MITIGATING DISTRIBUTIONAL BIASES IN CONTRASTIVE LEARNING

A PREPRINT

Lidia Troeshestova
Moscow Institute of Physics and Technology
troeshestova.ls@phystech.edu

Roman Isachenko

ABSTRACT

Recently there has been a renewed interest in using contrastive learning for self-supervised representation learning. A popular method for learning representations without labels is to compare similar (positive) and dissimilar (negative) pairs of samples. However, negative samples are often chosen randomly, which means that they could be of the same class. This paper analyzes various ways to eliminate these biases. Based on the fully-supervised case, we develop debiased contrastive models that account for same-label datapoints without requiring knowledge of true labels, and explore their properties. The experiments are performed on MNIST and CIFAR10 datasets. This will further improve availability of accurate models for classification in tasks where extensive labeling is expensive or inaccessible.

Keywords : Contrastive learning · Representation learning · Self-supervised learning

1 Introduction

Representation learning has become increasingly popular in recent years due to its ability to learn meaningful representations from large amounts of data, without the need for manual feature engineering. A widely used solution for this problem is contrastive learning, a technique that uses the principle of contrasting samples against each other to learn attributes that are common between data classes and attributes that set apart a data class from another. It encourages the representations of similar pairs $(\mathbf{x}, \mathbf{x}^+)$ to be close, and those of dissimilar pairs $(\mathbf{x}, \mathbf{x}^-)$ to be more orthogonal.

The basic metric for supervised similarity is *triplet loss* [Schroff et al., 2015], where $\mathbf{x}$ is called *anchor*, $\mathbf{x}^+$ is a sample with the same label as anchor, and $\mathbf{x}^-$ is a sample with the different label than anchor. An improvement of triplet loss is *Multi-class N-pair Loss* invented by [Sohn, 2016]:

$$\mathcal{L}(\mathbf{x}, \mathbf{x}^+, \{\mathbf{x}_i^-\}_{i=1}^{N-1}; f) = -\log \frac{\exp(f(\mathbf{x})^T f(\mathbf{x}_i^+))}{\exp(f(\mathbf{x})^T f(\mathbf{x}_i^+)) + \sum_{i=1}^{N-1} \exp(f(\mathbf{x})^T f(\mathbf{x}_i^-))} \tag{1}$$

Contrastive learning has proven to be effective for eliciting visual representations: SimCLR [Ting Chen, 2020] is a framework that considerably outperforms previous self-supervised and semi-supervised learning methods. Its authors developed NT-Xent loss (normalized temperature-scaled cross entropy loss), that uses cosine similarity instead of scalar products.

Due to the unavailability of actual labels during training, negative instances $\mathbf{x}^-$ are often randomly selected from the training data. However, this approach results in sampling bias, where $\mathbf{x}^-$ may actually be similar to $\mathbf{x}$. This bias causes a significant drop in performance.

To address this issue, [Prannay Khosla, 2021] extend the self-supervised batch contrastive approach to the fully-supervised setting, allowing to effectively leverage label information. According to the experiments, supervised contrastive loss outperforms the base cross entropy, but only by a small amount. It also outperforms the cross entropy on robustness benchmark (ImageNet-C, which applies common naturally occuring perturbations such as noise, blur and contrast changes to the ImageNet dataset) and is less sensitive to hyperparameter changes.

As for the case of self-supervised learning, a debiased contrastive loss was proposed to mitigate the bias of negative sampling by [Ching-Yao Chuang, 2020]. Given N negative and M positive samples, $\mathcal{L}^{N,M}_{\text{Debiased}}(f)$ is the loss function that corrects for False Negative samples. To remain unsupervised in practice, samples are acquired from the data distribution and a positive distribution mimicked by data augmentations, which can contatin False Positives.

This work aims to develop a new algorithm for debiased sampling in contrastive learning, which accounts for False Positive as well as False Negative samples.

2 Problem Statement

2.1 Debiased Contrastive Learning

In [Ching-Yao Chuang, 2020] the distribution over classes is denoted by $\rho(c)$, the joint distribution $p_{x,c}(\mathbf{x}, c) = p(\mathbf{x}|c)\rho(c)$. Let $h : X \to C$ be the function assigning the latent class labels. Then $p^+_x(\mathbf{x}') = p(\mathbf{x}'|h(\mathbf{x}') = h(\mathbf{x}))$ is the probability of observing $\mathbf{x}'$ as a positive example for $\mathbf{x}$ and $p^-_x(\mathbf{x}') = p(\mathbf{x}'|h(\mathbf{x}') \neq h(\mathbf{x}))$ the probability of a negative example. It is assumed that the class probabilities $\rho(\mathbf{x}) = \tau^+$ are uniform, and $\tau^- = 1 - \tau^+$ is the probability of observing any different class. Then the "ideal", or unbiased loss to optimize would be:

$$L^N_{\text{Unbiased}}(f) = \mathbb{E}_{\substack{\mathbf{x}\sim p, \mathbf{x}^+\sim p^+_x, \\ \mathbf{x}^-_i \sim p^-_x}} \left[-\log \frac{e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + \sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{x}^-_i)}} \right] \tag{2}$$

In practice, $p^-_x(\mathbf{x}^-_i)$ is unavailable. Thus, the standard solution is to sample negative examples $\mathbf{x}^-_i$ from the unlabeled $p(\mathbf{x})$ instead. When drawn from $p(\mathbf{x})$, the sample $\mathbf{x}^-_i$ will be of the same class as $\mathbf{x}$ with probability τ^+.

$$L^N_{\text{Biased}}(f) = \mathbb{E}_{\substack{\mathbf{x}\sim p, \mathbf{x}^+\sim p^+_x, \\ \mathbf{x}^-_i \sim p_x}} \left[-\log \frac{e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + \sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{x}^-_i)}} \right] \tag{3}$$

Problem statement: Derive a loss function that approximates the ideal L^N_{Unbiased}, only having access to M positive samples and samples from the distribution p.

The data distribution can be decomposed:

$$p(\mathbf{x}') = \tau^+ p^+_x(\mathbf{x}') + \tau^- p^-_x(\mathbf{x}') \tag{4}$$

And the missing $p^-_x(\mathbf{x}')$ is derived as

$$p^-_x(\mathbf{x}') = \frac{p(\mathbf{x}') - \tau^+ p^+_x(\mathbf{x}')}{\tau^-} \tag{5}$$

However, estimating $p(\mathbf{x}')$ and $p^+_x(\mathbf{x}')$ empirically is computationally expensive, and the authors propose a cheaper way to estimate $p^-_x(\mathbf{x}')$, using N samples $\{\mathbf{u}_i\}_{i=1}^N$ from p and M samples $\{\mathbf{v}_i\}_{i=1}^M$ from p^+_x, with temperature t:

$$L^{N,M}_{\text{DebiasedNeg}}(f) = \mathbb{E}_{\substack{\mathbf{x}\sim p; \mathbf{x}_+\sim p^+_x, \\ \{\mathbf{u}_i\}_{i=1}^N \sim p^N \\ \{\mathbf{v}_i\}_{i=1}^N \sim p^{+M}_x}} \left[-\log \frac{e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + Ng\big(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^N, \{\mathbf{v}_i\}_{i=1}^M\big)} \right], \tag{6}$$

where the empirical estimates of p and p^-_x are defined as

$$g\big(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^N, \{\mathbf{v}_i\}_{i=1}^M\big) = \max\left\{ \frac{1}{\tau^-}\left(\frac{1}{N}\sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{u}_i)} - \tau^+ \frac{1}{M}\sum_{i=1}^M e^{f(\mathbf{x})^T f(\mathbf{v}_i)} \right), e^{-1/t} \right\}. \tag{7}$$

Using this loss on SimCLR framework allowed to achieve higher accuracy compared to the regular biased loss. With a similar theoretical inference, this work is aimed at developing a new algorithm, that would additionally focus on eliminating bias caused by False Positives.

2.2 False Positive Samples

False Positive samples can be encountered in contrastive learning due to excessive or inappropriate augmentation. Similar to Debiased Contrastive Learning, we propose to estimate the positive sampling distribution in the loss function.

$$p_x^+(\mathbf{x}') = \frac{p(\mathbf{x}') - \tau^- p_x^-(\mathbf{x}')}{\tau^+} \tag{8}$$

Contrary to Debiased Contrastive Learning, drawing samples from $p(\mathbf{x}')$ will also include positive labelled datapoints, since some of them could be actually negative.

Lemma 1. *With $N \to \infty$:*

$$\mathbb{E}_{\substack{\boldsymbol{x}\sim p \\ \boldsymbol{x}^+\sim p_x^+ \\ \{\boldsymbol{x}_i^-\}_{i=1}^N \sim {p_x^-}^N}} \left[-\log \frac{e^{f(\boldsymbol{x})^T f(\boldsymbol{x}^+)}}{e^{f(\boldsymbol{x})^T f(\boldsymbol{x}^+)} + \sum_{i=1}^N e^{f(\boldsymbol{x})^T f(\boldsymbol{x}_i^-)}} \right] \longrightarrow \mathbb{E}_{\substack{\boldsymbol{x}\sim p \\ \boldsymbol{x}^-\sim p_x^-}} \left[-\log \frac{R}{R + N\mathbb{E}_{\boldsymbol{x}^-\sim p_x^-} e^{f(\boldsymbol{x})^T f(\boldsymbol{x}^-)}} \right], \tag{9}$$

where

$$R = \frac{1}{\tau^+}\left(\mathbb{E}_{\boldsymbol{x}'\sim p} e^{f(\boldsymbol{x})^T f(\boldsymbol{x}')} - \tau^- \mathbb{E}_{\boldsymbol{x}^-\sim p_x^-} e^{f(\boldsymbol{x})^T f(\boldsymbol{x}^-)}\right). \tag{10}$$

Proof. Since the expression inside the expectation is bounded, we can apply Dominated convergence theorem:

$$\lim_{N\to\infty} \mathbb{E}\left[-\log \frac{e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + \sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{x}_i^-)}} \right] = \mathbb{E}\left[\lim_{N\to\infty} -\log \frac{e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + \sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{x}_i^-)}} \right] =$$

$$= \mathbb{E}\left[-\log \frac{\mathbb{E}_{\mathbf{x}^+\sim p_x^+} e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{\mathbb{E}_{\mathbf{x}^+\sim p_x^+} e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + N\mathbb{E}_{\mathbf{x}^-\sim p_x^-} e^{f(\mathbf{x})^T f(\mathbf{x}^-)}} \right]$$

Using 8 and linearity of the expectation we complete the proof:

$$\mathbb{E}_{\mathbf{x}^+\sim p_x^+} e^{f(\mathbf{x})^T f(\mathbf{x}^+)} = \frac{1}{\tau^+}\left(\mathbb{E}_{\mathbf{x}'\sim p} e^{f(\mathbf{x})^T f(\mathbf{x}')} - \tau^- \mathbb{E}_{\mathbf{x}^-\sim p_x^-} e^{f(\mathbf{x})^T f(\mathbf{x}^-)}\right) = R.$$

■

We will denote the limit by $\tilde{L}^N_{\text{DebiasedPos}}$ and rewrite it, multiplying the nominator and the denominator of the under-logarithm expression by τ^+:

$$\tilde{L}^N_{\text{DebiasedPos}}(f) = \mathbb{E}_{\substack{\mathbf{x}\sim p \\ \mathbf{x}^-\sim p_x^-}} \left[-\log \frac{\mathbb{E}_{\mathbf{x}'\sim p} e^{f(\mathbf{x})^T f(\mathbf{x}')} - \tau^- \mathbb{E}_{\mathbf{x}^-\sim p_x^-} e^{f(\mathbf{x})^T f(\mathbf{x}_i^-)}}{\mathbb{E}_{\mathbf{x}'\sim p} e^{f(\mathbf{x})^T f(\mathbf{x}')} + \left(N\tau^+ - \tau^-\right)\mathbb{E}_{\mathbf{x}^-\sim p_x^-} e^{f(\mathbf{x})^T f(\mathbf{x}_i^-)}} \right] \tag{11}$$

With finite N:

$$L^N_{\text{DebiasedPos}}(f) = \mathbb{E}_{\substack{\mathbf{x}\sim p \\ \{\mathbf{u}_i\}_{i=1}^N \sim {p_x^-}^N \\ \mathbf{v}\sim p_x^+}} \left[-\log \frac{P_{\text{emp}}(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^N, \mathbf{v}) - \tau^- P^-_{\text{emp}}(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^N)}{P_{\text{emp}}(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^N, \mathbf{v}) + \left(N\tau^+ - \tau^-\right)P^-_{\text{emp}}(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^N)} \right], \tag{12}$$

where

$$P_{\text{emp}}(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^N, \mathbf{v}) = \frac{1}{N+2}\left(\sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{u}_i)} + e^{f(\mathbf{x})^T f(\mathbf{v})} + e^{f(\mathbf{x})^T f(\mathbf{x})} \right); \tag{13}$$

$$P^-_{\text{emp}}(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^N) = \frac{1}{N} \sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{u}_i)}. \tag{14}$$

Theorem 2.1. *If $M = 1$, for any embedding f and any $\delta > 0$ exists large enough N, that*

$$\left|\tilde{L}^N_{\textit{DebiasedPos}}(f) - L^N_{\textit{DebiasedPos}}(f)\right| \leq \left[\left(1 + \frac{\tau^-}{\tau^+} + \delta\right)\sqrt{\frac{\pi}{2N}} + \left(1 + \frac{1}{\tau^+}\right)\sqrt{\frac{\pi}{2N+2}} \right] e^{3/2} \tag{15}$$

When the number of positive samples $M > 1$, similar to [Prannay Khosla, 2021] we will aggregate using the summation over positives located outside of the log (which yields better accuracy for supervised contrastive losses rather than summation inside of the log).

$$L^{N,M}_{\text{DebiasedPos}}(f) = \mathbb{E}_{\substack{\mathbf{x}\sim p \\ \{\mathbf{u}_i\}_{i=1}^{N}\sim p^{-N} \\ \{\mathbf{v}_i\}_{i=1}^{M}\sim p^{+M}}} \frac{1}{M}\sum^{M}\left[-\log\frac{P_{\text{emp}} - \tau^- P^-_{\text{emp}}}{P_{\text{emp}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}}\right] \tag{16}$$

In addition, we will try a new method of combining losses, that takes all possible pairs of same-class (positive) elements, and averages all $L^N_{\text{DebiasedPos}}(f)$ losses.

3 Experiment

3.1 Baseline Reproduction

The goal of the experiment is to compare Multi-class N-pair Loss 1 with Negative-Debiased Loss 6. We use STL10 [Coates et al., 2011] dataset, an image recognition dataset for developing unsupervised feature learning algorithms. It contains 100000 unlabeled images of 96x96 pixels for unsupervised learning. The baseline negative-debiased model is SimCLR [Ting Chen, 2020] with ResNet-18 [He et al., 2016] as the encoder architecture, with the Adam optimizer [Diederik P. Kingma, 2014] with learning rate 0.001 and batch size 512. All models are trained for 50 epochs and evaluated by training a linear classifier after fixing the learned embedding.

The results are shown in Figure 1. While the difference between Top-5 accuracy for the debiased model and the biased model is insignificant, Top-1 accuracy of the debiased model is higher than that of the biased model.

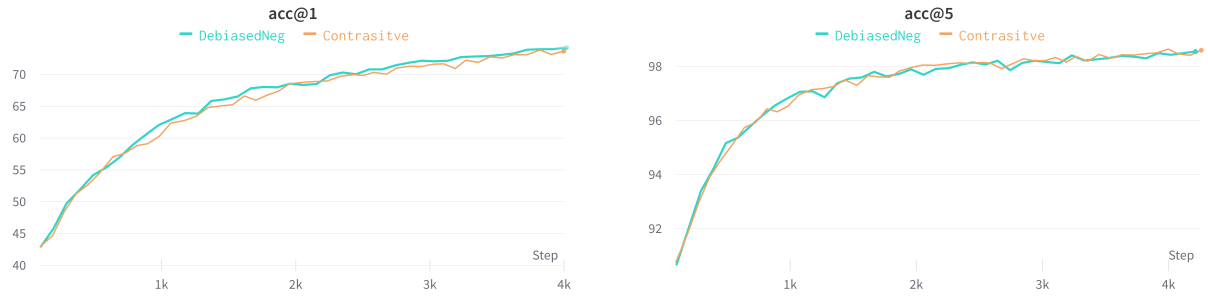

Figure 1: Classification accuracy on CIFAR10. Negative-Debiased loss is denoted as debiasedNeg.

Figure 2: Classification accuracy on CIFAR10, Positive-Debiased loss is denoted as debiasedPos.

3.2 Testing New Loss

We run the same experiment for Positive-Debiased loss 16. Figure 2 shows a tangible difference between the performances of the novel loss and the previous losses. The accuracy advantage is prevalent from the very first steps of

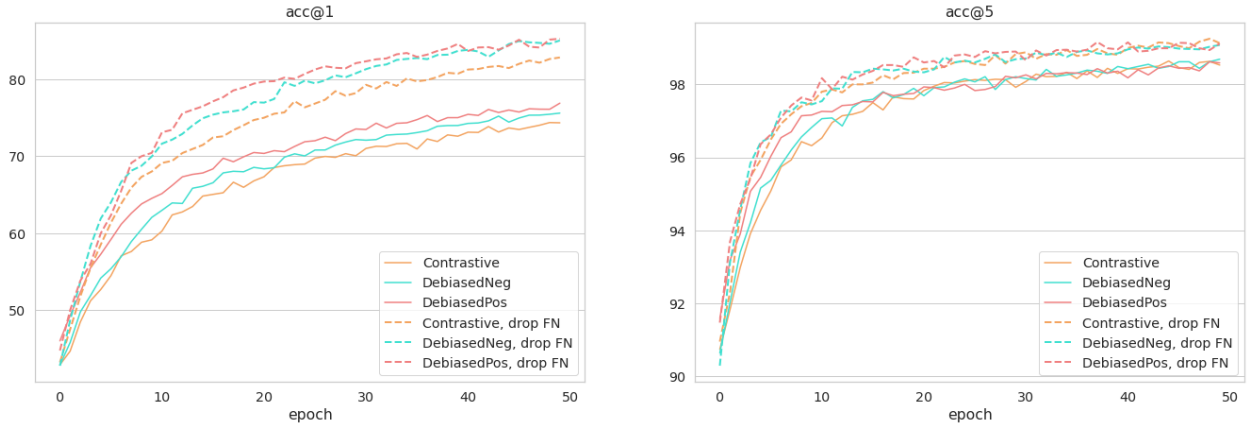

Figure 3: Classification accuracy on CIFAR10 with dropping False Negatives

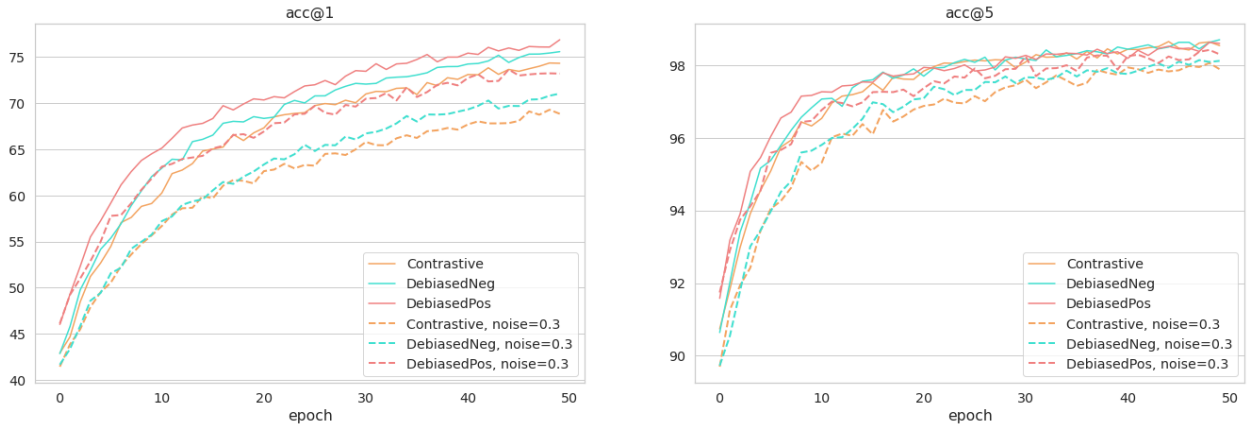

Figure 4: Classification accuracy on CIFAR10 with increased False Positive rate

training both for Top-1 and Top-5 accuracy, meaning Positive-Debiased loss is more efficient in the early stages of training.

To test the performance of the loss under the assumptions made in 2.2, we artificially exclude False Negative errors via accessing ground truth labels during training, and discover that Positive-Debiased loss still outperforms Negative-Debiased loss [Figure 3].

To understand if Positive-Debiased loss is robust to noise, we apply Gaussian blur on train images with probability 0.3, and thus increase the False Positive rate. Figure 4 demonstrates that Positive-Debiased loss is able to mitigate False Positive errors significantly better than Negative-Debiased and Contrastive losses.

Additionally we test all three losses with increased positive sample size M [Figure 5]. Larger M leads to better estimate of the loss both for Positive- and Negative-Debiased losses, yielding better accuracy as a result. It is worth noting that in all of the experiments Positive-Debiased loss achieves better accuracy in the earlier stages of training both for Top-1 and Top-5 scores.

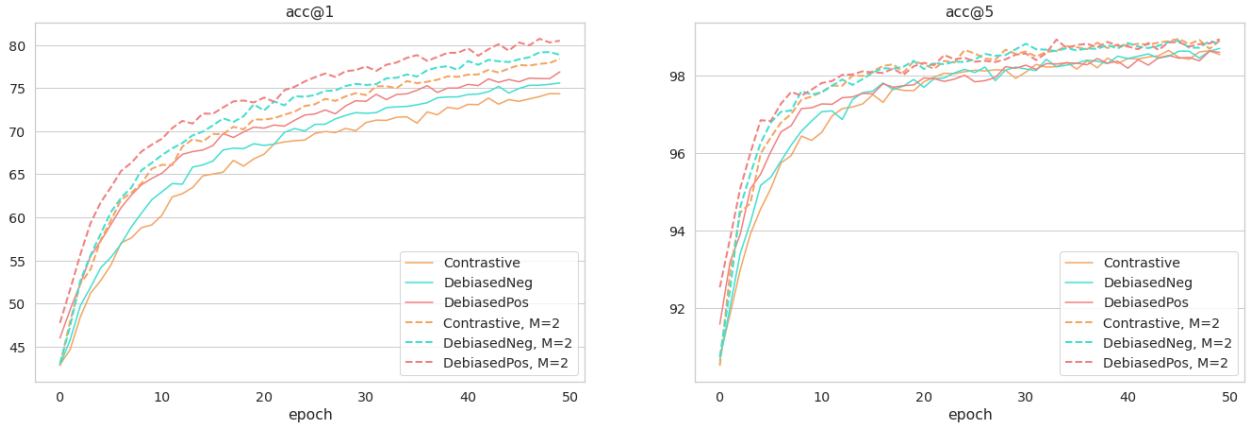

Figure 5: Classification accuracy on CIFAR10 with increased positive sample size M

References

Chen Wang Aaron Sarna Yonglong Tian Phillip Isola Aaron Maschinot Ce Liu Dilip Krishnan Prannay Khosla, Piotr Teterwak. Supervised contrastive learning. *arXiv:2004.11362*, 2021.

Lin Yen-Chen Antonio Torralba Stefanie Jegelka Ching-Yao Chuang, Joshua Robinson. Debiased contrastive learning. *arXiv:2007.00224*, 2020.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper/2016/file/6b180037abbebea991d8b1232f8a8ca9-Paper.pdf`.

Mohammad Norouzi Geoffrey Hinton Ting Chen, Simon Kornblith. A simple framework for contrastive learning of visual representations. *arXiv:2002.05709*, 2020.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 215–223, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi:10.1109/CVPR.2016.90.

Jimmy Ba Diederik P. Kingma. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Face recognition using triplet loss. *arXiv preprint arXiv:1503.03832*, 2015.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020. doi:10.1109/access.2020.3031549. URL `https://doi.org/10.1109%2Faccess.2020.3031549`.

Jeff Z. HaoChen and Tengyu Ma. A theoretical study of inductive biases in contrastive learning, 2022.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations, 2021.

Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. Use all the labels: A hierarchical multi-label contrastive learning framework, 2022.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples, 2021.

Tsai-Shien Chen, Wei-Chih Hung, Hung-Yu Tseng, Shao-Yi Chien, and Ming-Hsuan Yang. Incremental false negative detection for contrastive learning, 2022.

Ching-Yun Ko, Jeet Mohapatra, Sijia Liu, Pin-Yu Chen, Luca Daniel, and Lily Weng. Revisiting contrastive learning through the lens of neighborhood component analysis: an integrated framework, 2022.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views, 2019.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere, 2022.

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning, 2019.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020.

Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning, 2020.

A Proof of Theorem 2

We recall theorem 2.1:

Theorem 2.1. *If $M = 1$, for any embedding f and any $\delta > 0$ exists large enough N, that*

$$\left|\tilde{L}^N_{DebiasedPos}(f) - L^N_{DebiasedPos}(f)\right| \leq \left[\left(1 + \frac{\tau^-}{\tau^+} + \delta\right)\sqrt{\frac{\pi}{2N}} + \left(1 + \frac{1}{\tau^+}\right)\sqrt{\frac{\pi}{2N+2}}\right]e^{3/2} \tag{15}$$

Proof. Let us denote:

$$P_{\text{est}} := \mathbb{E}_{\mathbf{x}' \sim p} e^{f(\mathbf{x})^T f(\mathbf{x}')}$$

$$P^-_{\text{est}} := \mathbb{E}_{\mathbf{x}^- \sim p^-_x} e^{f(\mathbf{x})^T f(\mathbf{x}^-)}$$

$$A_1 := -\log \frac{P_{\text{est}} - \tau^- P^-_{\text{est}}}{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{est}}}$$

$$A_2 := -\log \frac{P_{\text{est}} - \tau^- P^-_{\text{est}}}{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}}$$

$$A_3 := -\log \frac{P_{\text{est}} - \tau^- P^-_{\text{est}}}{P_{\text{emp}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}}$$

$$A_4 := -\log \frac{P_{\text{emp}} - \tau^- P^-_{\text{est}}}{P_{\text{emp}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}}$$

$$A_5 := -\log \frac{P_{\text{emp}} - \tau^- P^-_{\text{emp}}}{P_{\text{emp}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}}$$

Note that $e^{-1} \leq P_{\text{est}}, P^-_{\text{est}} \leq e$. We extend the proofs of Lemma A.2 and Theorem 3 from [Ching-Yao Chuang, 2020] to get results for DebiasedPos Loss.

1.

$$\Delta_1 := |A_2 - A_1| = \left| -\log \frac{P_{\text{est}} - \tau^- P^-_{\text{est}}}{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}} + \log \frac{P_{\text{est}} - \tau^- P^-_{\text{est}}}{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{est}}} \right|.$$

$$\forall \varepsilon > 0 : \mathbb{P}(\Delta_1 \geq \varepsilon) = \mathbb{P}\left(\left| \log\{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}\} - \log\{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{est}}\} \right| \geq \varepsilon\right) =$$

$$= \mathbb{P}\left(\log\{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}\} - \log\{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{est}}\} \geq \varepsilon\right) +$$

$$= \mathbb{P}\left(\log\{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}\} - \log\{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{est}}\} \leq -\varepsilon\right)$$

The first term can be bounded as:

$$\mathbb{P}\left(\log\{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}\} - \log\{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{est}}\} \geq \varepsilon\right) =$$

$$= \mathbb{P}\left(\log \frac{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}}{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{est}}} \geq \varepsilon\right) \leq \mathbb{P}\left(\frac{(N\tau^+ - \tau^-)(P^-_{\text{emp}} - P^-_{\text{est}})}{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{est}}} \geq \varepsilon\right) =$$

$$= \mathbb{P}\left(P^-_{\text{emp}} - P^-_{\text{est}} \geq \varepsilon\left\{ \frac{1}{N\tau^+ - \tau^-} P_{\text{est}} + P^-_{\text{est}} \right\}\right) \leq \mathbb{P}(P^-_{\text{emp}} - P^-_{\text{est}} \geq \varepsilon e^{-1})$$

Here we used the fact that $\log x \leq x - 1$ for $x > 0$, and $1/(N\tau^+ - \tau^-)P_{\text{est}} + P^-_{\text{est}} \geq e^{-1}$. Likewise bounding the second term:

$$\mathbb{P}\left(\log\{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}\} - \log\{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{est}}\} \leq -\varepsilon\right) \leq \mathbb{P}(P^-_{\text{est}} - P^-_{\text{emp}} \geq \varepsilon e^{-1})$$

and applying Hoeffding's inequality, we have:

$$\mathbb{P}(\Delta_1 \geq \varepsilon) \leq \mathbb{P}(|P^-_{\text{emp}} - P^-_{\text{est}}| \geq \varepsilon e^{-1}) = \mathbb{P}\bigg(\bigg|\frac{1}{N}\sum_{i=1}^{N} e^{f(\mathbf{x})^T f(\mathbf{u}_i)} - \mathbb{E}_{\mathbf{x}^- \sim p_x^-} e^{f(\mathbf{x})^T f(\mathbf{x}^-)}\bigg| \geq \varepsilon e^{-1}\bigg) \leq$$
$$\leq 2\exp\bigg(-\frac{2N\varepsilon^2 e^{-2}}{e - e^{-1}}\bigg) \leq 2\exp\bigg(-\frac{2N\varepsilon^2}{e^3}\bigg)$$

Finally, we write the expectation of Δ_1 as the integral of its tail probability

$$\big|\mathbb{E}A_2 - \mathbb{E}A_1\big| \leq \mathbb{E}\Delta_1 \leq \int_0^\infty 2\exp\bigg(-\frac{2N\varepsilon^2}{e^3}\bigg)d\varepsilon = \sqrt{\frac{\pi}{2N}}e^{3/2}.$$

2.

$$\Delta_2 := |A_3 - A_2| = \bigg|-\log\frac{P_{\text{est}} - \tau^- P^-_{\text{est}}}{P_{\text{emp}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}} + \log\frac{P_{\text{est}} - \tau^- P^-_{\text{est}}}{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}}\bigg|.$$
$$\forall\varepsilon > 0 : \mathbb{P}(\Delta_2 \geq \varepsilon) = \mathbb{P}\bigg(\bigg|\log\{P_{\text{emp}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}\} - \log\{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}\}\bigg| \geq \varepsilon\bigg)$$

Like before, we split the probability into sum of two terms, and bound the first term as:

$$\mathbb{P}\bigg(\log\{P_{\text{emp}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}\} - \log\{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}\} \geq \varepsilon\bigg) =$$
$$= \mathbb{P}\bigg(\log\frac{P_{\text{emp}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}}{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}} \geq \varepsilon\bigg) \leq \mathbb{P}\bigg(\frac{P_{\text{emp}} - P_{\text{est}}}{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}} \geq \varepsilon\bigg) =$$
$$= \mathbb{P}\bigg(P_{\text{emp}} - P_{\text{est}} \geq \varepsilon\big\{P_{\text{est}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}\big\}\bigg) \leq \mathbb{P}(P_{\text{emp}} - P_{\text{est}} \geq \varepsilon e^{-1})$$

This holds because with N big enough, $N\tau^+ \geq \tau^-$, and thus $1/(N\tau^+ - \tau^-)P_{\text{est}} + P^-_{\text{est}} \geq e^{-1}$.

$$\mathbb{P}(\Delta_2 \geq \varepsilon) \leq \mathbb{P}(|P_{\text{emp}} - P_{\text{est}}| \geq \varepsilon e^{-1}) =$$
$$= \mathbb{P}\bigg(\bigg|\frac{1}{N+2}\bigg(\sum_{i=1}^{N} e^{f(\mathbf{x})^T f(\mathbf{u}_i)} + e^{f(\mathbf{x})^T f(\mathbf{v})} + e^{f(\mathbf{x})^T f(\mathbf{x})}\bigg) - \mathbb{E}_{\mathbf{x}' \sim p} e^{f(\mathbf{x})^T f(\mathbf{x}')}\bigg| \geq \varepsilon e^{-1}\bigg) \leq$$
$$\leq 2\exp\bigg(-\frac{2(N+2)\varepsilon^2}{e^3}\bigg)$$
$$\big|\mathbb{E}A_3 - \mathbb{E}A_2\big| \leq \mathbb{E}\Delta_2 \leq \int_0^\infty 2\exp\bigg(-\frac{2(N+2)\varepsilon^2}{e^3}\bigg)d\varepsilon = \sqrt{\frac{\pi}{2(N+2)}}e^{3/2}.$$

3.

$$\Delta_3 := |A_3 - A_4| = \bigg|-\log\frac{P_{\text{est}} - \tau^- P^-_{\text{est}}}{P_{\text{emp}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}} + \log\frac{P_{\text{emp}} - \tau^- P^-_{\text{est}}}{P_{\text{emp}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}}\bigg|.$$
$$\forall\varepsilon > 0 : \mathbb{P}(\Delta_3 \geq \varepsilon) = \mathbb{P}\bigg(\bigg|\log\{P_{\text{emp}} - \tau^- P^-_{\text{est}}\} - \log\{P_{\text{est}} - \tau^- P^-_{\text{est}}\}\bigg| \geq \varepsilon\bigg)$$

Like before, we split the probability into sum of two terms, and bound the first term as:

$$\mathbb{P}\Big(\log\{P_{\text{emp}} - \tau^- P^-_{\text{est}}\} - \log\{P_{\text{est}} - \tau^- P^-_{\text{est}}\} \geq \varepsilon\Big) =$$
$$= \mathbb{P}\Big(\log\frac{P_{\text{emp}} - \tau^- P^-_{\text{est}}}{P_{\text{est}} - \tau^- P^-_{\text{est}}} \geq \varepsilon\Big) \leq \mathbb{P}\Big(\frac{P_{\text{emp}} - P_{\text{est}}}{P_{\text{est}} - \tau^- P^-_{\text{est}}} \geq \varepsilon\Big) =$$
$$= \mathbb{P}\Big(P_{\text{emp}} - P_{\text{est}} \geq \varepsilon\big\{P_{\text{est}} - \tau^- P^-_{\text{est}}\big\}\Big) \leq \mathbb{P}(P_{\text{emp}} - P_{\text{est}} \geq \varepsilon\tau^+ e^{-1})$$

This time the bound is different, because $P_{\text{est}} - \tau^- P^-_{\text{est}} = \tau^+ P^+_{\text{est}} := \tau^+ \mathbb{E}_{\mathbf{x}^+\sim p^+_x} e^{f(\mathbf{x})^T f(\mathbf{x}^+)} \geq \tau^+ e^{-1}$.

$$\mathbb{P}(\Delta_3 \geq \varepsilon) \leq \mathbb{P}(|P_{\text{emp}} - P_{\text{est}}| \geq \varepsilon\tau^+ e^{-1}) =$$
$$= \mathbb{P}\Bigg(\Bigg|\frac{1}{N+2}\Bigg(\sum_{i=1}^{N} e^{f(\mathbf{x})^T f(\mathbf{u}_i)} + e^{f(\mathbf{x})^T f(\mathbf{v})} + e^{f(\mathbf{x})^T f(\mathbf{x})}\Bigg) - \mathbb{E}_{\mathbf{x}'\sim p} e^{f(\mathbf{x})^T f(\mathbf{x}')}\Bigg| \geq \varepsilon\tau^+ e^{-1}\Bigg) \leq$$
$$\leq 2\exp\Bigg(-\frac{2(N+2)\varepsilon^2(\tau^+)^2}{e^3}\Bigg)$$
$$\big|\mathbb{E}A_3 - \mathbb{E}A_4\big| \leq \mathbb{E}\Delta_3 \leq \int_0^\infty 2\exp\Bigg(-\frac{2(N+2)\varepsilon^2(\tau^+)^2}{e^3}\Bigg)d\varepsilon = \frac{1}{\tau^+}\sqrt{\frac{\pi}{2(N+2)}}e^{3/2}.$$

4.

$$\Delta_4 := |A_5 - A_4| = \Bigg|-\log\frac{P_{\text{emp}} - \tau^- P^-_{\text{emp}}}{P_{\text{emp}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}} + \log\frac{P_{\text{emp}} - \tau^- P^-_{\text{est}}}{P_{\text{emp}} + (N\tau^+ - \tau^-)P^-_{\text{emp}}}\Bigg|.$$
$$\forall\varepsilon > 0 : \mathbb{P}(\Delta_4 \geq \varepsilon) = \mathbb{P}\Big(\Big|\log\{P_{\text{emp}} - \tau^- P^-_{\text{est}}\} - \log\{P_{\text{emp}} - \tau^- P^-_{\text{emp}}\}\Big| \geq \varepsilon\Big)$$

Like before, we split the probability into sum of two terms, and bound the first term as:

$$\mathbb{P}\Big(\log\{P_{\text{emp}} - \tau^- P^-_{\text{est}}\} - \log\{P_{\text{emp}} - \tau^- P^-_{\text{emp}}\} \geq \varepsilon\Big) =$$
$$= \mathbb{P}\Big(\log\frac{P_{\text{emp}} - \tau^- P^-_{\text{est}}}{P_{\text{emp}} - \tau^- P^-_{\text{emp}}} \geq \varepsilon\Big) \leq \mathbb{P}\Big(\frac{\tau^-(P^-_{\text{emp}} - P^-_{\text{est}})}{P_{\text{emp}} - \tau^- P^-_{\text{emp}}} \geq \varepsilon\Big) =$$
$$= \mathbb{P}\Big(P^-_{\text{emp}} - P^-_{\text{est}} \geq \varepsilon\Big\{\frac{1}{\tau^-}P_{\text{emp}} - P^-_{\text{emp}}\Big\}\Big) \leq \mathbb{P}\Big(P^-_{\text{emp}} - P^-_{\text{est}} \geq \varepsilon\Big(\frac{\tau^+}{\tau^-}e^{-1} - \delta\Big)\Big)$$

To get a lower bound of the expression in braces for big N, we will take a slightly smaller value than its limit.

$$\frac{1}{\tau^-}P_{\text{emp}} - P^-_{\text{emp}} = \frac{1}{\tau^-}\frac{1}{N+2}\Bigg(\sum_{i=1}^{N} e^{f(\mathbf{x})^T f(\mathbf{u}_i)} + e^{f(\mathbf{x})^T f(\mathbf{v})} + e^{f(\mathbf{x})^T f(\mathbf{x})}\Bigg) - \frac{1}{N}\sum_{i=1}^{N} e^{f(\mathbf{x})^T f(\mathbf{u}_i)} =$$
$$= \Bigg(\frac{1}{\tau^-}\frac{1}{N+2} - \frac{1}{N}\Bigg)\Bigg(\sum_{i=1}^{N} e^{f(\mathbf{x})^T f(\mathbf{u}_i)}\Bigg) + \frac{e^{f(\mathbf{x})^T f(\mathbf{v})} + e^{f(\mathbf{x})^T f(\mathbf{x})}}{\tau^-(N+2)} \xrightarrow[N\to\infty]{} \frac{e^{-1}}{\tau^-} - e^{-1} = \frac{\tau^+}{\tau^-}e^{-1}$$

$$\mathbb{P}(\Delta_4 \geq \varepsilon) \leq \mathbb{P}\left(|P^-_{\text{emp}} - P^-_{\text{est}}| \geq \varepsilon\left(\frac{\tau^+}{\tau^-}e^{-1} - \delta\right)\right) =$$
$$= \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} e^{f(\mathbf{x})^T f(\mathbf{u}_i)} - \mathbb{E}_{\mathbf{x}^- \sim p_x^-} e^{f(\mathbf{x})^T f(\mathbf{x}^-)}\right| \geq \varepsilon\left(\frac{\tau^+}{\tau^-}e^{-1} - \delta\right)\right) \leq$$
$$\leq 2\exp\left(-\frac{2N\varepsilon^2((\tau^+/\tau^-)e^{-1} - \delta)^2}{e}\right) =: 2\exp\left(-\frac{2N\varepsilon^2(\tau^+/\tau^- - \delta)^2}{e^3}\right)$$
$$\left|\mathbb{E}A_5 - \mathbb{E}A_4\right| \leq \mathbb{E}\Delta_4 \leq \int_0^\infty 2\exp\left(-\frac{2N\varepsilon^2(\tau^+/\tau^- - \delta)^2}{e^3}\right)d\varepsilon =$$
$$= \frac{1}{\tau^+/\tau^- - \delta}\sqrt{\frac{\pi}{2N}}e^{3/2} =: \left(\frac{\tau^-}{\tau^+} - \delta\right)\sqrt{\frac{\pi}{2N}}e^{3/2}.$$

The result follows by summing all four absolute differences:

$$\left|\tilde{L}^N_{\text{DebiasedPos}}(f) - L^N_{\text{DebiasedPos}}(f)\right| = \left|\mathbb{E}A_1 - \mathbb{E}A_5\right| \leq \left|\mathbb{E}A_1 - \mathbb{E}A_2\right| + \left|\mathbb{E}A_2 - \mathbb{E}A_3\right| + \left|\mathbb{E}A_3 - \mathbb{E}A_4\right| + \left|\mathbb{E}A_4 - \mathbb{E}A_5\right| \leq$$
$$\leq \left[\left(1 + \frac{\tau^-}{\tau^+} + \delta\right)\sqrt{\frac{\pi}{2N}} + \left(1 + \frac{1}{\tau^+}\right)\sqrt{\frac{\pi}{2(N+2)}}\right]e^{3/2}$$

■