PAC-Bayesian Analysis of the Surrogate Relation between Joint Embedding and Supervised Downstream Losses

Theresa Wasserer THERESA.WASSERER@TUM.DE
Technical University of Munich

Maximilian Fleissner MAXIMILIAN.FLEISSNER@TUM.DE
Technical University of Munich

Debarghya Ghoshdastidar GHOSHDAS@CIT.TUM.DE
Technical University of Munich

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

In recent years, self-supervised representation learning (SSL) has become an important learning paradigm and a crucial component of foundation models. SSL-based training pipelines are typically formalized as a sequence of two tasks—a pretext task that learns representations from large amounts of augmented unlabeled data, and a downstream task, where a simple model is fit on the learned representations with the help of little labeled data. The strong empirical performance of SSL-based pipelines for prominent joint embedding loss functions is not yet well explained in theory due to two main reasons: a lack of non-vacuous generalization bounds for the models learned in the pretext task, and a lack of practically computable transfer bounds that describe how generalization bounds derived for the pretext task transfer to the downstream task. In this work, we first derive non-vacuous PAC-Bayesian generalization bounds for models optimized in the pretext task with prominent joint embedding SSL loss functions (VICReg, Barlow Twins, and Spectral Contrastive loss), accounting for their non-i.i.d. nature. Next, we provide the first practically computable transfer bounds for our considered loss functions by formally proving a surrogate relation that upper bounds the downstream squared L2 loss by the SSL pretext loss and a more accurate measure for the influence of the chosen augmentations than in previous work. In addition, our theoretical analysis identifies effective hyperparameter choices, thereby reducing the need for extensive hyperparameter tuning and offering principled guidance for model selection. We empirically validate our theoretical findings on CIFAR-10 and MNIST datasets.

Keywords: Self-supervised learning, representation learning, generalization, PAC-Bayes bounds

1. Introduction

Supervised learning is a prominent learning paradigm with a well-studied statistical framework. However, training state-of-the-art deep neural networks in a supervised way requires large amounts of labeled data, which are scarce and costly as labels are typically human-generated. Self-supervised representation learning (SSL) addresses this challenge by leveraging vast amounts of freely available unlabeled data. As a result of adopting SSL, the overall training pipeline is divided into two stages. First, a pretext task is solved using SSL, in which pseudo-labels are automatically generated from unlabeled data with the goal of learning useful feature representations. This is achieved by learning a representation function $f : \mathcal{X} \to \mathbb{R}^d$ that maps inputs from a domain $\mathcal{X}$ to d-dimensional feature vectors. Second, these learned feature representations are used to solve a supervised downstream task—the actual target task—which is expected to become simpler when leveraging the learned representations, thereby reducing the required amount of labeled data. Consequently, SSL has become the foundation of numerous popular learning frameworks (Wu et al., 2018; Oord et al., 2018; Caron et al., 2020; Tian et al., 2020; Henaff, 2020; Misra and Maaten, 2020). Nonetheless, the theoretical understanding of statistical properties in this two-stage training pipeline remains limited.

Of ultimate interest are generalization bounds that quantify how the model's performance on the training data generalizes to unseen data. To prove success of SSL-based two-stage training pipelines, we require a good generalization bound on the downstream task, which due to the split-up pipeline is composed of two different bounds: the pretext generalization bound, describing the performance of the model trained in the pretext task, and the transfer bound, describing how the performance on the pretext task transfers to the downstream task. However, for state-of-the-art joint embedding loss functions there is a lack of tight and practically computable generalization bounds for downstream tasks, which can be attributed to two main reasons. First, most existing pretext generalization bounds in SSL are vacuous, as they are stated in terms of Rademacher complexity, a standard measure in statistical learning theory (Arora et al., 2019; HaoChen et al., 2021; Cabannes et al., 2023; Shwartz-Ziv et al., 2023). Recent techniques have alleviated this issue by enabling the derivation of non-vacuous PAC-Bayesian generalization bounds (Perez-Ortiz et al., 2021b; Nozawa et al., 2020; van Elst and Ghoshdastidar, 2025). Second, deriving practical transfer bounds requires showing that the pretext task acts as a surrogate for the downstream task. This surrogate relation is based on an assumed semantic correspondence between pretext and downstream tasks, which must then be rigorously proven for specific loss functions.

We now introduce the semantic correspondence for which we aim to establish a surrogate relation, focusing on a prominent pretext–downstream-task pair that has been successfully applied in the vision domain. Therein, the pretext task of contrastive instance discrimination empirically has been found to simplify the downstream task of classification, subsequently requiring only linear classifiers to separate feature representations according to their class labels (Chen et al., 2020; Zbontar et al., 2021; Bardes et al., 2022). This indicates that the pretext task already produces class-separable features, despite no explicit class labels being available. In contrastive instance discrimination, augmented views of images are generated, e.g., through rotations, color jitter or cropping, and joint embedding loss functions encourage that representations of same-instance augmentations become closely aligned in the feature space, while representations of two random-instance augmentations become distant from one another (Ericsson et al., 2022). To not only encourage the representation alignment of same-instance augmentations, but also of same-class instances—as class labels do in supervised tasks—a strong augmentation pipeline is required (HaoChen et al., 2021; Wang et al., 2022): Aggressive cropping often yields similar augmentations for same-class instances, e.g., extracting wheels on different car images, which encourages the representation alignment of same-class instances through the loss function's representation alignment of same-instance augmentations (Wang et al., 2022). Hence, the semantic correspondence described by the surrogate relation depends on two properties of the pretext task: first, the loss function encoding the instance discrimination objective; and second, the clustering strength of the augmentation pipeline, measuring the representation alignment of same-class instances.

Finally, establishing a surrogate relation, and hence deriving a transfer bound between contrastive instance discrimination and classification, consists in upper bounding the classification loss—incurred by an optimal linear classifier—by the contrastive instance discrimination loss and a suitable measure for the clustering strength of the augmentation pipeline. Ultimately, we are interested in deriving transfer bounds for state-of-the-art joint embedding loss functions, which requires

	Pretext loss function	Downstream loss function	PAC-Bayes pretext bound	Mean classifier approach	Data augmentation	Distinguishability of augmentation strengths
Arora et al. (2019)	Contrastive (Hinge, Logistic)	Hinge, Logistic	×	✓	×	×
Nozawa et al. (2020)	Contrastive (Hinge, Logistic, 0–1)	Hinge, Logistic	✓	✓	×	×
HaoChen et al. (2021)	SCL	SL	×	×	✓	×
Han Bao et al. (2022)	InfoNCE	CE	×	✓	×	×
Wang et al. (2022)	InfoNCE	CE	×	✓	✓	×
Shwartz-Ziv et al. (2023)	VICReg-style	L2 norm	×	×	×	×
van Elst and Ghoshdastidar (2025)	SimCLR	CE	✓	✓	✓	×
Ours	SCL, VICReg, BT	SL	✓	✓	✓	✓

Table 1: Comparison of downstream generalization bounds for classification, decomposed into pretext and transfer bounds. Reported are the considered loss functions, the presence of a PAC-Bayesian pretext bound, whether the transfer bound builds on the mean classifier approach, the transfer bound's handling of data augmentation, and its ability to capture differences in clustering strength across augmentation pipelines.

finding classification loss functions that are both effective in practice and provably relatable to these pretext loss functions. Therefore, the two loss functions must share a similar form. For classification the cross-entropy loss (CE) is typically used in practice and has been provably related to the pretext SimCLR loss function (Chen et al., 2020) due to their similar LogSumExp form, yielding a practical transfer bound (van Elst and Ghoshdastidar, 2025). However, SimCLR comes with computational drawbacks compared to state-of-the-art joint embedding loss functions such as VICReg (Bardes et al., 2022) or Barlow Twins (BT) (Zbontar et al., 2021). VICReg and BT lack the required LogSumExp form, but are similar to the Spectral Contrastive loss (SCL) function, which has been provably related to the downstream squared L2 loss (SL) function (HaoChen et al., 2021). However, the resulting transfer bound is impractical as it depends on non-computable quantities derived from a so-called population augmentation graph. In contrast, a proof technique, to which we refer as mean classifier approach, allows for relating convex loss functions of the same form, yielding computable transfer bounds (Arora et al., 2019). While this approach initially relied on impractical assumptions (Arora et al., 2019; Han Bao et al., 2022; Nozawa et al., 2020), it was recently adapted to account for the role of augmentations (van Elst and Ghoshdastidar, 2025). Nonetheless, current transfer bounds struggle to properly capture the impact that augmentation pipelines of different clustering strength, e.g., with or without aggressive cropping, have on the classification performance (Wang et al., 2022; van Elst and Ghoshdastidar, 2025). We summarize the limitations of previous bounds in Table 1.

Contribution. We, for the first time, are able to theoretically explain the success of state-of-the-art joint embedding losses for the downstream task of classification via practically computable downstream generalization bounds that adequately capture the performance transfer between tasks. These bounds offer theoretical guidance on selecting effective hyperparameter ranges for pretext losses and provide insight into principled SSL loss function design, as further discussed in Section 5. We develop our results through the following steps:

- We present a *unifying perspective* on theory-friendly variants of the pretext losses VICReg, BT and SCL by leveraging similarities among them (Balestriero and LeCun, 2022; Garrido et al., 2023). This allows us to provide unified proofs throughout the paper.
- We relate these pretext losses to a *new downstream loss*, which we introduce as supervised joint embedding loss (SupJE). We show that this downstream loss (a) performs on par with the standard CE loss, (b) recovers the squared L2 loss as a special case, and (c) provides

insights into effective pretext hyperparameter combinations, which reduces the need for costly hyperparameter tuning.

- We derive practically computable *transfer bounds* for these losses, building on the mean classifier approach (Arora et al., 2019). Moreover, we refine existing bounds on the *augmentation strength*, which allows us to capture the performance transfer more precisely than previous work (Wang et al., 2022; van Elst and Ghoshdastidar, 2025).
- We leverage recent advances on PAC-Bayes bounds to provide *non-vacuous pretext generalization bounds*, accounting for their non-i.i.d. nature by building on techniques from van Elst and Ghoshdastidar (2025) and employing the PAC-Bayes with Backprop algorithm (Perez-Ortiz et al., 2021b).
- We empirically evaluate our resulting downstream bounds and show the effectiveness of special hyperparameter combinations through experiments on the CIFAR-10 and MNIST datasets.

2. Preliminaries

In the following, we first formalize the learning problem within the two-stage pipeline of representation learning and downstream classification. We then provide technical details on the pretext and downstream tasks, with a focus on the augmentation process and loss functions. Lastly, we present the PAC-Bayesian framework used to derive the pretext generalization bounds.

Formalization of the SSL-based Two-Stage Training Pipeline. Let $\mathcal{X}$ be the input domain with distribution $\mathcal{D}_{\mathcal{X}}$ and $X \sim \mathcal{D}_{\mathcal{X}}$ a random variable; $\mathcal{Y} = \{1, \ldots, C\}$ the output domain with C classes, Y a random variable taking values in $\mathcal{Y}$ with $\pi = [\mathbb{P}(Y = c)]_{c \in \mathcal{Y}}$ the prior distribution over class labels; $\mathcal{D}_c = \mathbb{P}\left(X|Y = c\right)$ the class-conditional distribution for each $c \in \mathcal{Y}$; and $\mathcal{D}_{\mathcal{X},\mathcal{Y}}$ the joint distribution on $\mathcal{X} \times \mathcal{Y}$. The overall goal is to learn a classifier $h : \mathcal{X} \to \mathcal{Y}$, predicting class labels for input data, by training neural networks. To this end, SSL models first learn a normalized representation function $f : \mathcal{X} \to \mathbb{S}^{d-1}$, mapping inputs to d-dimensional unit-norm vectors. Second, linear classifier weights $W \in \mathbb{R}^{C \times d}$ are learned on top, yielding $h(\cdot) = Wf(\cdot)$. While f is learned from unlabeled data, sampled from $\mathcal{D}_{\mathcal{X}}$, W is learned from labeled sample pairs $(x, y) \sim \mathcal{D}_{\mathcal{X},\mathcal{Y}}$. Learning is guided by a loss function $\mathcal{L}$ that measures the discrepancy between the model's prediction and the task objective: in supervised classification, predicted labels $\hat{y} = h(x)$ are compared with true labels y; in SSL, pseudo-labels are required to guide the learning of f. We outline this procedure in the following.

Pretext Task - Contrastive Instance Discrimination. In the case of contrastive instance discrimination, pseudo-labels $u \in \{0, 1\}$ are generated implicitly via data augmentations. These labels encode relative, rather than absolute, class information, indicating whether pairs of samples belong to the same or different classes. Let $\bar{x} \sim \mathcal{D}_{\mathcal{X}}$ denote unlabeled *anchor* instances. For each anchor, we sample two augmented views $x, x^+ \sim \mathcal{A}(\cdot \mid \bar{x})$ from the distribution of augmentations $\mathcal{A}$ conditioned on the anchor, and refer to the pair (x, x^+) as positive pair. In the following, we denote by $\mathcal{S}$ the distribution obtained by first sampling an anchor and then a positive pair. Augmentations are employed to induce invariance to specific transformations while preserving the semantic content of instances. Consequently, they are assumed not to change the underlying class label—a common assumption typically satisfied by the employed augmentation pipelines (Chen et al., 2020). The loss function measures the similarity between the representations of augmented views in the feature space, guiding the model to predict whether two views originate from the same instance ($u = 1$) or from different ones ($u = 0$) (Ericsson et al., 2022). In the following, we introduce a general joint embedding (JE) loss function, which encodes this task objective through two complementary terms, inspired by previous work (HaoChen et al., 2021; Garrido et al., 2023):

$$\mathcal{L}_{\text{JE}}(f) = \beta \mathcal{L}_{\text{inv}}(f) + \lambda \mathcal{L}_{\text{ctr}}(f) + c_{\lambda,\beta}, \tag{1}$$

with λ and β being non-negative hyperparameters, $\mathcal{L}_{\text{inv}}(f) = -\mathbb{E}_{(x,x^+)\sim\mathcal{S}}\left[f(x)^\top f(x^+)\right]$, $\mathcal{L}_{\text{ctr}}(f) = \mathbb{E}_{\bar{x},\bar{x}'\sim\mathcal{D}_\mathcal{X}}\,\mathbb{E}_{x\sim\mathcal{A}(\cdot|\bar{x}),\,x'\sim\mathcal{A}(\cdot|\bar{x}')}\left[\left(f(x)^\top f(x')\right)^2\right]$, and a constant $c_{\lambda,\beta}$, which depends on λ and β to ensure the non-negativity of the loss. For any two augmented views, $x \sim \mathcal{A}(\cdot \mid \bar{x})$ and $x' \sim \mathcal{A}(\cdot \mid \bar{x}')$, the similarity of their representations is measured via cosine similarity $\frac{f(x)^\top f(x')}{\|f(x)\|\|f(x')\|}$, which reduces to the numerator due to the normalized representation function, yielding unit Euclidean norm, $\|f(\cdot)\| = 1$. The invariance term $\mathcal{L}_{\text{inv}}$ encourages augmented views of the same instance, x and x^+, to obtain similar representations—invariant to the applied augmentations—guiding their cosine similarity toward 1, which represents the prediction goal $(u = 1)$. By squaring the cosine similarity, the contrastive term $\mathcal{L}_{\text{ctr}}$ encourages the representations of two random augmentations, x and x', to become orthogonal, guiding their cosine similarity toward 0, corresponding to the prediction goal $(u = 0)$.

Remark 1 *The prominent loss functions VICReg, BT, and SCL can all be expressed in terms of $\mathcal{L}_{inv}$. However, they vary in their contrastive criterion $\mathcal{L}_{ctr}$, employing different fourth-moment terms to discourage similar representations for pairs of random augmentations. Additionally, the constants $c_{\lambda,\beta}$ differ across these losses. Further details are provided in Section 3.*

Downstream Task - Linear Classification. In the downstream task, we consider the squared L2 loss

$$\mathcal{L}_{\text{SL}}(f, W) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathcal{X},\mathcal{Y}}}\left[\|Wf(x) - \boldsymbol{y}(x)\|_2^2\right], \tag{2}$$

where $W \in \mathbb{R}^{C\times d}$ denotes the classifier weight matrix with corresponding row vectors w_c for $c \in \mathcal{Y}$ and $\boldsymbol{y}(x) \in \mathbb{R}^C$ represents a one-hot encoding of the downstream label y for the data point x. In Section 3, we introduce a more general form of this loss function, which is closely related to the pretext loss function in Equation 1 and motivates the derivation of a transfer bound between them. To enable this derivation, we follow Arora et al. (2019) and introduce the mean classifier $\mathbf{W}^\mu = [\mu_1 \cdots \mu_C]^\top$, where each row $\mu_c = \mathbb{E}_{x\sim\mathcal{D}_c}[f(x)]$ corresponds to the mean representation of inputs from class $c \in \mathcal{Y}$. Note that an upper bound on the downstream loss incurred by the mean classifier also provides an upper bound on the downstream loss of the optimal classifier, by taking the infimum over all classifiers W. Consequently, establishing a transfer bound reduces to upper bounding the downstream loss of the mean classifier. Moreover, to assess downstream classification performance, we use top-1 accuracy, where the predicted label is defined as $\hat{y} = \arg\max_{c\in\mathcal{Y}}[Wf(x)]_c$. Given a labeled test set of size N, we define

$$\text{top-1}(f, W) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[y_i = \hat{y}_i].$$

PAC-Bayesian Framework. We employ PAC-Bayesian generalization bounds for the pretext task, following recent work that enables the derivation of non-vacuous generalization bounds within the PAC-Bayesian setting (Perez-Ortiz et al., 2021b; van Elst and Ghoshdastidar, 2025). Below, we briefly motivate the need for generalization bounds, introduce PAC-Bayesian theory, and present the PAC-Bayes with Backprop method (Perez-Ortiz et al., 2021b), yielding non-vacuous bounds.

Let $\mathcal{F}$ be the set of parametrized functions f_θ that can be learned by training a neural network with p learnable parameters $\theta \in \Theta$, from some parameter space $\Theta \subset \mathbb{R}^p$, with a learning algorithm $\mathcal{A}$. In the following, we omit the explicit parameter dependence for simplicity and use f and θ interchangeably, since f is fully specified by its parameters θ. Let $\mathcal{Z}$ be an example domain, such as $\mathcal{X}^2$ for pairs of inputs, with corresponding distribution $\mathcal{D}_\mathcal{Z}$, and let $\ell : \mathcal{Z} \times \Theta \to [0, 1]$ be a loss function defined on the example domain and parameter space Θ. The goal is to learn a function $f \in \mathcal{F}$, that minimizes the population risk $\mathcal{L}(\theta) = \mathbb{E}_{z\sim\mathcal{D}_\mathcal{Z}}\left[\ell\left(z, \theta\right)\right]$. Since we only have access to a finite training set $S = \{z_1, \ldots, z_n\} \in \mathcal{Z}^n$, we instead learn a function $\hat{f}$ from S, by minimizing the empirical risk $\hat{\mathcal{L}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(z_i, \theta\right)$. While the empirical risk typically becomes small during training—given a suitable hypothesis class $\mathcal{F}$—we are ultimately interested in the true risk, that is, the risk of the learned function on unseen data. In practice, this risk is estimated using a test set, whereas in theory, generalization bounds quantify the concentration of the empirical risk around the population risk through high-probability inequalities.

In contrast to classical generalization bounds, PAC-Bayesian bounds analyze distributions over functions instead of individual functions f. Let Q and P denote distributions over parameter vectors $\theta \in \Theta$, where Q is called the posterior and P the prior distribution. We adapt the population risk to be defined on the posterior distribution and obtain $\mathcal{L}(Q) = \mathbb{E}_{\theta\sim Q}\left[\mathcal{L}(\theta)\right]$. Likewise, the empirical risk is defined as $\hat{\mathcal{L}}(Q) = \mathbb{E}_{\theta\sim Q}\left[\hat{\mathcal{L}}(\theta)\right]$. Unlike classical generalization bounds, PAC-Bayesian bounds do not rely on complexity measures of the hypothesis class, such as Rademacher complexity, but instead measure the Kullback–Leibler (KL) divergence between the posterior and a prior distribution, defined as $\mathrm{KL}(Q \parallel P) = \mathbb{E}_{\theta\sim Q}\left[\log\left(\frac{dQ}{dP}(\theta)\right)\right]$. In contrast to Bayesian statistics, there are no restrictions on the posterior distribution Q, such as being related to the prior via a likelihood factor. While the prior distribution P serves only as a reference, choosing it in an informed manner has been shown to effectively control the size of the KL divergence in the bound (Perez-Ortiz et al., 2021b).

Classical PAC-Bayes bounds (McAllester, 1999) typically rely on Hoeffding's inequality (Hoeffding, 1994), which requires the empirical risk to be a mean of independent random variables. However, this does not hold for the pretext losses considered in this work. To address this, we follow the approach of van Elst and Ghoshdastidar (2025) and utilize a variant of McAllester's proof based on McDiarmid's inequality (McDiarmid et al., 1989). This concentration inequality generalizes Hoeffding's by applying to any function of independent random variables that satisfies the bounded differences property. This property guarantees that no single input has too much influence on the function's value, by requiring that the change of a single input variable can alter the output at most by a fixed constant. We state the bounded differences property as formalized in Boucheron et al. (2003):

Definition 2 (Bounded differences) *A function $g : \mathcal{Z}^n \to \mathbb{R}$ has the* bounded differences property *if there exist nonnegative constants $c_1, \ldots, c_n$, such that for all $1 \leq k \leq n$*

$$\sup_{z_1,\ldots,z_n,z'_k\in\mathcal{Z}} \left|g(z_1, \ldots, z_k, \ldots, z_n) - g(z_1, \ldots, z_{k-1}, z'_k, z_{k+1}, \ldots, z_n)\right| \leq c_k.$$

Leveraging this property, we state an extended version of McAllester's classic bound, as derived by van Elst and Ghoshdastidar (2025):

Theorem 3 (Extended PAC-Bayes-classic bound, van Elst and Ghoshdastidar, 2025) *Suppose the mapping $g : \mathcal{Z}^n \to \mathbb{R}$ such that $\hat{\mathcal{L}}(\theta) = g(S)$ for any dataset $S \in \mathcal{Z}^n$ and $\theta \in \Theta$. Assume g satisfies the bounded differences property with nonnegative constants $c_k = \frac{C}{n}$, for $k \in \{1, \ldots, n\}$. Then, for any data-free distribution P over weight space Θ, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the random draw of $S \sim_{iid} (\mathcal{D}_{\mathcal{Z}})^n$, for all distributions Q over Θ, we have*

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

Finally, we aim to find a distribution Q that minimizes the right-hand-side of the PAC-Bayes bound. We restrict our analysis to families of Gaussian distributions for both the prior and the posterior, which allows the bound to be minimized by training a probabilistic neural network (Langford and Caruana, 2001; Dziugaite and Roy, 2017; Perez-Ortiz et al., 2021b) using a training objective f_{classic} derived from the PAC-Bayes-classic bound. Methods following this approach are generally referred to as PAC-Bayes with Backprop (PBB) (Perez-Ortiz et al., 2021b). To control the influence of the KL divergence in f_{classic}, the KL term is scaled by a regularization coefficient η.

3. Generalization Analysis in the SSL-Based Two-Stage Training Pipeline

We now present our theoretical results, namely the derivation of generalization bounds for the downstream task of classification using representations learned from the task of contrastive instance discrimination. We proceed in three steps. First, we provide a unifying perspective on prominent joint embedding instance discrimination loss functions: Variance-Invariance-Covariance Regularization (VICReg) (Bardes et al., 2022), Barlow Twins (BT) (Zbontar et al., 2021), and Spectral Contrastive loss (SCL) (HaoChen et al., 2021). This permits us to derive largely unified proofs in subsequent sections. Second, we establish PAC-Bayesian pretext generalization bounds for the considered joint embedding losses, while accounting for their non-i.i.d. nature. The established pretext generalization bounds can then be instantiated with prior and posterior distributions P and Q that we learn through the PAC-Bayes with Backprop algorithm, yielding non-vacuous pretext generalization bounds. Third, we present our main contribution, that is, deriving practical transfer bounds for a new family of downstream loss functions, termed SupJE losses, which include the squared L2 loss as a special case, while accounting for varying clustering strengths of different augmentation pipelines. This derivation provides insights into effective hyperparameter values for the pretext losses, allowing us to derive practical guidelines for their configuration.

3.1. Unification of Pretext Losses

Following prior work (Cabannes et al., 2023; Simon et al., 2023), we consider theory-friendly variants of the loss functions to facilitate subsequent analysis. We now present the unification of the empirical-risk versions, while the population-risk unification is stated for reference in the Appendix in Proposition 10.

Proposition 4 (Unification of empirical pretext loss functions) *Let β and λ be non-negative hyperparameters, and let $S = \{(x_i, x_i^+)\}_{i=1}^n \sim \mathcal{S}^n$ be an i.i.d. sample of n positive pairs. Then, each* *method $m \in \{\textsc{vicreg}, \textsc{bt}, \textsc{scl}\}$, defined in Appendix A, can be expressed, up to an additive constant c_m, as a weighted combination of an invariance term $\mathcal{L}_{inv}$ and a contrastive term $\mathcal{L}_{ctr}$, scaled by β and λ, respectively:*

$$\hat{\mathcal{L}}_m(f) = \beta \mathcal{L}_{inv}(f) + \lambda \mathcal{L}_{ctr}(f) + c_m, \tag{3}$$

with $\mathcal{L}_{inv}(f) = -\frac{2}{n}\sum_{i=1}^{n} f(x_i)^\top f(x_i^+)$, $\mathcal{L}_{ctr}(f) = \frac{1}{n(n-1)}\sum_{i \neq j} t_{i,j}$, and

$$\begin{aligned}
\textit{VICReg}\colon t_{i,j} &= \tfrac{1}{4}\Big(\big(f(x_i)^\top f(x_j)\big)^2 + 2\big(f(x_i)^\top f(x_j^+)\big)^2 + \big(f(x_i^+)^\top f(x_j^+)\big)^2\Big), c_{\textit{VICReg}} = 2\beta - \tfrac{\lambda}{d},\\
\textit{BT}\colon t_{i,j} &= \tfrac{1}{2}\Big(f(x_i^+)^\top f(x_j^+) f(x_j)^\top f(x_i) + f(x_i^+)^\top f(x_j) f(x_j^+)^\top f(x_i)\Big), \lambda = 1, c_{\textit{BT}} = \beta^2 d,\\
\textit{SCL}\colon t_{i,j} &= \Big(f(x_i)^\top f(x_j^+)\Big)^2, \beta = 1, c_{\textit{SCL}} = 2 - \tfrac{\lambda}{d}.
\end{aligned}$$

Proof The unification proof is provided in Appendix B. ■

For BT and SCL, some hyperparameters do not appear and are therefore set to 1, whereas for VICReg both hyperparameters are actually tunable. While at the population level, SCL represents a special case of VICReg for $\beta = 1$ (see Appendix D), the empirical versions further differ in how the symmetrization of augmented views is handled: VICReg considers explicit symmetrization (see Appendix B), while SCL simply restricts samples x_i, x_j^+ in $\mathcal{L}_{\text{ctr}}$ to be drawn from different views.

3.2. PAC-Bayesian Pretext Generalization Bounds

Proposition 4 shows that the summands of the contrastive terms are dependent, which prevents the application of PAC-Bayesian bounds that rely on Hoeffding's inequality, requiring the empirical loss to be expressed as a sum of independent random variables. However, our pretext losses satisfy the bounded differences assumption from Definition 2, which allows us to derive a Corollary of Theorem 3.

Corollary 5 (Extended PAC-Bayes-classic bound for pretext loss functions) *Fix a data-free distribution P over the function space $\mathcal{F}$ and a method $m \in \{\textsc{vicreg}, \textsc{bt}, \textsc{scl}\}$. Let the empirical loss $\hat{\mathcal{L}}_m$ be defined as in Proposition 4, let $\mathcal{L}_m$ denote the corresponding population loss, and extend both to distributions Q over $\mathcal{F}$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over size-n i.i.d. random samples $S \sim \mathcal{S}^n$, for all Q, we have*

$$\mathcal{L}_m(Q) \leq \hat{\mathcal{L}}_m(Q) + C_m \cdot \sqrt{\frac{KL(Q\|P) + \log\frac{2n}{\delta}}{2(n-1)}},$$

with $C_{\textit{VICReg}} = 4(\beta + \frac{\lambda}{2})$, $C_{\textit{BT}} = 4(\beta + \frac{1}{2})$ and $C_{\textit{SCL}} = 4(1 + \frac{\lambda}{2})$.

Proof We first prove that the bounded differences property holds for our considered pretext loss functions with respective constants C_m. The rest of the proof proceeds as in Theorem 3.2 from van Elst and Ghoshdastidar (2025). The detailed proof is provided in Appendix E. ■

3.3. Practical Transfer Bounds

In the following, we derive practically computable transfer bounds by building on the mean classifier approach (Arora et al., 2019), which requires that the population pretext and downstream loss functions share a similar form. To this end, we design a new family of downstream loss functions, which we call supervised joint embedding (SupJE) loss family, and which can be regarded as a supervised variant of the joint embedding pretext loss defined in Equation 1. Moreover, we can rewrite the pretext population losses from Proposition 10 to admit the form of Equation 1. Given the established similarity between the pretext and downstream losses, we can employ the mean classifier approach to derive a transfer bound. Additionally, we show that the squared L2 loss (SL) represents a special case of the SupJE loss by fixing a specific ratio between the involved hyperparameters, denoted as the SL-ratio. From this observation, we not only obtain a transfer bound for SL, but also show in Section 4 that SupJE variants selecting hyperparameter ratios near the SL-ratio achieve the best downstream performance. Since, in our derivation, the hyperparameter configuration of the downstream loss carries over to the pretext loss, we are able to infer effective hyperparameter configurations for pretext training.

Below, we offer the intuition behind the proof of the transfer bound, before presenting the theorems along with their formal proofs. We first note that VICReg and SCL take the same form as the pretext loss in Equation 1, since no symmetrization needs to be considered at the population level. However, the contrastive term in Barlow Twins differs. To further unify the framework, we rewrite the contrastive term of Barlow Twins as follows:

Lemma 6 (BT contrastive term reformulation) *Let*

$$\zeta = \frac{1}{2}\mathbb{E}_{(x,x^+),\,(x',x'^+)\sim\mathcal{S}^2}\left[\left(f(x^+)^\top f(x') - f(x'^+)^\top f(x)\right)^2\right].$$

By rewriting the expression, we obtain:

$$\mathbb{E}_{(x,x^+),\,(x',x'^+)\sim\mathcal{S}^2}\left[f(x^+)^\top f(x') f(x'^+)^\top f(x)\right] = \mathbb{E}_{x,x'}\left[\left(f(x)^\top f(x')\right)^2\right] - \zeta.$$

Proof The proof is provided in Appendix F. ■

The additional term ζ is absorbed into the final bound, and its magnitude and computation are discussed in Remark 14. With this reformulation, all considered pretext losses are now expressed through the invariance and contrastive terms of Equation 1, which we restate below for direct comparison with the SupJE loss:

$$\mathcal{L}_{\text{inv}}(f) = -\mathbb{E}_{x,x^+}\left[f(x)^\top f(x^+)\right] \quad \text{and} \quad \mathcal{L}_{\text{ctr}}(f) = \mathbb{E}_{x,x'}\left[\left(f(x)^\top f(x')\right)^2\right].$$

Next, we introduce the SupJE loss:

$$\mathcal{L}_{\text{SupJE}}(f, W) = \mathbb{E}_{(x,y)}\left[-2\beta\left(f(x)^\top w_y\right) + \frac{\lambda}{C}\sum_{c\in[C]}\left(f(x)^\top w_c\right)^2\right] + c_m, \tag{4}$$

with $m \in \{\text{VICReg},\ \text{BT},\ \text{SCL}\}$, nonnegative hyperparameters β and λ, and c_m from Proposition 10 for the non-negativity of the loss. The first term in the expectation of SupJE resembles $\mathcal{L}_{\text{inv}}$, encouraging alignment between $f(x)$ and the classifier's weight vector w_y, instead of $f(x^+)$, where y denotes the class label of input x. The second term in the expectation of SupJE resembles $\mathcal{L}_{\text{ctr}}$, contrasting the representation $f(x)$ with the classifier's weight vectors for all classes instead of $f(x')$—a random representation corresponding to some underlying class y'. The similar form of the losses suggests to think of the representations $f(x^+)$ and $f(x')$ in the pretext loss as classifier weights guiding the classification of $f(x)$. The key idea behind the mean classifier approach is to use the mean representation of a class, $\mu_{y'} = \mathbb{E}_{x \sim \mathcal{D}_{y'}}[f(x)]$, rather than random class representatives, such as $f(x')$ for class y', as classifier. Intuitively, this should improve classification guidance and yield a lower bound on the pretext loss, which is required to establish the subsequent transfer bound.

Theorem 7 (Transfer bound for SupJE loss) *Let $\mathcal{L}_{SupJE}$ denote the supervised joint embedding loss defined in Equation 4, and let $\mathcal{L}_m$ be the population counterpart of the empirical pretext loss defined in Proposition 4, with $m \in \{\textsc{VICReg}, \textsc{BT}, \textsc{SCL}\}$. Moreover, let π be a uniform class prior. Then, for all $f : \mathcal{X} \to \mathbb{S}^{d-1}$, we obtain:*

$$\inf_{W \in \mathbb{R}^{C \times d}} \mathcal{L}_{SupJE}(f, W) \leq \mathcal{L}_m(f) + \gamma\sigma' + \mathbf{1}_{\{m=BT\}}\zeta,$$

with $\gamma = 2\beta$, $\sigma' = \mathbb{E}_{x,x^+,y}\left[\left|f(x)^\top(f(x^+) - \mu_y)\right|\right]$, $\mathbf{1}_{\{\cdot\}}$ being the indicator function and ζ as defined in Lemma 6.

Proof sketch. We begin by providing lower bounds for the invariance and contrastive terms of the pretext loss separately. For the invariance term, we refine the bound established by van Elst and Ghoshdastidar (2025), which in turn refines the earlier bound introduced by Wang et al. (2022). These refinements arise from a more careful application of the Cauchy–Schwarz inequality. For the contrastive term, we adopt the mean classifier approach from Arora et al. (2019): formally, the intuition provided above can be proven by establishing the loss lower bound in terms of the mean classifier through Jensen's inequality. The proof follows by combining both parts and taking the infimum over all linear classifiers W. The complete derivation is provided in Appendix G.

Next, we state the transfer bound for the squared L2 loss as a special case of the SupJE loss by introducing the SL-ratio hyperparameter restriction:

Corollary 8 (Transfer bound for squared L2 loss) *Let $\mathcal{L}_{SL}$ denote the squared L2 loss defined in Equation 2, let $\mathcal{L}_m$ be the population counterpart of the empirical pretext loss defined in Proposition 4, with $m \in \{\textsc{VICReg}, \textsc{BT}, \textsc{SCL}\}$, and impose the following restrictions on hyperparameter combinations (SL-ratio): $\beta = \frac{\lambda}{C}$ and $\lambda_{SCL} = C$. Moreover, let π be a uniform class prior. Then, for all $f : \mathcal{X} \to \mathbb{S}^{d-1}$, we obtain:*

$$\inf_{W \in \mathbb{R}^{C \times d}} \mathcal{L}_{SL}(f, W) \leq \frac{\mathcal{L}_m(f) + \gamma\sigma' + \Delta_m}{\beta},$$

with $\sigma' = \mathbb{E}_{x,x^+,y}\left[\left|f(x)^\top(f(x^+) - \mu_y)\right|\right]$, $\gamma = 2\beta$, $\Delta_{VICReg} = \lambda(\frac{1}{d} - \frac{1}{C})$, $\Delta_{SCL} = -1 + \frac{C}{d}$ and $\Delta_{BT} = \frac{(1-\frac{d}{C})}{C} + \zeta$, with ζ as defined in Lemma 6.

Proof The proof follows by employing the hyperparameter constraint. Details are provided in Appendix H. ■

Corollary 8 provides insights into which hyperparameter combinations enable the pretext losses to translate into the squared L2 loss for linear downstream classification, namely all combinations where the weighting ratio between the invariance and contrastive terms is $2 : C$, with C being the number of classes in the downstream task. Thus, increasing the number of downstream classes requires a proportionally greater emphasis on the contrastive term.

Remark 9 *In Corollary 8, we assume a uniform class prior distribution π. To address class imbalance in the downstream task, we can reweight the squared L2 loss function to assign more weight to underrepresented classes, as is typically supported by standard implementations of the cross-entropy loss. We introduce such a balanced SL function and accordingly adapt the proof of Corollary 8 in Appendix I. For a detailed discussion of class-imbalanced settings, we refer to Section 5.*

4. Experiments

In this section, we (i) empirically evaluate the tightness of our downstream generalization bounds across augmentation pipelines of varying strength, specifically considering the presence or absence of random cropping, and (ii) show that based on the hyperparameter SL-ratio introduced in Corollary 8, a range of effective hyperparameter combinations can be identified.

4.1. Experimental Setup

We briefly outline the experimental setup used to derive downstream generalization bounds, with further technical details and additional experiments on the MNIST dataset provided in Appendix K. Our implementation is based on the publicly available code provided by van Elst and Ghoshdastidar (2025). In the following, we conduct experiments on the CIFAR-10 dataset (Krizhevsky and Hinton, 2009), using the following data split:

$$|D_{\text{train}}| = 50{,}000, \quad |D_{\text{test}}| = 10{,}000, \quad |D_{\text{prior}}| = 0.8 \cdot |D_{\text{train}}|, \quad |D_{\text{posterior}}| = 0.2 \cdot |D_{\text{train}}|.$$

Moreover, we follow standard preprocessing practices by normalizing image pixels channel-wise. For the pretext task, if not stated otherwise, we apply the data augmentation pipeline used in van Elst and Ghoshdastidar (2025), following standard SimCLR augmentations (Chen et al., 2020), see Appendix K.

The derivation of downstream generalization bounds involves four steps, with the first three focused on obtaining the pretext generalization bounds using the PAC-Bayes with Backprop algorithm. First, we learn a prior Gaussian distribution over model weights by training a probabilistic neural network on D_{prior}, minimizing the PAC-Bayes training objective f_{classic}, with a KL regularization coefficient $\eta = 10^{-6}$, effectively reducing the training to empirical risk minimization. This has been shown to be more effective than randomly initializing the prior (Dziugaite and Roy, 2018; Perez-Ortiz et al., 2021b,a; van Elst and Ghoshdastidar, 2025). Second, we learn a posterior Gaussian distribution over model weights by minimizing the PAC-Bayes training objective f_{classic} on the entire training dataset D_{train}, without using KL regularization. Third, we evaluate the pretext generalization bounds from Corollary 5 using the learned distributions Q and P on $D_{\text{posterior}}$, ensuring independence from the data used to learn the prior. The bounds are then compared to the pretext test loss, to assess their tightness. The test loss is evaluated using a randomized function f_θ, where the weights θ are sampled from the probabilistic posterior network P. Fourth, we follow the linear evaluation protocol (see Section 5 for a discussion of different evaluation protocols) and train a linear layer on top of the frozen pretrained network using D_{train} with corresponding labels. The final layers of the pretrained network, called the projection head, are often discarded after pretraining. We consider both scenarios in our experiments: one in which the projection head is discarded, and one in which it is included. The transfer bounds in Section 3.3 are derived under the assumption that the projection head is included.

4.2. Results

We start by evaluating the tightness of the downstream generalization bounds when considering a standard augmentation pipeline including random cropping. To this end, we first look at the *pretext generalization bound* in isolation and then plugged into the *transfer bound* to obtain a bound on the downstream risk.[1] In the upper part of Table 2, we observe that **our pretext generalization bound is consistently tighter than previous ones and reasonably close to the test loss in the last three columns.** We note that in the first three columns, the bounds remain loose. This can be attributed to two main factors: (i) small hyperparameter values yield low overall loss values, but the classic PAC-Bayes training objective becomes loose when the pretext population risk $\mathcal{L}_m(Q) < \frac{1}{4}$ (see Discussion in Section 5), and (ii) large hyperparameter values increase the bounded differences coefficient, which in turn puts more emphasis on the KL divergence.

		VICReg			SCL	BT	VICReg
		$\beta = 0.1, \lambda = 1$	$\beta = 0.5, \lambda = 5$	$\beta = 1, \lambda = 10$	$\beta = 1, \lambda = 1$	$\beta = 1, \lambda = 1$	$\beta = 1, \lambda = 1$
	Pretext test loss	0.075	0.363	0.726	0.419	260.917	0.426
	Classic bound (iid) (16)	0.456	2.441	5.475	1.876	337.957	1.938
	Catoni's bound (iid) (17)	0.288	1.687	4.361	1.37	275.273	1.465
	Cor. 5 (ours)	0.137	0.686	1.418	0.583	261.081	0.592
	KL/n	2.060×10^{-5}	1.333×10^{-4}	3.498×10^{-4}	1.009×10^{-4}	1.321×10^{-4}	1.539×10^{-4}
Th. 7	Risk certificate	0.252	1.271	2.589	1.246	261.797	1.238
	Test loss	0.190	0.948	1.896	1.082	261.634	1.072
Proj.	SupJE loss	0.148	0.734	1.476	0.967	255.014	0.971
	Top-1	0.706	0.717	0.706	0.485	0.471	0.463

Table 2: *Upper Part* — Comparison of PAC-Bayes risk certificates for pretext losses from Proposition 4 across different hyperparameter combinations to the pretext test loss as well as to other bounds established in the literature (see Appendix J). The Kullback–Leibler (KL) divergence between prior and posterior distributions is reported per dataset size n. *Lower Part* — Comparison of the transfer bound from Theorem 7 with the corresponding downstream test loss. Top-1 accuracy is reported to assess downstream performance. The first three columns correspond to hyperparameter combinations that satisfy the constraints of Corollary 8, whereas the last three columns present an exemplary deviating hyperparameter combination.

In the lower part of the table, we compare the downstream SupJE test loss with the guarantees provided by the transfer bound from Theorem 7, instantiated either with the pretext risk certificate or the pretext test loss. Using the latter in the transfer bound yields **downstream bounds that are consistently close to the SupJE test loss across all different hyperparameter combinations, validating that our transfer bound adequately captures the surrogate relation.** This is also reflected by the fact that, using the pretext generalization certificates in the transfer bound, the tightness of the pretext certificates translates to the downstream certificates. Hence, while the

[1] Note that, to obtain an end-to-end generalization bound and provide guarantees on labeled sample complexity, we would also need to account for learning the linear classifier from data. Here, we focus on evaluating the transfer bound.

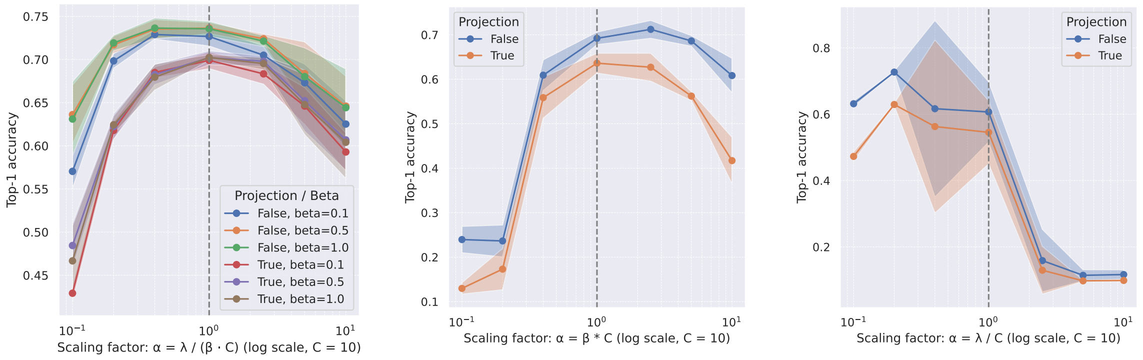
bounds in the first three columns inherit their looseness from the pretext bounds, certificates on the downstream performance in the last three columns remain reasonably tight.

Our refinement of the bound on the invariance term, together with the resulting measure of the augmentations' clustering strength, σ' (as defined in Theorem 7), constitutes a key advance toward more accurate transfer bounds. Since no other computable downstream bounds exist for our specific loss combination, we compare σ' in isolation with prior work. van Elst and Ghoshdastidar (2025) capture the clustering strength of the augmentation pipeline via the intra-class feature deviation, $\sigma = \mathbb{E}_{(x,y)}\left[\|f(x) - \mu_y\|_2\right]$. Effectively, σ and σ' quantify the inverse clustering strength of the augmentation pipeline; that is, lower intra-class feature deviations indicate stronger clustering. In Table 3, we compare these measures in two settings: once employing the full augmentation pipeline during pretraining, and once excluding random cropping, which is considered a key contributor to clustering strength. While σ can barely distinguish between the two different augmentation settings, **our refined σ' aligns more closely with the observed downstream performance.** Specifically, when cropping is applied, the stronger top-1 accuracy is reflected by a smaller σ', resulting in tighter risk certificates compared to those computed with σ. Moreover, when random cropping is excluded, our downstream bounds—though slightly underestimated when approximated via the test loss—still capture the observed performance reasonably well. This is noteworthy, as Saunshi et al. (2022) show that previous bounds, which do not account for the inductive bias of the function class, become vacuous under disjoint augmentation distributions—a scenario we simulate by excluding cropping from the augmentation pipeline. Consequently, compared to previous work, **our explanation of performance transfer is more general**.

		$\beta = 0.1, \lambda = 1$		$\beta = 0.5, \lambda = 5$		$\beta = 1, \lambda = 10$	
	Cropping applied	×	✓	×	✓	×	✓
Th. 7	Risk certificate	0.2604	0.2519	1.3014	1.2709	2.6055	2.5889
	Test loss	0.1991	0.1899	0.9939	0.9480	1.9892	1.8964
Proj.	SupJE loss	0.2047	0.1483	1.0077	0.7339	2.0324	1.4759
	Top-1	0.3463	0.7057	0.3695	0.7168	0.3591	0.7065
	σ	0.9917	0.9198	0.9924	0.9201	0.9926	0.9211
	σ' (refined, ours)	0.9565	0.5732	0.9642	0.5853	0.9653	0.5853

Table 3: Comparison of downstream bounds for VICReg across different hyperparameter combinations computed once using the full augmentation pipeline and once without random cropping. We report the intra-class feature deviation σ provided in Theorem 3.15 of van Elst and Ghoshdastidar (2025) alongside our refined σ'.

Finally, we highlight a key finding derived from our theoretical analysis: In Figure 1, we show that **the SL-ratio required in Corollary 8, to relate the pretext losses to the downstream squared L2 loss, represents an effective heuristic for setting hyperparameters**, since the best hyperparameters lie within a small range around that ratio. We explore this by evaluating downstream performance across different scalings of the SL-ratio. Since this hyperparameter restriction leaves VICReg with one degree of freedom, each ratio is evaluated for three distinct β values, showing only slight performance differences and indicating that the hyperparameter ratio, rather than the absolute values matter. While the trend is clear for VICReg—**moving further away from the SL-ratio worsens downstream performance**—we observe a slight shift in the optimal hyperparameters for BT, which may be attributed to the different form of BT's contrastive term that required a reformulation in our bound (see Lemma 6). For SCL, we cannot draw conclusions due to the high variance resulting from training instabilities and the limited number of run iterations imposed by computational constraints. Lastly, we show in Table 6 (Appendix L) that the **SupJE loss performs on par with the standard cross-entropy loss**, both under the SL-ratio and when deviating from this "ideal ratio". Hence, the observed performance drops cannot be attributed to the specific form of the SupJE loss, indicating that these hyperparameter ratios impair classification performance regardless of the employed downstream loss, which ultimately highlights the significance of effective pretext hyperparameter ratios.



VICReg BT SCL

Figure 1: Downstream performance is evaluated in terms of top-1 accuracy for different hyperparameter configurations obtained by scaling the ratio from Corollary 8 (SL-ratio) with $\alpha \in \{0.1, 0.2, 0.4, 1, 2.5, 5, 10\}$ for different pretext loss functions. The dashed line denotes use of the SL-ratio ($\alpha = 1$). Each result is averaged over five runs with different random seeds. The shaded area around the curves represents the standard deviation. We plot downstream top-1 accuracy with and without the projection head.

5. Discussion and Future Work

For the first time, we provide practically computable generalization bounds for the downstream task of classification when using state-of-the-art SSL joint embedding methods. Moreover, we adequately capture the performance transfer under varying augmentation strength. In addition, we identify a heuristic for selecting pretext hyperparameters based on the SL-ratio (see Corollary 8), which is particularly effective for the most commonly employed VICReg loss, thereby mitigating the need for extensive hyperparameter tuning. In summary, our bounds offer (1) guarantees on the model's true risk through non-vacuous generalization bounds and (2) guidance for model selection through the proposed hyperparameter configurations. This paves the way for *self-certified learning* (Perez-Ortiz et al., 2021b), eliminating the need for held-out test and validation datasets, thus allowing all available data to be used for training. In the following, we discuss remaining challenges, and outline directions for future research.

Loss function design. The unification of pretext loss functions provides insights into the key structural elements of these losses. In particular, we observe that the considered loss functions primarily differ in their contrastive terms by considering different fourth-moments (see Proposition 10). This observation suggests the exploration of novel loss functions based on alternative fourth-moment terms or suitable higher-order moments. Beyond these insights into specific pretext loss structures, our derivation of downstream-informed pretext hyperparameters motivates a shift toward a *backwards-driven paradigm* for SSL loss function design. Rather than designing a pretext task and its corresponding loss function heuristically, we propose to start from a downstream task and a corresponding downstream loss function, chosen by the practitioner, and apply mean-classifier-style approaches backwards, thereby deriving a pretext loss that provably bounds the loss of the downstream task. We see great potential for principled SSL loss-function design through this backwards-driven perspective.

Uniform prior assumption. We assume a uniform class prior in the transfer bounds of Section 3.3. This is not a strong constraint in supervised learning, as class imbalance is typically addressed during training by up- or downsampling certain classes to restore balance and during evaluation by adjusting metrics accordingly. Alternatively, the loss function can be reweighted as shown in Appendix I. However, some regularization-based joint embedding losses are inherently limited in handling imbalanced datasets. Assran et al. (2022) show that VICReg induces a bias toward representations that partition the data into nearly uniformly sized groups. They propose reformulating pretext loss functions to favor more realistic long-tailed, power-law distributions. Therefore, adapting pretext loss functions such as VICReg in this manner represents an important direction for future work.

Linear evaluation. SSL methods are typically evaluated on a variety of downstream tasks using different protocols (Marks et al., 2025). There are two perspectives on SSL: one views it as feature extraction, and the other as a means of providing a better network initialization than random initialization or supervised pretraining (Goyal et al., 2019; Marks et al., 2025). While feature extraction is typically evaluated through linear probing, the initialization perspective is evaluated through fine-tuning, where the pretrained network is allowed to update during downstream training. In this work, we focus on the feature extraction perspective to rigorously examine the relationship between the pretext and downstream tasks, without allowing information to flow in both directions.

Tighter PAC-Bayes bounds in the low-loss regime. Achieving self-certified learning requires tight generalization bounds for models trained with effective hyperparameter combinations. However, in the low-loss regime, where the population loss $\mathcal{L}_m(Q) < \frac{1}{4}$, the classic PAC-Bayes training objective is known to be looser than a refined quadratic objective (Rivasplata et al., 2019; Perez-Ortiz et al., 2021b). This objective can be derived from the PAC-Bayes-kl bound (Langford and Seeger, 2001) by applying a tighter version of Pinsker's inequality (Perez-Ortiz et al., 2021b). However, due to the non-i.i.d. nature of the pretext losses, we cannot directly derive the quadratic training objective from the PAC-Bayes-kl bound. Hence, it is necessary to investigate how the quadratic objective can be derived in the non-i.i.d. setting. While one could compute the bound over i.i.d. batches, this typically results in weaker certificates, particularly in scenarios with a small number of large batches (van Elst and Ghoshdastidar, 2025).

Acknowledgments

This work has been supported by the German Research Foundation (DFG) through the Research Training Group GRK 2428 ConVeY and by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research (BMBF).

References

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. In *36th International Conference on Machine Learning, ICML 2019*, pages 9904–9923. International Machine Learning Society (IMLS), 2019.

Mahmoud Assran, Randall Balestriero, Quentin Duval, Florian Bordes, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, and Nicolas Ballas. The hidden uniform cluster prior in self-supervised learning. *arXiv preprint arXiv:2210.07277*, 2022.

Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 26671–26685. Curran Associates, Inc., 2022.

Adrien Bardes, Jean Ponce, and Yann Lecun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR 2022-International Conference on Learning Representations*, 2022.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Summer school on machine learning*, pages 208–240. Springer, 2003.

Vivien Cabannes, Bobak Kiani, Randall Balestriero, Yann LeCun, and Alberto Bietti. The ssl interplay: Augmentations, inductive bias, and generalization. In *International conference on machine learning*, pages 3252–3298. PMLR, 2023.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020.

Olivier Catoni. Pac-bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*, 2007.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iii. *Communications on pure and applied Mathematics*, 29(4):389–461, 1976.

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

Gintare Karolina Dziugaite and Daniel M Roy. Data-dependent pac-bayes priors via differential privacy. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M. Hospedales. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62, 2022. doi: 10.1109/MSP.2021.3134634.

Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann Lecun. On the duality between contrastive and non-contrastive self-supervised learning. In *ICLR 2023-Eleventh International Conference on Learning Representations*, 2023.

Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the ieee/cvf International Conference on computer vision*, pages 6391–6400, 2019.

Han Bao, Yoshihiro Nagano, and Kento Nozawa. On the surrogate gap between contrastive and supervised losses. *International Conference on Machine Learning*, pages 1585–1606, 2022. ISSN 2640-3498.

Jeff Z HaoChen and Tengyu Ma. A theoretical study of inductive biases in contrastive learning. *arXiv preprint arXiv:2211.14699*, 2022.

Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 5000–5011. Curran Associates, Inc., 2021.

Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pages 4182–4192. PMLR, 2020.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

John Langford and Rich Caruana. (not) bounding the true error. In *Proceedings of the 15th International Conference on Neural Information Processing Systems: Natural and Synthetic*, pages 809–816, 2001.

John Langford and Matthias Seeger. Bounds for averaging classifiers. Technical Report CMU-CS-01-102, Carnegie Mellon University, 2001.

Yann LeCun, Corinna Cortes, and Christopher J. C. Burges. MNIST handwritten digit database, 2010. URL https://www.kaggle.com/datasets/hojjatk/mnist-dataset.

Markus Marks, Manuel Knott, Neehar Kondapaneni, Elijah Cole, Thijs Defraeye, Fernando Perez-Cruz, and Pietro Perona. A closer look at benchmarking self-supervised pre-training with image classification. *International Journal of Computer Vision*, pages 1–13, 2025.

David A McAllester. Pac-bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170, 1999.

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

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6707–6717, 2020.

Kento Nozawa, Pascal Germain, and Benjamin Guedj. Pac-bayesian contrastive unsupervised representation learning. In *Conference on Uncertainty in Artificial Intelligence*, pages 21–30. PMLR, 2020.

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

Maria Perez-Ortiz, Omar Rivasplata, Benjamin Guedj, Matthew Gleeson, Jingyu Zhang, John Shawe-Taylor, Miroslaw Bober, and Josef Kittler. Learning pac-bayes priors for probabilistic neural networks. *arXiv preprint arXiv:2109.10304*, 2021a.

Maria Perez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22(227):1–40, 2021b.

Omar Rivasplata, Vikram M Tankasali, and Csaba Szepesvari. Pac-bayes with backprop. *arXiv preprint arXiv:1908.07380*, 2019.

Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pages 19250–19286. PMLR, 2022.

Ravid Shwartz-Ziv, Randall Balestriero, Kenji Kawaguchi, Tim G. J. Rudner, and Yann LeCun. An information theory perspective on variance-invariance-covariance regularization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 33965–33998. Curran Associates, Inc., 2023.

James B Simon, Maksis Knutins, Liu Ziyin, Daniel Geisz, Abraham J Fetterman, and Joshua Albrecht. On the stepwise nature of self-supervised learning. In *International Conference on Machine Learning*, pages 31852–31876. PMLR, 2023.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020.

Anna van Elst and Debarghya Ghoshdastidar. Tight pac-bayesian risk certificates for contrastive learning. *SIAM Journal on Mathematics of Data Science*, 7(4):1904–1927, 2025.

Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. *arXiv preprint arXiv:2203.13457*, 2022.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.

Appendix A. Definition of pretext loss functions

Variance-Invariance-Covariance Regularization loss (VICReg). We consider a theory-friendly version of VICReg (Bardes et al., 2022), similar to the one proposed by Cabannes et al. (2023). It differs from the original loss in the following points: (i) it replaces the hinge loss variance term by a least squares loss as introduced in Balestriero and LeCun (2022), which consequently allows to combine variance and covariance criteria, (ii) scales the identity matrix w.r.t the embedding dimension d to account for feature normalization and (iii) considers the second moment instead of the covariance:

$$\mathcal{L}_{\text{VICReg}}(f) = \lambda \left\| \mathbb{E}_x \left[f(x) f(x)^\top \right] - \frac{1}{d} I_d \right\|_F^2 + \beta \mathbb{E}_{x,x^+} \left[\left\| f(x) - f(x^+) \right\|_2^2 \right]. \tag{5}$$

Barlow Twins loss (BT). Again, we consider a simplified version of the original BT loss (Zbontar et al., 2021), similar to the one proposed by Simon et al. (2023), where the hyperparameter for weighting the off-diagonal terms is set to 1 and batch normalization is removed. We reintroduce a hyperparameter, β, to account for feature normalization. As shown later, it also plays a role similar to the β parameter in VICReg, hence the naming. The loss is defined as

$$\mathcal{L}_{\text{BT}}(f) = \left\| \mathbb{E}_{x,x^+} \left[f(x) f(x^+)^\top \right] - \beta I_d \right\|_F^2. \tag{6}$$

Spectral Contrastive loss (SCL). We consider a slightly modified version of SCL proposed in HaoChen et al. (2021), which can be seen as a theory-friendly version of the contrastive loss used in the SimCLR framework (Chen et al., 2020). In line with the generalized spectral contrastive loss (HaoChen and Ma, 2022), we introduce a hyperparameter λ for the contrastive term. Moreover, we add a constant factor of $2 - \frac{\lambda}{d}$ to ensure the non-negativity of the loss (see Appendix D). This simplifies the analysis and does not affect gradient-based optimization. The loss is defined as

$$\mathcal{L}_{\text{SC}}(f) = -2\mathbb{E}_{x,x^+} \left[f(x)^\top f(x^+) \right] + \lambda \mathbb{E}_{x,x'} \left[\left(f(x)^\top f(x') \right)^2 \right] + 2 - \frac{\lambda}{d}. \tag{7}$$

Appendix B. Proof of Proposition 4

In the following, we demonstrate the unification of the loss functions presented in Appendix A in their empirical form.

Spectral Contrastive Loss. Since the population version of SCL is already presented in expanded form, we directly state the empirical version used by HaoChen et al. (2021), incorporating our modifications. This empirical version is obtained from the population form by replacing expectations with empirical averages and omitting terms where $i = j$:

$$\mathcal{L}_{\text{SC}}(f) = -\frac{2}{n}\sum_{i=1}^{n} f(x_i)^\top f(x_i^+) + \frac{\lambda}{n(n-1)}\sum_{i\neq j} f(x_i)^\top f(x_j^+) + 2 - \frac{\lambda}{d}. \tag{8}$$

For VICReg and BT, we begin by stating the empirical versions in their symmetrized form. We then expand these expressions and subsequently drop the dependent terms where $i = j$. The empirical VICReg loss is given by

$$\begin{aligned} \mathcal{L}_{\text{VICReg}}(f) &= \lambda \left\| C - \frac{1}{d} I \right\|_F^2 + \frac{\beta}{n}\sum_{i=1}^{n} \left\| f(x_i) - f(x_i^+) \right\|_2^2, \\ \text{with } C &= \frac{1}{2n}\sum_{i=1}^{n}\left(f(x_i)f(x_i)^\top + f(x_i^+)f(x_i^+)^\top \right), \end{aligned} \tag{9}$$

while the empirical BT loss takes the form

$$\begin{aligned} \mathcal{L}_{\text{BT}}(f) &= \left\| C - \beta I_d \right\|_F^2, \\ \text{with } C &= \frac{1}{2n}\sum_{i=1}^{n}\left(f(x_i)f(x_i^+)^\top + f(x_i^+)f(x_i)^\top \right). \end{aligned} \tag{10}$$

To facilitate a unified derivation, we introduce some notation in the following. Let $\mathcal{L}_{\text{ctr}}(f) = \|C - \gamma I\|_F^2$, where we set $\gamma = \frac{1}{d}$ for VICReg and $\gamma = \beta$ for BT. Moreover, we define $\mathcal{B}_i^\ell$ with $\ell \in \{\text{vicreg}, \text{bt}\}$ as $\mathcal{B}_i^{\text{vicreg}} = f(x_i)f(x_i)^\top + f(x_i^+)f(x_i^+)^\top$ and $\mathcal{B}_i^{\text{bt}} = f(x_i)f(x_i^+)^\top + f(x_i^+)f(x_i)^\top$. We begin by expanding $\mathcal{L}_{\text{ctr}}$:

$$\begin{aligned} \mathcal{L}_{\text{ctr}}(f) &= \|C - \gamma I\|_F^2 \\ &= \operatorname{Tr}\left((C-\gamma I)^\top (C - \gamma I) \right) \\ &= \operatorname{Tr}(C^2) - 2\gamma \operatorname{Tr}(C) + \gamma^2 \operatorname{Tr}(I) \\ &= \frac{1}{4n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{Tr}\left((\mathcal{B}_i^\ell)^\top \mathcal{B}_j^\ell \right) - \frac{\gamma}{n}\sum_{i=1}^{n} \operatorname{Tr}(\mathcal{B}_i^\ell) + \gamma^2 d \end{aligned} \tag{11}$$

VICReg. We now focus exclusively on VICReg and expand the terms in Equation 11 for $\mathcal{B}_i^{\text{vicreg}}$. We begin with the single-sum term:

$$\begin{aligned}
\sum_{i=1}^{n} \mathrm{Tr}(\mathcal{B}_i^{\text{vicreg}}) &= \sum_{i=1}^{n} \mathrm{Tr}\left(f(x_i) f(x_i)^\top + f(x_i^+) f(x_i^+)^\top \right) \\
&= \sum_{i=1}^{n} \left(\mathrm{Tr}\left(f(x_i)^\top f(x_i) \right) + \mathrm{Tr}\left(f(x_i^+)^\top f(x_i^+) \right) \right) \\
&= \sum_{i=1}^{n} \|f(x_i)\|_2^2 + \left\|f(x_i^+)\right\|_2^2 \\
&= 2n.
\end{aligned}$$

Next, we expand the double-sum term:

$$\begin{aligned}
&\sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Tr}\left((\mathcal{B}_i^{\text{vicreg}})^\top \mathcal{B}_j^{\text{vicreg}} \right) \\
&= \sum_{i,j} \Big[\mathrm{Tr}\big(f(x_i)\, f(x_i)^\top\, f(x_j)\, f(x_j)^\top \big) \;+\; \mathrm{Tr}\big(f(x_i)\, f(x_i)^\top\, f(x_j^+)\, f(x_j^+)^\top \big) \\
&\quad +\; \mathrm{Tr}\big(f(x_i^+)\, f(x_i^+)^\top\, f(x_j)\, f(x_j)^\top \big) \;+\; \mathrm{Tr}\big(f(x_i^+)\, f(x_i^+)^\top\, f(x_j^+)\, f(x_j^+)^\top \big) \Big] \\
&= \sum_{i,j} \Big[\big\langle f(x_i),\, f(x_j) \big\rangle^2 \;+\; \big\langle f(x_i),\, f(x_j^+) \big\rangle^2 \;+\; \big\langle f(x_i^+),\, f(x_j) \big\rangle^2 \;+\; \big\langle f(x_i^+),\, f(x_j^+) \big\rangle^2 \Big] \\
&= \sum_{i,j} \Big[\big\langle f(x_i),\, f(x_j) \big\rangle^2 \;+\; 2\big\langle f(x_i),\, f(x_j^+) \big\rangle^2 \;+\; \big\langle f(x_i^+),\, f(x_j^+) \big\rangle^2 \Big].
\end{aligned}$$

Combining all terms for the contrastive part, we obtain:

$$\mathcal{L}_{\text{ctr}}(f) = \frac{1}{4n^2} \sum_{i,j} \Big[\big\langle f(x_i),\, f(x_j) \big\rangle^2 \;+\; 2\big\langle f(x_i),\, f(x_j^+) \big\rangle^2 \;+\; \big\langle f(x_i^+),\, f(x_j^+) \big\rangle^2 \Big] - \frac{1}{d},$$

since $\gamma = \frac{1}{d}$ for VICReg. Next, we expand the invariance term:

$$\begin{aligned}
\mathcal{L}_{\text{inv}}(f) &= \frac{1}{n} \sum_{i=1}^{n} \left\| f(x_i) - f(x_i^+) \right\|_2^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\|f(x_i)\|^2 + \|f(x_i^+)\|^2 - 2 f(x_i)^\top f(x_i^+) \right) \\
&= 2 - \frac{2}{n} \sum_{i=1}^{n} f(x_i)^\top f(x_i^+). \qquad (12)
\end{aligned}$$

We combine all parts and obtain

$$\begin{aligned}\mathcal{L}_{\text{VICreg}}(f) &= -\frac{2\beta}{n}\sum_{i=1}^{n} f(x_i)^\top f(x_i^+) \\ &\qquad + \frac{\lambda}{4n^2}\sum_{i,j}\Big[\big\langle f(x_i),\, f(x_j)\big\rangle^2 + 2\big\langle f(x_i),\, f(x_j^+)\big\rangle^2 + \big\langle f(x_i^+),\, f(x_j^+)\big\rangle^2\Big] + 2\beta - \frac{\lambda}{d}.\end{aligned}$$

Finally, we remove terms where $i = j$, and adjust the normalization to obtain an unbiased estimate, thereby arriving at the form stated in Proposition 4:

$$\begin{aligned}\mathcal{L}_{\text{VICreg}}(f) = \frac{\lambda}{4n(n-1)}&\sum_{i\neq j}\Big[\big\langle f(x_i),\, f(x_j)\big\rangle^2 \;+\; 2\big\langle f(x_i),\, f(x_j^+)\big\rangle^2 \;+\; \big\langle f(x_i^+),\, f(x_j^+)\big\rangle^2\Big] \\ &- \frac{2\beta}{n}\sum_{i=1}^{n} f(x_i)^\top f(x_i^+) + 2\beta - \frac{\lambda}{d}.\end{aligned}$$

Barlow Twins. Again, we start from Equation 11 and expand the terms for $\mathcal{B}_i^{\text{bt}}$. We begin with the single-sum term:

$$\begin{aligned}\sum_{i=1}^{n}\operatorname{Tr}(\mathcal{B}_i^{\text{bt}}) &= \sum_{i=1}^{n}\operatorname{Tr}\left(f(x_i)f(x_i^+)^\top + f(x_i^+)f(x_i)^\top\right) \\ &= \sum_{i=1}^{n}\left(\operatorname{Tr}\left(f(x_i^+)^\top f(x_i)\right) + \operatorname{Tr}\left(f(x_i)^\top f(x_i^+)\right)\right) \\ &= 2\sum_{i=1}^{n} f(x_i)^\top f(x_i^+).\end{aligned}$$

Next, we expand the double-sum term:

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{Tr}\left((\mathcal{B}_i^{\text{bt}})^\top\mathcal{B}_j^{\text{bt}}\right) \\ &\quad= \sum_{i,j}\Big[\operatorname{Tr}\big(f(x_i)\,f(x_i^+)^\top\,f(x_j^+)\,f(x_j)^\top\big) \;+\; \operatorname{Tr}\big(f(x_i)\,f(x_i^+)^\top\,f(x_j)\,f(x_j^+)^\top\big) \\ &\qquad+\; \operatorname{Tr}\big(f(x_i^+)\,f(x_i)^\top\,f(x_j^+)\,f(x_j)^\top\big) \;+\; \operatorname{Tr}\big(f(x_i^+)\,f(x_i)^\top\,f(x_j)\,f(x_j^+)^\top\big)\Big] \\ &\quad= \sum_{i,j}\Big[\big\langle f(x_i^+),\, f(x_j^+)\big\rangle\big\langle f(x_j),\, f(x_i)\big\rangle \;+\; \big\langle f(x_i^+),\, f(x_j)\big\rangle\big\langle f(x_j^+),\, f(x_i)\big\rangle \\ &\qquad+\; \big\langle f(x_i),\, f(x_j^+)\big\rangle\big\langle f(x_j),\, f(x_i^+)\big\rangle \;+\; \big\langle f(x_i),\, f(x_j)\big\rangle\big\langle f(x_j^+),\, f(x_i^+)\big\rangle\Big] \\ &\quad= 2\sum_{i,j}\Big[\big\langle f(x_i^+),\, f(x_j^+)\big\rangle\big\langle f(x_j),\, f(x_i)\big\rangle \;+\; \big\langle f(x_i^+),\, f(x_j)\big\rangle\big\langle f(x_j^+),\, f(x_i)\big\rangle\Big].\end{aligned}$$

Combining all terms, we obtain:

$$\mathcal{L}_{\mathrm{BT}}(f) = \frac{1}{2n^2}\sum_{i,j}\Big[\big\langle f(x_i^+),\, f(x_j^+)\big\rangle\big\langle f(x_j),\, f(x_i)\big\rangle \;+\; \big\langle f(x_i^+),\, f(x_j)\big\rangle\big\langle f(x_j^+),\, f(x_i)\big\rangle\Big]$$
$$-\;\frac{2\beta}{n}\sum_{i=1}^{n} f(x_i)^\top f(x_i^+) + \beta^2 d.$$

Finally, we remove terms where $i = j$ and adjust the normalization to obtain an unbiased estimate and hence arrive also for BT at the form stated in Proposition 4:

$$\mathcal{L}_{\mathrm{BT}}(f) = \frac{1}{2n(n-1)}\sum_{i\neq j}\Big[\big\langle f(x_i^+),\, f(x_j^+)\big\rangle\big\langle f(x_j),\, f(x_i)\big\rangle \;+\; \big\langle f(x_i^+),\, f(x_j)\big\rangle\big\langle f(x_j^+),\, f(x_i)\big\rangle\Big]$$
$$-\;\frac{2\beta}{n}\sum_{i=1}^{n} f(x_i)^\top f(x_i^+) + \beta^2 d.$$

Appendix C. Unification at the population level

In the following, we state the unification of the pretext losses in their population form.

Proposition 10 (Unification of population-level pretext loss functions) *Let β and λ be hyperparameters. Then, each method $m \in \{\textsc{vicreg},\ \textsc{bt},\ \textsc{scl}\}$, defined in Appendix A, can be expressed, up to an additive constant c_m, as a weighted combination of an invariance term $\mathcal{L}_{inv}$ and a contrastive term $\mathcal{L}_{ctr}$, scaled by β and λ, respectively:*

$$\mathcal{L}_m(f) = \beta\mathcal{L}_{inv}(f) + \lambda\mathcal{L}_{ctr}(f) + c_m, \tag{13}$$

with $\mathcal{L}_{inv}(f) = -2\mathbb{E}_{(x,x^+)\sim\mathcal{S}}\left[f(x)^\top f(x^+)\right]$, $\mathcal{L}_{ctr}(f) = \mathbb{E}_{\bar{x},\bar{x}'\sim\mathcal{D}_\mathcal{X}}\left[\phi\left(\bar{x},\bar{x}'\right)\right]$, where $(\bar{x},\bar{x}')$ is a pair of independently sampled anchor data points, and

$$SCL : \phi\left(\bar{x},\bar{x}'\right) = \mathbb{E}_{x\sim\mathcal{A}(\cdot|\bar{x}),x'\sim\mathcal{A}(\cdot|\bar{x}')}\left[\left(f(x)^\top f(x')\right)^2\right], \beta = 1, c_{SCL} = 2 - \frac{\lambda}{d},$$
$$BT : \phi\left(\bar{x},\bar{x}'\right) = \mathbb{E}_{x,x^+\sim\mathcal{A}(\cdot|\bar{x}),\,x',x'^+\sim\mathcal{A}(\cdot|\bar{x}')}\left[f(x^+)^\top f(x')f(x'^+)^\top f(x)\right], \lambda = 1, c_{BT} = \beta^2 d,$$
$$VICReg : \phi\left(\bar{x},\bar{x}'\right) = \mathbb{E}_{x\sim\mathcal{A}(\cdot|\bar{x}),x'\sim\mathcal{A}(\cdot|\bar{x}')}\left[\left(f(x)^\top f(x')\right)^2\right], c_{VICReg} = 2\beta - \frac{\lambda}{d}.$$

Proof This directly follows from the definitions of the loss functions in Appendix A and the unification proof of Proposition 4. ■

Appendix D. Proof of SCL as a Special Case of VICReg

As pointed out by Cabannes et al. (2023), the population version of SCL can be recovered from the theory-friendly VICReg by setting $\beta = 1$. We illustrate this again below and motivate our modified definition of SCL, which ensures the loss remains non-negative. We expand VICReg as in the previous section, but this time at the population level. We begin by expanding the first term in Equation 5:

$$\lambda \left\| \mathbb{E}_x[f(x)f(x)^\top] - \frac{1}{d}I \right\|_F^2 = \lambda \mathbb{E}_{x,x'}\left[\left(f(x)^\top f(x')\right)^2\right] - 2\frac{\lambda}{d}\mathbb{E}_x\left[\|f(x)\|_2^2\right] + \frac{\lambda}{d}.$$

Next, we expand the second part:

$$\beta \mathbb{E}_{x,x^+}\left[\left\|f(x) - f(x^+)\right\|_2^2\right] = -2\beta \mathbb{E}_{x,x^+}\left[f(x)^\top f(x^+)\right] + 2\beta \mathbb{E}_x\left[\|f(x)\|_2^2\right].$$

Combining both parts, we obtain

$$\begin{aligned}\mathcal{L}_{\text{VICReg}}(f) = -2\beta \mathbb{E}_{x,x^+}\left[f(x)^\top f(x^+)\right] + \lambda \mathbb{E}_{x,x'}\left[\left(f(x)^\top f(x')\right)^2\right] \\ + 2\left(\beta - \frac{\lambda}{d}\right)\mathbb{E}_x\left[\|f(x)\|^2\right] + \frac{\lambda}{d}.\end{aligned}$$

Finally, setting $\beta = 1$ yields the form of the Spectral Contrastive loss:

$$\mathcal{L}_{\text{SC}}(f) = -2\mathbb{E}_{x,x^+}\left[f(x)^\top f(x^+)\right] + \lambda \mathbb{E}_{x,x'}\left[\left(f(x)^\top f(x')\right)^2\right] + 2 - \frac{\lambda}{d}. \tag{14}$$

Notably, the equality between the losses holds even without feature normalization for $\lambda = d$.

Appendix E. Proof of Corollary 5

In the following, we provide the proof of Corollary 5. We begin by proving that the bounded differences property (Definition 2) holds for the losses in Proposition 4 under exchange of anchor samples.

Lemma 11 (Bounded differences property for the losses of Proposition 4) *We consider the loss $\hat{\mathcal{L}}_m$ from Equation 3, computed over B equally sized batches, each containing r i.i.d. positive pairs, forming a random partition of S. Let $\bar{S} = \{\bar{x}_i\}_{i=1}^n$ denote the set of anchors sampled from $\mathcal{D}_{\mathcal{X}}$, and let $\ell_m : \mathcal{X}^n \to \mathbb{R}$ be defined so that the loss can be expressed as a function of the anchor samples, $\hat{\mathcal{L}}_m(f) = \ell_m(\bar{x}_1, \ldots, \bar{x}_k, \ldots, \bar{x}_n)$. Then, ℓ_m satisfies the bounded differences property with $c_k = \frac{C_m}{n}$, where $C_m = 4(\beta + \frac{\lambda}{2})$.*

Proof
We first state $\hat{\mathcal{L}}_m$ computed over B batches:

$$\hat{\mathcal{L}}_m(f) = \frac{1}{B}\sum_{b=1}^{B}\left(\beta\mathcal{L}^{(b)}_{\text{inv}}(f) + \lambda\mathcal{L}^{(b)}_{\text{ctr}}(f)\right) + c,$$

where $\mathcal{L}^{(b)}_{\text{inv}}$ and $\mathcal{L}^{(b)}_{\text{ctr}}$ are computed over batch b. We consider the exchanged sample x_k to be in the first batch, and thus fix $k \in [r]$ to bound the difference $\Delta\ell_m(\bar{x}_k) = \ell_m(\bar{x}_1, \dots, \bar{x}_k, \dots, \bar{x}_n) - \ell_m(\bar{x}_1, \dots, \bar{x}'_k, \dots, \bar{x}_n)$, without loss of generality. Moreover, we extend this definition to functions $\ell_{\text{inv}} : \mathcal{X}^n \to \mathbb{R}$ and $\ell_{\text{ctr}} : \mathcal{X}^n \to \mathbb{R}$, representing the invariance and contrastive terms over anchor samples, respectively, and obtain:

$$|\Delta\ell_m(\bar{x}_k)| \leq \frac{1}{n}\left[\beta\left|\Delta\ell^{(1)}_{\text{inv}}(\bar{x}_k)\right| + \lambda\left|\Delta\ell^{(1)}_{\text{ctr}}(\bar{x}_k)\right|\right]. \tag{15}$$

We start with bounding the difference on the invariance term:

$$\left|\Delta\ell^{(1)}_{\text{inv}}(\bar{x}_k)\right| \overset{\text{(a)}}{=} \left|-2\langle f(x_k), f(x_k^+)\rangle + 2\langle f(x'_k), f({x'}_k^{+})\rangle\right| \overset{\text{(b)}}{\leq} 4, \tag{16}$$

where (a) follows since terms that do not involve $\bar{x}_k$ cancel out and (b) follows from the normalization of embeddings $\|f(x)\|_2 = 1$ from which we obtain $\forall i, j, -1 \leq \langle f(x_i), f(x_j)\rangle \leq 1$. Next, we bound $\Delta\ell^{(1)}_{\text{ctr}}(\bar{x}_k)$. Note that, since the bound $\langle f(x_i), f(x_j)\rangle \leq 1$ holds for any augmented samples x_i, x_j, and since we consider exchange at the level of anchor samples $\bar{x}_k$, generating the positive pairs $(x_k, x_k^+) \sim \mathcal{A}(\cdot \mid \bar{x}_k)$, the terms $t_{i,j}$ in Proposition 4 are indistinguishable from a bounded differences perspective. Hence, we can provide the proof for all losses together, considering a non-symmetrized, general form for the contrastive term: $\mathcal{L}^{(1)}_{\text{ctr}}(f) = \frac{1}{r(r-1)}\sum_{i\neq j}\left(f(x_i)^\top f(x_j)\right)^2$. We get that:

$$\begin{aligned}|\Delta\ell^{(1)}_{\text{ctr}}(\bar{x}_k)| = \frac{1}{r-1}\Bigg|&\left(\sum_{j=1,j\neq k}^{r}\left(f(x_k)^\top f(x_j)\right)^2 + \sum_{i=1,i\neq k}^{r}\left(f(x_i)^\top f(x_k)\right)^2\right)\\ &-\left(\sum_{j=1,j\neq k}^{r}\left(f(x'_k)^\top f(x_j)\right)^2) + \sum_{i=1,i\neq k}^{r}\left(f(x_i)^\top f(x'_k)\right)^2\right)\Bigg| \overset{\text{(a)}}{\leq} 2,\end{aligned} \tag{17}$$

where (a) follows from $\forall i, j, 0 \leq \langle f(x_i), f(x_j)\rangle^2 \leq 1$. Using the bounds from Equation 16 and 17 in Equation 15, we obtain $c_k = \frac{4(\beta + \frac{\lambda}{2})}{n}$. ■

Next, we follow a modified version of the proof from McAllester (1999), as presented in van Elst and Ghoshdastidar (2025). We simply restate Lemma 3.5 therein and combine their Lemmas 3.6 and 3.7 in Lemma 13.

Lemma 12 (Lemma 3.5, van Elst and Ghoshdastidar, 2025) *Let X be a real valued random variable. If for $n \in \mathbb{N}^*$ and for $x > 0$, $\mathbb{P}(|X| \geq x) \leq 2e^{-nx^2}$, then*

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

Lemma 13 (Lemma 3.6-3.7, van Elst and Ghoshdastidar, 2025) *Let* $h : x \mapsto \frac{2x^2}{C_m^2}$ *and fix* $f \sim P$*. Then, we have with probability at least* $1-\delta$ *over i.i.d. dataset S:*

$$\forall Q, \quad h\left(\hat{\mathcal{L}}_m(Q) - \mathcal{L}_m(Q)\right) \leq \frac{\mathrm{KL}(Q \parallel P) + \log(\frac{2n}{\delta})}{n-1}.$$

Proof We start by deriving the concentration guarantee for a fixed predictor f. Using the bounded difference property from Lemma 11, we can apply McDiarmid's inequality and obtain for $\epsilon > 0$:

$$\begin{aligned}\mathbb{P}_S\left(\frac{\sqrt{2}}{C_m}\left|\mathcal{L}_m(f) - \hat{\mathcal{L}}_m(f)\right| \geq \epsilon\right) &= \mathbb{P}_S\left(\left|\mathcal{L}_m(f) - \hat{\mathcal{L}}_m(f)\right| \geq \frac{C_m}{\sqrt{2}}\epsilon\right) \\ &\leq 2\exp\left(-2\frac{\frac{C_m^2}{2}\epsilon^2}{\sum_{i=1}^n c_i^2}\right) \\ &\leq 2\exp\left(-n\epsilon^2\right).\end{aligned}$$

Observing that $\left(\frac{\sqrt{2}}{C_m}(\mathcal{L}_m(f) - \hat{\mathcal{L}}_m(f))\right)^2 = h(\mathcal{L}_m(f) - \hat{\mathcal{L}}_m(f))$, we apply Lemma 12 and obtain

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

Next, we integrate both sides w.r.t P

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

and change the order of integration due to Fubini

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

Next, Markov's inequality is applied and we obtain with probability at least $1-\delta$ over S:

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

Then, we apply the variational formula from Donsker and Varadhan (1976), plug in the previous bound, and obtain for all Q:

$$\begin{aligned}\mathbb{E}_{f\sim Q}\left[(n-1)h(\mathcal{L}_m(f) - \hat{\mathcal{L}}_m(f))\right] &\leq \mathrm{KL}(Q \parallel P) + \log \mathbb{E}_{f\sim P}\left[e^{(n-1)h(\mathcal{L}_m(f) - \hat{\mathcal{L}}_m(f))}\right] \\ &\leq \mathrm{KL}(Q \parallel P) + \log\frac{2n}{\delta}.\end{aligned}$$

Next, Jensen's inequality is applied, exploiting convexity of h:

$$\forall Q, \quad (n-1)h\left(\mathcal{L}_m(Q) - \hat{\mathcal{L}}_m(Q)\right) \leq \mathrm{KL}(Q \parallel P) + \log\frac{2n}{\delta}.$$

■

Finally, unfolding the definition of h and rearranging the terms, we obtain the statement of Corollary 5: with probability at least $1-\delta$ over i.i.d. dataset S,

$$\forall Q, \quad \mathcal{L}_m(Q) \leq \hat{\mathcal{L}}_m(Q) + C_m\sqrt{\frac{\mathrm{KL}(Q \parallel P) + \log\frac{2n}{\delta}}{2(n-1)}}.$$

Appendix F. Proof of Lemma 6

In the following, we present the proof of Lemma 6. Let $a = f(x^+)^\top f(x')$ and $b = f(x'^+)^\top f(x)$. Then, we obtain:

$$\begin{aligned}
\mathbb{E}_{(x,x^+),\,(x',x'^+)\sim\mathcal{S}^2}\left[ab\right] &= \frac{1}{2}\mathbb{E}_{x,x^+,x',x'^+}\left[2ab\right] \\
&= \frac{1}{2}\mathbb{E}_{x,x^+,x',x'^+}\left[a^2 + b^2 - (a-b)^2\right] \\
&= \frac{1}{2}\mathbb{E}_{x,x^+,x',x'^+}\left[a^2 + b^2\right] - \frac{1}{2}\mathbb{E}_{x,x^+,x',x'^+}\left[(a-b)^2\right] \\
&= \frac{1}{2}(\mathbb{E}_{x^+,x'}\left[a^2\right] + \mathbb{E}_{x,x'^+}\left[b^2\right]) - \frac{1}{2}\mathbb{E}_{x,x^+,x',x'^+}\left[(a-b)^2\right] \\
&= \mathbb{E}_{x,x'}\left[a^2\right] - \frac{1}{2}\mathbb{E}_{x,x^+,x',x'^+}\left[(a-b)^2\right].
\end{aligned}$$

Plugging in the definition of ζ completes the proof.

Remark 14 *Note that due to feature normalization, we can bound the expectation as* $\mathbb{E}_{(x,x^+),\,(x',x'^+)\sim\mathcal{S}^2}\left[\left(f(x^+)^\top f(x') - f(x'^+)^\top f(x)\right)^2\right] \leq 4$ *and hence we could employ the bound* $\phi_{BT}(\bar{x},\bar{x}') \geq \mathbb{E}_{x,x'}\left[\left(f(x)^\top f(x')\right)^2\right] - 2$*. However, this bound is usually pessimistic in practice, since the loss function implicitly enforces ζ to be small, due to the alignment of augmentations of the same instance. Therefore, we empirically estimate ζ from the training data.*

Appendix G. Proof of Theorem 7

In the following, we present the proof of Theorem 7. The proof is carried out using the population pretext losses $\mathcal{L}_m$ as defined in Proposition 10, with the contrastive term given by $\mathcal{L}_{\text{ctr}} = \mathbb{E}_{x,x'}\left[\left(f(x)^\top f(x')\right)^2\right]$. For Barlow Twins, we apply the reformulation from Lemma 6 to align it with this unified form. Moreover, we state the batchwise form of the population loss, computed over batches of size r:

$$\begin{aligned}
\mathcal{L}_m(f) = -2\beta\mathbb{E}_{\{x_i,x_i^+\}_{i=1}^r}&\left[\frac{1}{r}\sum_{i=1}^r f(x_i)^\top f(x_i^+)\right] \\
&+ \lambda\mathbb{E}_{\{x_i,x_i^+\}_{i=1}^r}\left[\frac{1}{r(r-1)}\sum_{i\neq j}\left(f(x_i)^\top f(x_j^+)\right)^2\right] + c_m. \qquad (18)
\end{aligned}$$

We note that, due to the linearity of expectation, the i.i.d. nature of the pairs (x_i, x_i^+) and the independence of x_i, x_j^+ for $i \neq j$, we obtain the form stated in Equation 13 by moving the summations outside the expectation. We therefore proceed with the formulation from Equation 13.

Next, we separately provide lower bounds for $\mathcal{L}_{\text{ctr}} = \mathbb{E}_{x,x'}\left[(f(x)^\top f(x'))^2\right]$ and $\mathcal{L}_{\text{inv}} = -\mathbb{E}_{x,x^+}\left[f(x)^\top f(x^+)\right]$. We start with the lower bound on $\mathcal{L}_{\text{Inv}}$. To this end, we refine a bound introduced by Wang et al. (2022) in Theorem A.3. Let $\mathbf{u} = f(x)$ and $\mathbf{v} = f(x^+) - \mu_y$. Moreover, let $\mathbf{u}$ and $\mathbf{v}$ be vectors of the same dimension, and let ϑ denote the angle between them, provided both are nonzero. Then, we define the similarity between $\mathbf{u}$ and $\mathbf{v}$ as

$$\mathrm{sim}(\mathbf{u},\mathbf{v}) = \begin{cases} \cos\vartheta, & \text{if } \|\mathbf{u}\|_2 \neq 0 \text{ and } \|\mathbf{v}\|_2 \neq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

Based on these definitions, we derive:

$$\begin{aligned}
\mathcal{L}_{\text{Inv}} &= -\mathbb{E}_{x,x^+}\left[f(x)^\top f(x^+)\right] \\
&\overset{\text{(a)}}{=} -\mathbb{E}_{x,x^+,y}\left[f(x)^\top(\mu_y + f(x^+) - \mu_y)\right] \\
&= -\mathbb{E}_{x,x^+,y}\left[f(x)^\top \mu_y + \mathbf{u}^\top\mathbf{v}\right] \\
&\geq -\mathbb{E}_{x,x^+,y}\left[f(x)^\top \mu_y + \left|\mathbf{u}^\top\mathbf{v}\right|\right] \\
&\overset{\text{(b)}}{=} -\mathbb{E}_{x,x^+,y}\left[f(x)^\top \mu_y + \|\mathbf{u}\|_2\|\mathbf{v}\|_2\ |\mathrm{sim}(\mathbf{u},\mathbf{v})|\right] \\
&\overset{\text{(c)}}{=} -\mathbb{E}_{(x,y)}\left[f(x)^\top \mu_y\right] - \mathbb{E}_{x,x^+,y}\left[\|\mathbf{v}\|_2\ |\mathrm{sim}(\mathbf{u},\mathbf{v})|\right] \\
&\overset{\text{(d)}}{=} -\sigma' - \mathbb{E}_{(x,y)}\left[f(x)^\top \mu_y\right],
\end{aligned}$$

where (a) introduces the mean classifier assuming that y is the label of the augmented pair (x, x^+), (b) uses the geometric definition of the dot product in Euclidean space as well as the definition of $\mathrm{sim}(\mathbf{u},\mathbf{v})$ to ensure that ϑ is well-defined, (c) uses linearity of expectation and feature normalization $\|\mathbf{u}\|_2 = 1$ and (d) uses the definition of σ' as the cosine-scaled intra-class feature deviation.

Remark 15 *Wang et al. (2022) use the Cauchy-Schwarz inequality in place of Equation (b), leading to the intra-class feature deviation $\sigma = \mathbb{E}_{(x,y)}\left[\|f(x) - \mu_y)\|_2\right]$, which is subsequently employed by van Elst and Ghoshdastidar (2025) in the transfer bound. Wang et al. (2022) further derive an upper bound on σ, ultimately using $Var(f(x)|y) = \sqrt{\mathbb{E}_{(x,y)}\left[\|f(x) - \mu_y)\|_2^2\right]}$ in their bound. Hence, the following relation holds: $\sigma' \leq \sigma \leq Var(f(x)|y)$. These quantities can be interpreted as measures of the inverse clustering strength of the augmentation pipeline: lower intra-class feature deviations correspond to stronger clustering. While this may not provide the most accurate quantification of the augmentation pipeline's clustering strength, it follows directly from the pretext losses, enabling its use in the transfer bound. For an alternative metric that is not directly tied to the pretext loss, we refer the reader to Wang et al. (2022). Empirically, we observe that σ yields overly pessimistic bounds when a strong augmentation pipeline, including random cropping, is employed, motivating our refinement. Beyond the derivation above, which highlights the connection to prior work, we can directly state our refined term as $\sigma' = \mathbb{E}_{x,x^+,y}\left[\left|f(x)^\top(f(x^+) - \mu_y)\right|\right]$, thereby avoiding the introduction of the cosine-based similarity measure.*

Next, we establish a lower bound for the contrastive term, following the approach of Arora et al. (2019), which introduces the mean classifier via Jensen's inequality:

$$\begin{aligned}\mathbb{E}_{x,x'}\left[(f(x)^\top f(x'))^2\right] &\overset{\text{(a)}}{=} \mathbb{E}_{(x,y)}\mathbb{E}_{(x',y')}\left[\left(f(x)^\top f(x')\right)^2\right] \\ &\overset{\text{(b)}}{\geq} \mathbb{E}_{(x,y)}\mathbb{E}_{y'}\left[\left(f(x)^\top \mathbb{E}_{x'\sim\mathcal{D}_{y'}}\left[f(x')\right]\right)^2\right] \\ &\overset{\text{(c)}}{=} \mathbb{E}_{(x,y)}\left[\sum_{c\in[C]}\frac{1}{C}\left(f(x)^\top\mu_c\right)^2\right],\end{aligned} \tag{20}$$

where (a) uses independence of samples x, x', (b) uses Jensen's inequality due to convexity of $u \mapsto (a^\top u)^2$ and (c) uses the uniform class prior assumption. Next, we combine both parts:

$$\begin{aligned}\mathcal{L}_m(f) &\overset{\text{(a)}}{\geq} -\gamma\left(\sigma' + \mathbb{E}_{(x,y)}\left[f(x)^\top\mu_y\right]\right) + \lambda\mathbb{E}_{(x,y)}\left[\sum_{c\in[C]}\frac{1}{C}\left(f(x)^\top\mu_c\right)^2\right] + c_m - \mathbf{1}_{\{m=\text{BT}\}}\zeta \\ &= -\gamma\sigma' + \mathbb{E}_{(x,y)}\left[-\gamma\left(f(x)^\top\mu_y\right) + \frac{\lambda}{C}\sum_{c\in[C]}\left(f(x)^\top\mu_c\right)^2\right] + c_m - \mathbf{1}_{\{m=\text{BT}\}}\zeta \\ &\overset{\text{(b)}}{=} -\gamma\sigma' + \mathcal{L}_{\text{SupJE}}(f, W^\mu) - \mathbf{1}_{\{m=\text{BT}\}}\zeta,\end{aligned}$$

where (a) combines the bounds on $\mathcal{L}_{\text{ctr}}$ and $\mathcal{L}_{\text{inv}}$, uses the definition of $\mathcal{L}_m$ and introduces the additional term for Barlow Twins, ζ, as defined in Lemma 6, (b) uses the definition $\gamma = 2\beta$ and the definition of $\mathcal{L}_{\text{SupJE}}$. Finally, we take the infimum over all linear classifiers and obtain

$$\inf_{W\in\mathbb{R}^{C\times d}}\mathcal{L}_{\text{SupJE}}(f, W) \leq \mathcal{L}_m(f) + \gamma\sigma' + \mathbf{1}_{\{m=\text{BT}\}}\zeta.$$

Appendix H. Proof of Corollary 8

We begin by combining the bounds on the invariance and contrastive terms established in the proof of Theorem 7:

$$\begin{aligned}\mathcal{L}_m(f) &\overset{\text{(a)}}{\geq} -\gamma\left(\sigma' + \mathbb{E}_{(x,y)}\left[f(x)^\top\mu_y\right]\right) + \lambda\mathbb{E}_{(x,y)}\left[\sum_{c\in[C]}\frac{1}{C}\left(f(x)^\top\mu_c\right)^2\right] + \beta - \Delta_m \\ &= -\gamma\sigma' + \mathbb{E}_{(x,y)}\left[-\gamma\left(f(x)^\top\mu_y\right) + \frac{\lambda}{C}\sum_{c\in[C]}\left(f(x)^\top\mu_c\right)^2\right] + \beta - \Delta_m \\ &\overset{\text{(b)}}{=} -\gamma\sigma' + \mathbb{E}_{(x,y)}\left[\frac{\lambda}{C}\sum_{c\in[C]}\left(f(x)^\top\mu_c\right)^2 - \gamma\left(f(x)^\top\mu_y\right) + \beta\right] - \Delta_m \\ &\overset{\text{(c)}}{=} -\gamma\sigma' + \mathbb{E}_{(x,y)}\left[\frac{\lambda}{C}\left\|W^\mu f(x)\right\|_2^2 - 2\beta\left(f(x)^\top\mu_y\right) + \beta\left\|\boldsymbol{y}(x)\right\|_2^2\right] - \Delta_m \\ &\overset{\text{(d)}}{=} -\gamma\sigma' + \beta\mathbb{E}_{(x,y)}\left[\left\|W^\mu f(x) - \boldsymbol{y}(x)\right\|_2^2\right] - \Delta_m \\ &\overset{\text{(e)}}{=} -\gamma\sigma' + \beta\mathcal{L}_{\text{SL}}(f, W^\mu) - \Delta_m,\end{aligned}$$

where (a) combines the bounds on $\mathcal{L}_{\text{ctr}}$ and $\mathcal{L}_{\text{inv}}$ and uses the definition of $\mathcal{L}_m$ and Δ_m, (b) moves one factor of β inside the expectation, (c) uses the definition $\gamma = 2\beta$, the definition of the mean classifier and introduces the squared ℓ_2-norm of the one-hot encoding of the label y, (d) uses $\frac{\lambda}{C} = \beta$ and (e) applies the definition of $\mathcal{L}_{\text{SL}}$. Finally, we take the infimum over all linear classifiers and obtain

$$\inf_{W \in \mathbb{R}^{C \times d}} \mathcal{L}_{\text{SL}}(f, W) \leq \frac{\mathcal{L}_m(f) + \gamma\sigma' + \Delta_m}{\beta}.$$

Appendix I. Proof of Remark 9

In the following, we introduce a balanced version of SL and adapt the proof of Corollary 8. The weights of the balanced loss function can either be estimated from training data or derived from π. Here, we consider the case where the weights are estimated from the training data. Let $\mathcal{L}_{\text{BalSL}} = \mathbb{E}_{(x,y)} \left[b_y \| W^\mu f(x) - \boldsymbol{y}(x) \|_2^2 \right]$, where the class-specific weight is given by $b_y = \frac{\omega_y}{\sum_{c \in [C]} \omega_c}$ with $\omega_y = \frac{N}{N_y C}$ and where N denotes the total number of training samples, and N_y the number of samples belonging to class y. Moreover, we assume $\mathbb{P}(Y = c) > 0$ for all $c \in [C]$. We begin by restating the bound on the contrastive term (20):

$$\begin{aligned} \mathbb{E}_{x,x'} \left[\left(f(x)^\top f(x') \right)^2 \right] &\geq \mathbb{E}_{(x,y)} \mathbb{E}_{y'} \left[\left(f(x)^\top \mathbb{E}_{x'|y'} \left[f(x') \right] \right)^2 \right] \\ &= \mathbb{E}_{(x,y)} \left[\sum_{c \in [C]} \mathbb{P}(Y = c) \left(f(x)^\top \mu_c \right)^2 \right]. \end{aligned}$$

Again, we combine bounds on the invariance and contrastive terms:

$$\begin{aligned} \mathcal{L}_m(f) &\geq -\gamma\sigma' + \mathbb{E}_{(x,y)} \left[-\gamma \left(f(x)^\top \mu_y \right) + \lambda \sum_{c \in C} \mathbb{P}(Y = c) \left(f(x)^\top \mu_c \right)^2 \right] + \beta - \Delta_m \\ &\geq -\gamma\sigma' + \mathbb{E}_{(x,y)} \left[-\gamma \left(f(x)^\top \mu_y \right) + \lambda \sum_{c \in C} \min_{c \in [C]} \mathbb{P}(Y = c) \left(f(x)^\top \mu_c \right)^2 \right] + \beta - \Delta_m \\ &\overset{\text{(a)}}{=} -\gamma\sigma' + \mathbb{E}_{(x,y)} \left[\lambda \, p_{min} \sum_{c \in [C]} \left(f(x)^\top \mu_c \right)^2 - 2\beta \left(f(x)^\top \mu_y \right) + \beta \right] - \Delta_m \\ &\overset{\text{(b)}}{=} -\gamma\sigma' + \beta \mathbb{E}_{(x,y)} \left[\| W^\mu f(x) - \boldsymbol{y}(x) \|_2^2 \right] - \Delta_m \\ &\overset{\text{(c)}}{\geq} -\gamma\sigma' + \beta \mathbb{E}_{(x,y)} \left[b_y \, \| W^\mu f(x) - \boldsymbol{y}(x) \|_2^2 \right] - \Delta_m \\ &\overset{\text{(d)}}{=} -\gamma\sigma' + \beta \mathcal{L}_{\text{BalSL}}(f, W^\mu) - \Delta_m, \end{aligned}$$

where (a) uses the definition $p_{min} = \min_{c \in [C]} \mathbb{P}(Y = c)$, which can be estimated from train data as well, and $\gamma = 2\beta$, (b) imposes the hyperparameter ratio $\lambda \, p_{min} = \beta$, (c) introduces b_y and (d) uses the definition of $\mathcal{L}_{BalSL}$. Again, the proof for $\mathcal{L}_{\text{BalSL}}$ follows by taking the infimum over all linear classifiers.

Appendix J. List of PAC-Bayes Bounds

In the following, we list the PAC-Bayes bounds that we used in Section 4 for comparison with Corollary 5. These bounds assume independence across the summands of $\hat{\mathcal{L}}(f)$. However, as discussed in Section 3.2, the summands are dependent within each batch. To address this, the bounds are not applied over n i.i.d. samples, but over i.i.d. batches. Let r denote the number of samples per batch, and B the dataset containing $p = \frac{n}{r}$ batches. Moreover, let $C = 2\beta + \lambda + c_m$ be the upper bound of the losses presented in Proposition 4, using the parameters defined therein. The PAC-Bayes bounds can then be applied over the p i.i.d. batches in B. We now follow the presentation of the modified bounds as given in van Elst and Ghoshdastidar (2025):

Proposition 16 *(Classic PAC-Bayes bound over i.i.d. batches, originally by McAllester, 1999 and adapted in van Elst and Ghoshdastidar, 2025) For any prior P over Θ, and any $\delta \in (0,1)$, with a probability of at least $1-\delta$ over size-p i.i.d. random batches B, simultaneously for all posterior distributions Q over Θ, the following inequality holds:*

$$\mathcal{L}(Q) \leq \hat{\mathcal{L}}(Q) + C\sqrt{r\frac{\mathrm{KL}(Q \parallel P) + \log(\frac{2\sqrt{n}}{\delta\sqrt{r}})}{2n}}.$$

Proposition 17 *(Catoni's PAC-Bayes bound over i.i.d. batches, originally by Catoni, 2007 and adapted in van Elst and Ghoshdastidar, 2025) For any prior P over Θ, and any $\delta \in (0,1)$, with a probability of at least $1-\delta$ over size-p i.i.d. random batches B, simultaneously for all posterior distributions Q over Θ, the following inequality holds:*

$$\frac{1}{C}\mathcal{L}(Q) \leq \inf_{\gamma>0}\left\{\frac{1-\exp\left(-\frac{\gamma}{C}\hat{\mathcal{L}}(Q) - r\frac{\mathrm{KL}(Q\|P)+\log(\frac{1}{\delta})}{n}\right)}{1-\exp(-\gamma)}\right\}.$$

Appendix K. Experimental Details

In the following, we provide additional details on the experimental setup and present the results on the MNIST dataset.

Computational resources. All experiments were conducted on the NVIDIA HGX H100 architecture, equipped with NVIDIA H100 GPUs. We use a single GPU and 40 GB of memory.

Data augmentation. We follow the augmentation pipeline described in van Elst and Ghoshdastidar (2025), applying random cropping, color jittering (with strength 0.5 and probability 0.8), color dropping and random horizontal flipping (with probability 0.5). Additionally, the augmented images are normalized channel-wise using the mean and standard deviation computed from the training dataset.

Training pipeline. The prior and posterior probabilistic networks are trained using the AdamW optimizer for 100 epochs with a cosine annealing learning rate schedule. The mean parameters μ_0 of the Gaussian prior distribution are initialized randomly from a truncated, centered Gaussian with standard deviation set to $\frac{1}{\sqrt{n_{\text{in}}}}$, where n_{in} denotes the input dimension of the respective layer. The distribution is truncated at ± 2 standard deviations. The standard deviation parameters σ_0 are initialized from $\{0.01, 0.05, 0.1\}$, following Perez-Ortiz et al. (2021b) and van Elst and Ghoshdastidar (2025). The posterior is always initialized at the prior. For the evaluation of the pretext generalization bounds, Monte Carlo sampling is employed to approximate the expectation over Q in $\hat{\mathcal{L}}_S(Q)$ using an average over $p = 100$ samples drawn i.i.d. from the posterior. For downstream linear classification, we consider 10 classes, and train for 20 epochs using the Adam optimizer and a learning rate of 0.01. Moreover, we empirically estimate the intra-class feature deviation σ' from Section 3.3 in the downstream bound by applying Monte Carlo sampling to the augmentation process. For the computation of the mean classifier, we use $p = 10$ samples, and for other sampling steps, we use $p = 5$. Additionally, the ζ term for the Barlow Twins loss, introduced in Lemma 6, is empirically estimated from the training data.

Model architectures. For the prior and posterior distribution, we use a 7-layer convolutional neural network, max-pooling at every second layer and a 2-layer MLP projection head. ReLU activations are used in the hidden layers. The embedding dimension is 256, while the representation dimension is 2048.

Hyperparameters. The parameter δ used in the PAC-Bayes objective is set to 0.04. We use a batch size of $r = 250$. For the pretext task, we select the optimizer and scheduler hyperparameters that yield the best risk certificates. Moreover, we experiment with different values for λ and β.

Experiments on MNIST. The MNIST dataset contains 60,000 train images and 10,000 test images (LeCun et al., 2010). The experimental setup largely follows that of CIFAR-10, with two main differences: a 3-layer convolutional neural network is used for both the prior and posterior, and the representation dimension is reduced to 512. We provide the results on pretext and downstream generalization bounds in Table 4, while Table 5 illustrates how the refined term σ' captures the effects of different augmentation pipelines.

		VICReg			SCL	BT	VICReg
		$\beta = 0.1, \lambda = 1$	$\beta = 0.5, \lambda = 5$	$\beta = 1, \lambda = 10$	$\beta = 1, \lambda = 1$	$\beta = 1, \lambda = 1$	$\beta = 1, \lambda = 1$
	Pretext test loss	0.080	0.395	0.790	0.395	260.899	0.394
	Classic bound (iid) (16)	0.422	0.298	4.783	1.699	328.249	1.698
	Catoni's bound (iid) (17)	0.255	1.423	3.386	1.174	271.7	1.172
	Cor. 5 (ours)	0.128	0.643	1.317	0.519	259.917	0.517
	KL/n	1.710×10^{-5}	8.290×10^{-5}	1.958×10^{-4}	7.620×10^{-5}	7.190×10^{-5}	7.660×10^{-5}
Th. 7	Risk certificate	0.233	1.175	2.378	1.055	260.462	1.065
	Test loss	0.185	0.927	1.852	0.931	261.444	0.942
Proj.	SupJE loss	0.104	0.522	1.050	0.446	254.598	0.463
	Top-1	0.963	0.961	0.958	0.846	0.771	0.844

Table 4: *Upper Part* — Comparison of PAC-Bayes risk certificates for pretext losses from Proposition 4 across different hyperparameter combinations to the pretext test loss as well as to other bounds established in the literature (see Appendix J) on MNIST. The Kullback–Leibler (KL) divergence between prior and posterior distributions is reported per dataset size n. *Lower Part* — Comparison of the transfer bound from Theorem 7 with the corresponding downstream test loss. Top-1 accuracy is reported to assess downstream performance. The first three columns correspond to hyperparameter combinations that satisfy the constraints of Corollary 8, whereas the last three columns present an exemplary deviating hyperparameter combination.

		$\beta = 0.1, \lambda = 1$		$\beta = 0.5, \lambda = 5$		$\beta = 1, \lambda = 10$	
	Cropping applied	×	✓	×	✓	×	✓
Th. 7	Risk certificate	0.2476	0.2332	1.2419	1.1746	2.4850	2.3781
	Test loss	0.1953	0.1848	0.9792	0.9265	1.9593	1.8516
Proj.	SupJE loss	0.1220	0.1044	0.6203	0.5224	1.2394	1.0501
	Top-1	0.8582	0.9628	0.8447	0.9613	0.8407	0.9579
	σ	0.9590	0.8536	0.9637	0.8571	0.9629	0.8570
	σ' (refined, ours)	0.9172	0.5242	0.9275	0.5313	0.9265	0.5307

Table 5: Comparison of downstream bounds for VICReg across different hyperparameter combinations computed once using the full augmentation pipeline and once without random cropping on MNIST. We report the intra-class feature deviation σ as provided in Theorem 3.15 of van Elst and Ghoshdastidar (2025) alongside our refined σ'.

Appendix L. Empirical Comparison of Cross-Entropy and SupJE loss

Table 6 shows that the SupJE loss performs on par with the standard cross-entropy loss, both under the SL-ratio and when deviating from this "ideal ratio".

		VICReg			SCL	BT	VICReg
	Top-1 (%)	$\beta = 0.1, \lambda = 1$	$\beta = 0.5, \lambda = 5$	$\beta = 1, \lambda = 10$	$\beta = 1, \lambda = 1$	$\beta = 1, \lambda = 1$	$\beta = 1, \lambda = 1$
Proj.	SupJE loss	69.10	69.18	70.28	45.81	44.22	50.73
	CE loss	69.42	69.39	70.11	47.68	47.94	51.49
	SupJE loss	71.96	72.60	73.84	62.92	65.34	66.64
	CE loss	72.21	72.95	73.96	63.37	65.70	67.31

Table 6: Comparison of Top-1 accuracy between the cross-entropy loss and the SupJE loss defined in Equation 4 across different hyperparameter combinations on CIFAR-10. Results are reported both with and without a projection head. The first three columns correspond to hyperparameter combinations that satisfy the constraints of Corollary 8, whereas the last three columns present an exemplary deviating hyperparameter combination.