{\{x',y'\}}\left[\mathrm{LSE}(\mathbf{z})\right]\geq\mathbb{E}_{\{y'\}}\left[\mathrm{LSE}(\mathbf{z}^\mu)\right].$$

Similar to (Bao, Nagano, and Nozawa 2022), we will employ the following lemma:

Lemma. *For* $\mathbf{z}\in[-L^2,L^2]^N$*,*

$$2\log N\leq\mathrm{LSE}(\mathbf{z})+\mathrm{LSE}(-\mathbf{z})\leq 2\log\left(N\cosh(L^2)\right).$$

We recall the number of negative samples for the simplified SimCLR loss $K:=m-1$ and for the SimCLR loss $K:=2(m-1)$, and we have

$$\mathbb{E}_{\{y'\}}\left[\mathrm{LSE}(\mathbf{z}^\mu)\right]\geq-\mathbb{E}_{\{y'\}}\left[\mathrm{LSE}(-\mathbf{z}^\mu)\right]+2\log(K).$$

Applying Jensen's inequality, we have

$$-\mathbb{E}_{\{y'\}}\left[\mathrm{LSE}(-\mathbf{z}^\mu)\right]\geq-\left[\log\sum_{\{y'\}}\mathbb{E}_{y'}\left[\exp(-f(x)^\top\mu_{y'})\right]\right].$$

In addition, we have, for a fixed x,

$$\mathbb{E}_{y'}\left[\exp(-f(x)^\top \mu_{y'})\right] = \sum_{c\in\mathcal{C}} \exp(-f(x)^\top \mu_c)\pi(c),$$

where $\mathcal{C}$ represents the set of classes and for convenience, we assume $\pi(c) = \frac{1}{C}$.

Combining all the previous steps, we obtain

$$\mathbb{E}_{\{x',y'\}}\left[\mathrm{LSE}(\mathbf{z})\right] \geq -\log \frac{K}{C}\sum_{c\in\mathcal{C}} \exp(-f(x)^\top \mu_c) + 2\log(K),$$

which implies

$$\mathbb{E}_{\{x',y'\}}\left[\mathrm{LSE}(\mathbf{z})\right] \geq -\,\mathrm{LSE}\left(\left\{-f(x)^\top \mu_c\right\}_{c\in\mathcal{C}}\right) + \log(K) + \log(C).$$

We can apply the lemma again and obtain:

$$\mathbb{E}_{\{x',y'\}}\left[\mathrm{LSE}(\mathbf{z})\right] \geq \mathrm{LSE}\left(\left\{f(x)^\top \mu_c\right\}_{c\in\mathcal{C}}\right) - 2\log(C\cosh(\frac{1}{\tau})) + \log(K) + \log(C).$$

All in all, we have proved that

$$\mathbb{E}_{X^-}\left[\mathrm{LSE}(\mathbf{z})\right] \geq \mathrm{LSE}\left(\left\{f(x)^\top \mu_c\right\}_{c\in\mathcal{C}}\right) - \Delta,$$

where $\Delta := \log\left(\frac{C}{K}\cosh^2(\frac{1}{\tau})\right)$.

Final Expression. We recall the definition of the Cross-Entropy loss of a linear classifier:

$$L_{\mathrm{CE}}(f,W) = \mathbb{E}_{(x,y)}\left[-\log \frac{\exp\left(f(x)^\top w_y\right)}{\sum_{i=1}^{C}\exp\left(f(x)^\top w_i\right)}\right],$$

Combining the previous steps, we obtain

$$L(f) \geq -\frac{\sigma}{\tau} - \mathbb{E}_{(x,c)}\left[f(x)^\top\left(\frac{\mu_c}{\tau}\right)\right] + \mathbb{E}_{(x,c)}\,\mathrm{LSE}\left(\left\{\exp\left(f(x)^\top\left(\frac{\mu_c}{\tau}\right)\right)\right\}_{c\in[C]}\right) - \Delta.$$

Rearranging the terms and noting that $W^{\mu} := [\frac{\mu_1}{\tau}\cdots\frac{\mu_C}{\tau}]^\top$ is a linear classifier completes the proof:

$$\min_{W\in\mathbb{R}^{C\times d}} L_{\mathrm{CE}}(f,W) \leq L(f) + \frac{\sigma}{\tau} + \Delta.$$

Theorem 3: Bound B

To prove the second part of the theorem, we extend our approach from part 1 but focus on the complex dependency on temperature in the log-sum-exp term using the following lemma:

Lemma: Let $\mathbf{x}\in\mathbb{R}^n$ with $\mathbf{x} = (x_1,\ldots,x_n)$. We have

$$\max_i x_i \leq \mathrm{LSE}(\mathbf{x}) \leq \max_i x_i + \log(n)$$

and for any $t > 0$,

$$t\max_i x_i < \mathrm{LSE}(t\mathbf{x}) \leq t\max_i x_i + \log(n).$$

The LSE function acts as a smooth maximum—a smooth approximation to the maximum function. We can leverage this lemma to establish lower bounds on the log-sum-exp term in two different ways.

First Approach. Given $\tau > 0$, it follows that

$$\mathrm{LSE}(\mathbf{z}/\tau) > \frac{1}{\tau}\max_i z_i \geq \frac{1}{\tau}\left(\mathrm{LSE}(\mathbf{z}) - \log(K)\right).$$

Following the proof of part one, we derive:

$$\mathbb{E}_{X^-}\left[\mathrm{LSE}(\mathbf{z}/\tau)\right] \geq \frac{1}{\tau}\left(\mathrm{LSE}\left(\left\{f(x)^\top \mu_c\right\}_{c\in[C]}\right) - \Delta - \log(K)\right).$$

Combining the previous steps, we obtain

$$\tau L(f) \geq -\sigma - \mathbb{E}_{(x,c)}\left[f(x)^\top \mu_c\right] + \mathbb{E}_{(x,c)} \operatorname{LSE}\left(\left\{\exp(f(x)^\top \mu_c)\right\}_{c\in[C]}\right) - \Delta - \log(K)$$

which completes the proof:

$$L_{\mathrm{CE}}^{\mu}(f) \leq \tau L(f) + \sigma + \Delta'.$$

where $\Delta' := \log(C\cosh^2(1))$ and the loss of the mean classifier is defined as

$$L_{\mathrm{CE}}^{\mu}(f) = \mathbb{E}_{(x,y)}\left[-\log\frac{\exp\left(f(x)^\top \mu_y\right)}{\sum_{i=1}^{C}\exp\left(f(x)^\top \mu_i\right)}\right].$$

Since the mean classifier is a linear classifier, we obtain a bound on the optimal linear classifier:

$$\inf_{\mathbf{W}\in\mathbb{R}^{C\times d}} L_{\mathrm{CE}}(f, W) \leq \tau L(f) + \sigma + \Delta'$$

Second Approach. Similarly to the proof of part one, we establish:

$$\mathbb{E}_{X^-}\left[\operatorname{LSE}(\mathbf{z}/\tau)\right] \geq \operatorname{LSE}\left(\left\{f(x)^\top \mu_c/\tau\right\}_{c\in[C]}\right) - \Delta$$

where $\Delta := \log(\frac{C}{K}\cosh^2(\frac{1}{\tau}))$.

Applying the lemma to $\operatorname{LSE}\left(\left\{f(x)^\top \mu_c/\tau\right\}_{c\in[C]}\right)$yields:

$$\operatorname{LSE}\left(\left\{f(x)^\top \mu_c/\tau\right\}_{c\in[C]}\right) \geq \frac{1}{\tau}\left(\operatorname{LSE}\left(\left\{f(x)^\top \mu_c\right\}_{c\in[C]}\right) - \log(C)\right).$$

Combining these steps, we derive:

$$\mathbb{E}_{X^-}\left[\operatorname{LSE}(\mathbf{z}/\tau)\right] \geq \frac{1}{\tau}\operatorname{LSE}\left(\left\{f(x)^\top \mu_c\right\}_{c\in[C]}\right) - \frac{\log(C) - \tau\Delta}{\tau}.$$

This results in a similar expression as in the first approach:

$$\tau L(f) \geq -\sigma - \mathbb{E}_{(x,c)}\left[f(x)^\top \mu_c\right] + \mathbb{E}_{(x,c)} \operatorname{LSE}\left(\left\{\exp(f(x)^\top \mu_c)\right\}_{c\in[C]}\right) - \tau\Delta - \log(C).$$

Thus, we conclude the proof:

$$\inf_{\mathbf{W}\in\mathbb{R}^{C\times d}} L_{\mathrm{CE}}(f, W) \leq \tau L(f) + \sigma + \Delta'.$$

where $\Delta' := \tau\log(\frac{C}{K}\cosh^2(\frac{1}{\tau})) + \log(C)$.

Proof Theorem 4

We follow the steps of the proof of theorem 1 and adapt it to the zero-one contrastive risk.

Concentration bound on the negative samples. For convenience, we have$|X^-| = m - 1$. We define $S_{m-1}(x) := \sum_{x'\in X^-}\mathbb{I}_{\{f(x)^\top f(x^+) < f(x)^\top f(x')\}}$ and similarly to the proof of Theorem 1, we apply Hoeffding's inequality: for all $\varepsilon > 0$,

$$\mathbb{P}\left(S_{m-1}(x) - \mathbb{E}[S_{m-1}(x)] \geq \varepsilon \mid x\right) \leq \exp\left(-\frac{2\varepsilon^2}{m-1}\right) = \delta_c,$$

where in this case $c = 1$. Solving for ε:

$$\varepsilon = \sqrt{\frac{(m-1)\log\left(\frac{1}{\delta_c}\right)}{2}}.$$

Upper Bound on the contrastive zero-one risk by an Intermediate Loss. Similarly, we can show

$$R(Q) \leq \widetilde{R}(Q) + \frac{\varepsilon}{m-1} + \delta_c,$$

where $\widetilde{R}(Q)$ represents the intermediate population risk where the inner loss was replaced by $r((x,x^+),X^-) = \mathbb{E}_{x'}\left[\mathbb{I}_{\{f(x)^\top f(x^+)<f(x)^\top f(x')\}} \mid x\right]$.

PAC-Bayes-kl Bound on the Intermediate Loss. We can apply the PAC-Bayes-kl bound: given a prior $\mathcal{P}$ over $\mathcal{F}$ and $\delta_{\text{pb}} \in (0,1)$, with probability at least $1-\delta_{\text{pb}}$ over training i.i.d. samples $S \sim \mathcal{S}^n$, for all $\mathcal{Q}$ over $\mathcal{F}$, we have:

$$\widetilde{R}(Q) \leq \mathrm{kl}^{-1}\left(\widetilde{R}_S(Q), \frac{\mathrm{KL}\left(Q\|P\right)+\log\left(\frac{2\sqrt{n}}{\delta_{\text{pb}}}\right)}{n}\right).$$

Upper-bound on the Intermediate Loss. We can show that the intermediate empirical risk can be upper-bounded by an expression containg the contrastive zero-one empirical risk with probability at least $1-\delta_c^\alpha$ over dataset S:

$$\widetilde{R}_S(Q) \leq \delta_c^{1-\alpha} + \widehat{R}_S(Q) + \frac{\varepsilon}{m-1}.$$

Final Expression of the PAC-Bayes Bound on the contrastive zero-one risk. Using the union bound, we inject the previous upper-bound into the PAC-Bayes bound to finish the proof: with probability at least $1-\delta_{\text{pb}}-\delta_c^\alpha$ over dataset S,

$$R(Q) \leq \mathrm{kl}^{-1}\left(\widehat{R}_S(Q) + \frac{\varepsilon}{m-1} + \delta_c^{1-\alpha}, \frac{\mathrm{KL}\left(Q\|P\right)+\log\left(\frac{2\sqrt{n}}{\delta_{\text{pb}}}\right)}{n}\right) + \frac{\varepsilon}{m-1} + \delta_c.$$

Let $\delta \in (0,1)$ be a confidence parameter and $\alpha \in (0,1)$. Setting $\delta_c^\alpha = \frac{\delta}{2}$,$\delta_{\text{pb}} = \frac{\delta}{2}$, and $\gamma = \frac{\varepsilon}{m-1}$the bound can be rewritten as: with probability at least $1-\delta$ over dataset S:

$$R(Q) \leq \mathrm{kl}^{-1}\left(\widehat{R}_S(Q) + \gamma + \left(\frac{\delta}{2}\right)^{\frac{1-\alpha}{\alpha}}, \frac{\mathrm{KL}\left(Q\|P\right)+\log\left(\frac{\sqrt{n}}{\delta}\right)}{n}\right) + \gamma + \left(\frac{\delta}{2}\right)^{\frac{1}{\alpha}}.$$

with $\gamma = \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2(m-1)\alpha}}$.

Proof Theorem 5

We follow the steps of the proof of theorem 2 and adapt it to the zero-one contrastive risk.

Bounded difference assumption. We can show that the contrastive zero-one risk satisfies the bounded difference assumption with $c_i = \frac{2}{n}, \forall i$. Similarly, we can show this by bounding the following quantity

$$\Delta\phi(x_i) = \frac{1}{n}\left[\delta_i(x_i) + \sum_{\substack{j=1\\ j\neq i}}^{m} \delta_j(x_i)\right]$$

where

$$\delta_i(x_i) := r((x_i,x_i^+),X_i^-) - r((x_i',x_i'^+),X_i^-)$$

and

$$\delta_j(x_i) := r((x_j,x_j^+),X_j^-) - r((x_j,x_j^+),\widetilde{X}_j^-)$$

with $\widetilde{X}_j^-$ being the set of negative samples perturbed with x_i'.

Since $\delta_i(x_i) \leq 1$ and $\delta_j(x_i) \leq \frac{1}{m-1}$, we observe that

$$\sup_{x_1,\ldots,x_n,x_i'\in X} |\Delta\phi(x_i)| \leq \frac{2}{n}.$$

Proof of theorem 5. We can directly apply the previous result with $C=2$ and obtain: with probability at least $1-\delta$ over an i.i.d. sample S,

$$\forall Q, \quad R(Q) \leq \widehat{R}_S(Q) + 2\sqrt{\frac{\mathrm{KL}(Q \parallel P) + \log\frac{2n}{\delta}}{2(n-1)}}.$$

Additional Experimental Details

Data Pre-processing Details

For data augmentation, we apply random cropping (`transforms.RandomResizedCrop`), random horizontal flip with probability 0.5 (`transforms.RandomHorizontalFlip`), color jittering with strength 0.5 and probability 0.8 (`transforms.ColorJitter`), and color dropping (`transforms.RandomGrayscale`), leaving out gaussian blur (Chen et al. 2020; Marcel and Rodriguez 2010). We then normalize the augmented images per channel using the mean and standard deviation of the training data. The code for data augmentation using Pytorch is as follows (Paszke et al. 2017) :

```
def create_simclr_data_augmentation(strength=0.5, name="mnist"):
"""
    Create a data augmentation pipeline for SimCLR.

    Args:
        strength (float): Intensity of the augmentation.
        name (str): Name of the dataset.

    Returns:
        transforms.Compose
"""
    if name == "mnist":
        size = 28
        mean = (0.1307,)
        std_dev = (0.3081,)
    if name == "cifar10":
        size = 32
        mean = (0.4914, 0.4822, 0.4465)
        std_dev = (0.2470, 0.2435, 0.2616)
    scale = (0.08, 1.0)
    color_jitter = transforms.ColorJitter(
        brightness=0.8 * strength,
        contrast=0.8 * strength,
        saturation=0.8 * strength,
        hue=0.2 * strength,
    )

    common_transforms = [
        transforms.RandomResizedCrop(size=size, scale=scale),
        transforms.RandomHorizontalFlip(p=0.5),
        transforms.RandomApply(transforms=[color_jitter], p=0.8),
        transforms.RandomGrayscale(0.2),
        transforms.ToTensor(),
        transforms.Normalize(mean, std_dev)
    ]

    return transforms.Compose(common_transforms)
```

Computing Infrastructure

The experiments are run using three different resource types:

- CPU + Nvidia Tesla P100 (16GB, No Tensor Cores)
- CPU + Nvidia Tesla V100 (16GB, Tensor Cores)
- CPU + Nvidia Ampere A100 (20G MIG, Ampere Tensor Cores)

We use a small memory (40GB) and the Nvidia NGC container image *Pytorch 2.4.0*.

Experiments on MNIST

		SimCLR Loss			
		$\tau = 1$	$\tau = 0.7$	$\tau = 0.5$	$\tau = 0.2$
	Test Loss	4.9318	4.6575	4.3130	2.8574
Risk Certificate	kl bound (iid)	6.97	7.008	7.258	8.108
	Catoni's bound (iid)	6.875	6.848	7.007	7.618
	Classic bound (iid)	7.904	7.426	7.475	8.324
	Nozawa et al.	23.469	20.448	22.275	35.05
	Th. 1 (ours)	5.465	5.383	5.368	6.011
	Th. 2 (ours)	5.099	5.093	5.729	36.033
	KL $/n$	0.0009	0.0005	0.0005	0.0006

Table 5: Comparison of risk certificates for the SimCLR loss on the MNIST dataset.

		Contrastive 0-1 Risk			
		$\tau = 1$	$\tau = 0.7$	$\tau = 0.5$	$\tau = 0.2$
	Test Loss	0.0571	0.0529	0.0478	0.0357
Risk Certificate	kl bound (iid)	0.408	0.327	0.324	0.319
	Catoni's bound (iid)	0.419	0.333	0.331	0.309
	Classic bound (iid)	0.466	0.396	0.394	0.399
	Nozawa et al.	2.535	1.95	2.097	2.097
	Th. 4 (ours)	0.356	0.347	0.343	0.333
	Th. 5 (ours)	0.113	0.101	0.098	0.088

Table 6: Comparison of risk certificates for the contrastive zero-one risk on the MNIST dataset.

KL Divergence

The KL divergence between one-dimensional Gaussian distributions is given by:

$$\mathrm{KL}\left(\mathrm{Gauss}\left(\mu_1, b_1\right) \| \mathrm{Gauss}\left(\mu_0, b_0\right)\right) = \frac{1}{2}\left(\log\left(\frac{b_0}{b_1}\right) + \frac{\left(\mu_1 - \mu_0\right)^2}{b_0} + \frac{b_1}{b_0} - 1\right).$$

For multi-dimensional Gaussian distributions with diagonal covariance matrices, the KL divergence is the sum of the KL divergences of the independent components.

Extension of Theorem 1 and 4 to Catoni's Bound

Theorem 1 using Catoni's bound:

$$\frac{1}{B}L(Q) \leq \inf_{\substack{\alpha\in(0,1)\\ \lambda>0}} \left\{ \frac{1 - \exp\left\{-\lambda\left(\frac{L'_S(Q)}{B} + \left(\frac{\delta}{2}\right)^{\frac{1-\alpha}{\alpha}}\right) - \frac{\mathrm{KL}(Q\|P)+\log\left(\frac{1}{\delta}\right)}{n}\right\}}{1 - \exp(-\lambda)} + \left(\frac{\delta}{2}\right)^{\frac{1}{\alpha}} \right\}$$

Theorem 4 using Catoni's bound:

$$R(Q) \leq \inf_{\substack{\alpha\in(0,1)\\ \lambda>0}} \left\{ \frac{1 - \exp\left\{-\lambda\left(R_S(Q) + \gamma + \left(\frac{\delta}{2}\right)^{\frac{1-\alpha}{\alpha}}\right) - \frac{\mathrm{KL}(Q\|P)+\log\left(\frac{1}{\delta}\right)}{n}\right\}}{1 - \exp(-\lambda)} + \gamma + \left(\frac{\delta}{2}\right)^{\frac{1}{\alpha}} \right\},$$

with $\gamma = \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2(m-1)\alpha}}$.

References

Alquier, P.; and Guedj, B. 2018. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107(5): 887–902.

		$\tau = 1$	$\tau = 0.7$	$\tau = 0.5$	$\tau = 0.2$
	Bao et al.	3.0212	3.7607	5.0141	12.5462
	Th. 3 (ours)	3.0212	3.7607	4.8096	4.5996
Proj.	Sup. Loss	1.5453	1.5317	1.5063	1.5163
	top-1	0.8874	0.9057	0.9425	0.9418
	Sup. Loss	1.4783	1.4769	1.4732	1.4793
	top-1	0.9779	0.9809	0.9829	0.9778

Table 7: Comparison of upper bounds on downstream classification loss with MNIST.

Arora, S.; Khandeparkar, H.; Khodak, M.; Plevrakis, O.; and Saunshi, N. 2019. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*.

Bao, H.; Nagano, Y.; and Nozawa, K. 2022. On the surrogate gap between contrastive and supervised losses. In *International Conference on Machine Learning*, 1585–1606. PMLR.

Catoni, O. 2007. PAC-Bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, W.; Liu, T.-Y.; Lan, Y.; Ma, Z.-M.; and Li, H. 2009. Ranking measures and loss functions in learning to rank. *Advances in Neural Information Processing Systems*, 22.

Chérief-Abdellatif, B.-E.; Shi, Y.; Doucet, A.; and Guedj, B. 2022. On PAC-Bayesian reconstruction guarantees for VAEs. In *International conference on artificial intelligence and statistics*, 3066–3079. PMLR.

Dziugaite, G. K.; and Roy, D. M. 2017. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*.

Germain, P.; Lacasse, A.; Laviolette, F.; and Marchand, M. 2009. PAC-Bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 353–360.

HaoChen, J. Z.; Wei, C.; Gaidon, A.; and Ma, T. 2021. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34: 5000–5011.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Hoeffding, W. 1994. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, 409–426.

Jing, L.; Vincent, P.; LeCun, Y.; and Tian, Y. 2021. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*.

Krizhevsky, A.; and Hinton, G. 2009. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto.

Le-Khac, P. H.; Healy, G.; and Smeaton, A. F. 2020. Contrastive representation learning: A framework and review. *Ieee Access*, 8: 193907–193934.

LeCun, Y.; Cortes, C.; and Burges, C. J. C. 2010. MNIST handwritten digit database.

Marcel, S.; and Rodriguez, Y. 2010. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM international conference on Multimedia*, 1485–1488.

McAllester, D. 2003a. Simplified PAC-Bayesian margin bounds. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, 203–215. Springer.

McAllester, D. A. 2003b. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1): 5–21.

McDiarmid, C.; et al. 1989. On the method of bounded differences. *Surveys in combinatorics*, 141(1): 148–188.

Nozawa, K.; Germain, P.; and Guedj, B. 2020. PAC-Bayesian contrastive unsupervised representation learning. In *Conference on Uncertainty in Artificial Intelligence*, 21–30. PMLR.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in PyTorch.

Perez-Ortiz, M.; Rivasplata, O.; Guedj, B.; Gleeson, M.; Zhang, J.; Shawe-Taylor, J.; Bober, M.; and Kittler, J. 2021a. Learning PAC-Bayes priors for probabilistic neural networks. *arXiv preprint arXiv:2109.10304*.

Perez-Ortiz, M.; Rivasplata, O.; Shawe-Taylor, J.; and Szepesvári, C. 2021b. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22(227): 1–40.

Seeger, M. 2002. PAC-Bayesian generalisation error bounds for Gaussian process classification. *Journal of machine learning research*, 3(Oct): 233–269.

Sohn, K. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.

Wang, T.; and Isola, P. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, 9929–9939. PMLR.

Wang, Y.; Zhang, Q.; Wang, Y.; Yang, J.; and Lin, Z. 2022. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. *arXiv preprint arXiv:2203.13457*.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3733–3742.

Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; and Deny, S. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, 12310–12320. PMLR.

Zhou, W.; Veitch, V.; Austern, M.; Adams, R. P.; and Orbanz, P. 2018. Non-vacuous generalization bounds at the imagenet scale: a PAC-bayesian compression approach. *arXiv preprint arXiv:1804.05862*.