MITIGATING DISTRIBUTIONAL BIASES IN CONTRASTIVE LEARNING

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

Roman Isachenko
Moscow Institute of Physics and Technology
Yandex
roman.isachenko@phystech.edu

ABSTRACT

Recently contrastive learning has regained popularity as a self-supervised representation learning technique. It involves comparing positive (similar) and negative (dissimilar) pairs of samples to learn representations without labels. However, false negative and false positive errors in sampling lead to the loss function bias. 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. Using the debiased representations, we measure accuracy of predictions in the classification task. The experiments are carried out on the CIFAR10 dataset, demonstrating the applicability and robustness of the proposed method in scenarios where extensive labeling is expensive or not feasible.

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 popular solution for this problem is contrastive learning, a technique that uses the principle of contrasting samples against each other to learn shared and distinctive attributes between different data classes. 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 a 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 [Chen et al., 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.

During training, the lack of true labels often necessitates the random selection of negative instances $\mathbf{x}^-$ from the training data. However, this approach results in sampling bias, where $\mathbf{x}^-$ may actually be similar to $\mathbf{x}$. This bias leads to a significant drop in performance [Hong and Yang, 2021].

To address this issue, [Khosla et al., 2021] extend the self-supervised batch contrastive approach to the fully-supervised setting, allowing to effectively leverage label information. The experiments demonstrate that the supervised contrastive

loss outperforms the base cross-entropy, albeit only by a marginal degree. It also outperforms the cross-entropy on robustness benchmark (ImageNet-C, which applies common naturally occurring 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 [Chuang et al., 2020]. Given N negative and M positive samples, $L_{\text{DebiasedNeg}}^{N,M}(f)$ 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 contain False Positive samples.

This work aims to develop a novel 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

Let $\mathcal{X}$ be the object space, and let $p(x)$ be the data distribution over $\mathcal{X}$, then the target embedding is $f : \mathcal{X} \to \mathbb{R}^d$. In [Chuang et al., 2020] the distribution over a set of the discrete latent classes $\mathcal{C}$ is denoted by $\rho(c)$, hence the joint distribution $p_{x,c}(\mathbf{x}, c) = p(\mathbf{x}|c)\rho(c)$. Let $h : \mathcal{X} \to \mathcal{C}$ be the function assigning the 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:

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

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^+.

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

<u>Problem statement:</u>   Derive a loss function that approximates the ideal L_{Unbiased}^N, 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 suggest a cost-effective approach for estimating $p_x^-(\mathbf{x}')$, 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:

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

where the empirical estimate of the term corresponding to p_x^- is defined as

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

The implementation of this loss for the SimCLR framework resulted in improved accuracy compared to the regular biased loss. Building on this theoretical foundation, our study aims to develop an algorithm that specifically addresses the issue of eliminating False Positive bias.

2.2 False Positive Samples

Excessive or inappropriate augmentation leads to False Positive samples in contrastive learning. Similar to Debiased Contrastive Learning, we propose estimating the positive sampling distribution within the loss function.

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

Unlike Debiased Contrastive Learning, drawing samples from $p(\mathbf{x}')$ will include positive labeled data points that may be, in fact, negative.

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

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

where

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

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 equation 8 and the 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 $\tilde{L}_{\text{DebiasedPos}}^N$ and rewrite it, multiplying the numerator and the denominator of the under-logarithm expression by τ^+:

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

With finite N:

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

where

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

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

When N is finite and $M = 1$, theorem 2.1 bounds the error as decreasing with rate $O(N^{-1/2})$.

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

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

2.3 Positive Sample Aggregation

When the number of positive samples $M > 1$, similar to [Khosla et al., 2021], we will aggregate losses using the summation over positive samples located outside of the log. In our case this means taking all possible pairs of same-class (positive) elements, and averaging all corresponding $L_{\text{DebiasedPos}}^{N}(f)$ losses. We will refer to this method as *loss-combination*.

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

Another way to aggregate positive samples is to do so inside the estimates P_{emp}, following the example of [Chuang et al., 2020], which we will refer to as *pos-grouping*. Empirically, the experiments in Section 3.3 show that while *pos-grouping* is computationally less expensive, *loss-combination* yields better accuracy with larger M.

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 CIFAR10 [Krizhevsky and Hinton, 2009] dataset, an image recognition dataset for image classification. It consists of 60000 32x32 color images in 10 classes. The baseline Negative-Debiased model is SimCLR [Chen et al., 2020] with ResNet-18 [He et al., 2015] as the encoder architecture, with the Adam optimizer [Kingma and Ba, 2017], 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 primary 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.

3.2 Testing New Loss

We run the same experiment for Positive-Debiased loss [14]. The code is available at our GitHub. Figure 1 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 training both for Top-1 and Top-5 accuracy, meaning Positive-Debiased loss is more efficient in the early stages of training.
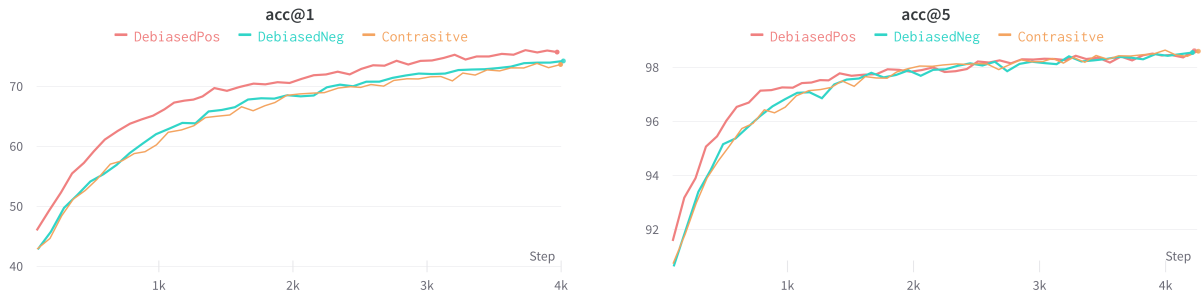

Figure 1: Classification accuracy on CIFAR10, Negative- and Positive-Debiased losses are denoted as debiasedNeg and debiasedPos.

To test the performance of the loss under the assumptions made in Section 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 2].

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 3 demonstrates that Positive-Debiased loss is able to mitigate False Positive errors significantly better than Negative-Debiased and Contrastive losses.
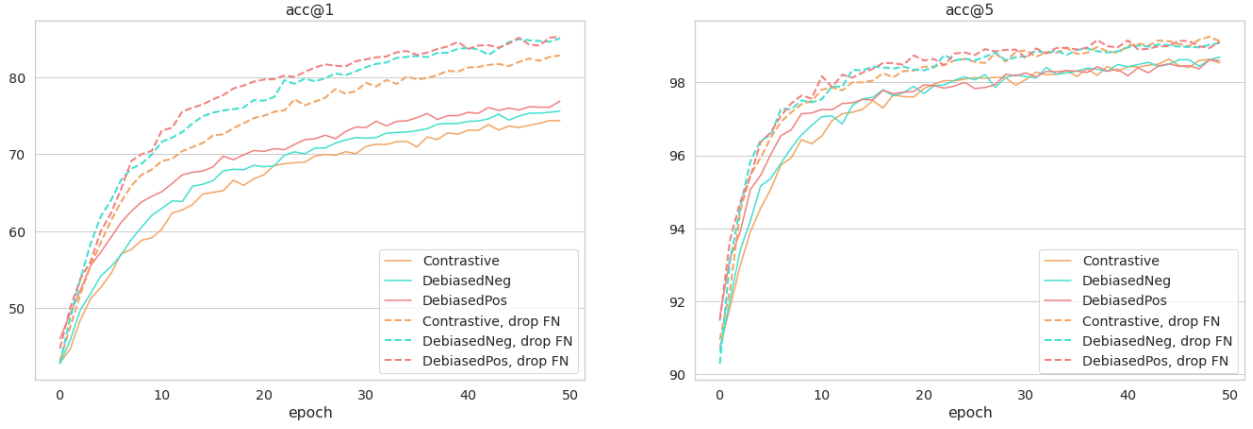
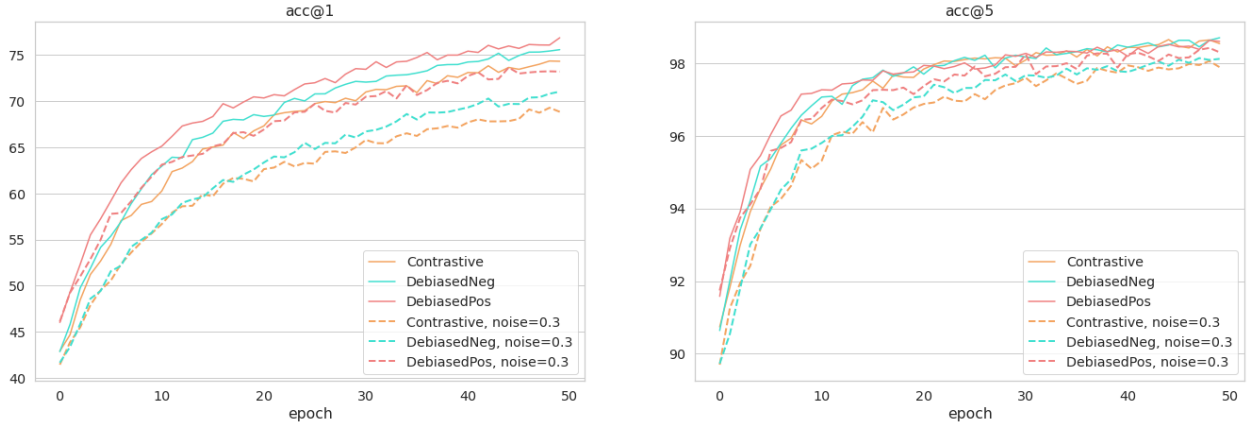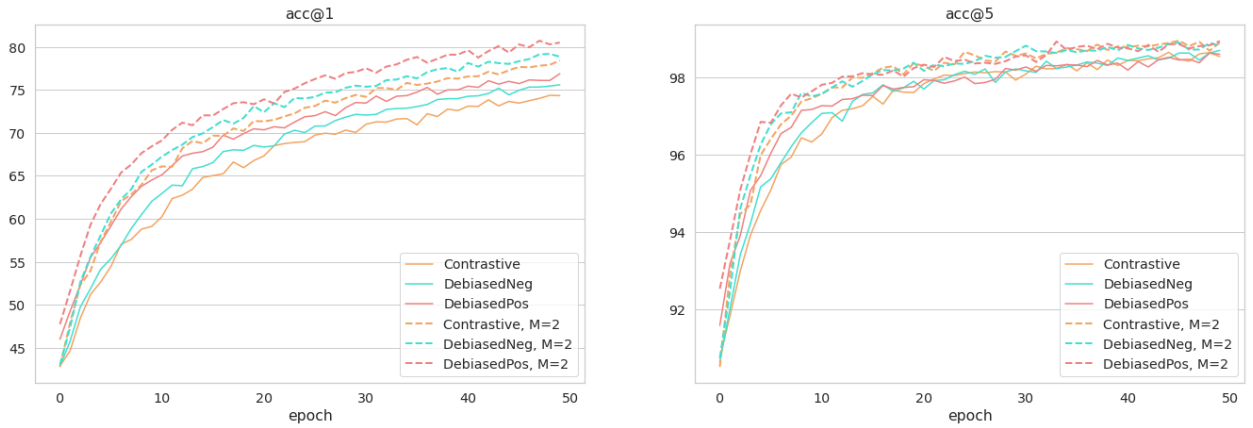Figure 2: Classification accuracy on CIFAR10 with dropping False Negatives

Figure 3: Classification accuracy on CIFAR10 with an increased False Positive rate

Figure 4: Classification accuracy on CIFAR10 with an increased positive sample size M
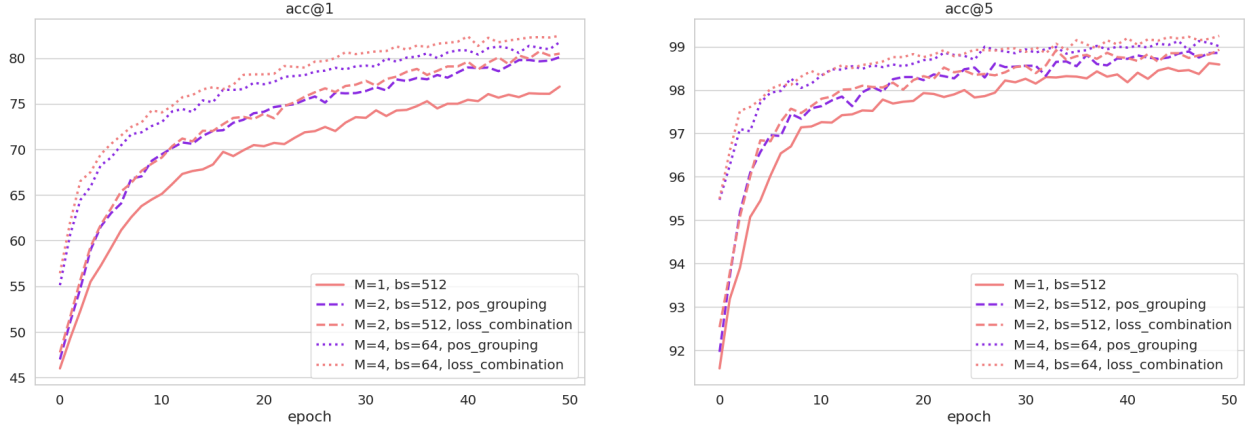
Figure 5: Classification accuracy on CIFAR10 with various positive sample aggregations

Table 1: Aggregation over M positive samples for DebiasedPos

Method	M	bs	Training time (min)	acc@1	acc@5
-	1	512	122.85	76.76	98.61
pos-grouping	2	512	**181.87**	80.1	98.86
loss-combination	2	512	192.82	**80.5**	**98.93**
pos-grouping	4	64	**278.88**	81.75	99.01
loss-combination	4	64	308.25	**82.44**	**99.25**

Additionally, we test all three losses with increased positive sample size M [Figure 4]. 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.

3.3 Positive Sample Aggregation

In this experiment, we compare methods of aggregation over positive samples introduced in Section 2.3. We train SimCLR with the same parameters, except for setting batch size to 64 when testing $M = 4$ for faster training. As expected, there is a trade-off between accuracy and speed. The results in Table 1 show that *pos-grouping* improves over *loss-combination* in accuracy, but is computationally slower. Figure 5 shows that with bigger M the advantage of *loss-combination* over *pos-grouping* gets more significant.

3.4 Discussion

1. Positive-Debiased loss is based on the assumption of true negative samples, i.e., with no False Negative errors; however, it still improves accuracy of classification when this assumption does not hold.

2. Another implicit assumption we make is that the classes have uniform distribution, which is not true for many datasets, and it is an interesting direction of further research to additionally account for the biases in the class distribution.

4 Conclusion

In this paper, we propose a new method for debiasing contrastive loss based on mitigating biases caused by False Positive errors. Theoretical analysis shows that computationally Positive-Debiased and baseline losses are equivalent. However, in a scenario with an increased False Positive rate our loss yields better accuracy, which shows that it is more robust. Moreover, Positive-Debiased loss is more dominant in the early epochs. We further explore 2 methods of aggregating loss over positive samples, and infer that *pos-grouping* trains faster, while *loss-combination* leads to better performance.

References

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

Ching-Yao Chuang, Joshua Robinson, Lin Yen-Chen, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning, 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_files/paper/2016/file/6b180037abbebea991d8b1232f8a8ca9-Paper.pdf`.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 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, 2015.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2015. doi:10.1109/cvpr.2015.7298682. URL `https://doi.org/10.1109%2Fcvpr.2015.7298682`.

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.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

Youngkyu Hong and Eunho Yang. Unbiased classification through bias-contrastive and bias-balanced learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26449–26461. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/de8aa43e5d5fa8536cf23e54244476fa-Paper.pdf`.

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 [Chuang et al., 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:

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

$$\leq 2\exp\left(-\frac{2N\varepsilon^2 e^{-2}}{e - e^{-1}}\right) \leq 2\exp\left(-\frac{2N\varepsilon^2}{e^3}\right)$$

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

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

2.

$$\Delta_2 := |A_3 - A_2| = \left|-\log\frac{P_{\text{est}} - \tau^- P_{\text{est}}^-}{P_{\text{emp}} + (N\tau^+ - \tau^-)P_{\text{emp}}^-} + \log\frac{P_{\text{est}} - \tau^- P_{\text{est}}^-}{P_{\text{est}} + (N\tau^+ - \tau^-)P_{\text{emp}}^-}\right|.$$

$$\forall \varepsilon > 0 : \mathbb{P}(\Delta_2 \geq \varepsilon) = \mathbb{P}\left(\left|\log\{P_{\text{emp}} + (N\tau^+ - \tau^-)P_{\text{emp}}^-\} - \log\{P_{\text{est}} + (N\tau^+ - \tau^-)P_{\text{emp}}^-\}\right| \geq \varepsilon\right)$$

We split the probability into a sum of two terms, and bound the first term as:

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

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

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

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

$$\mathbb{P}(\Delta_2 \geq \varepsilon) \leq \mathbb{P}(|P_{\text{emp}} - P_{\text{est}}| \geq \varepsilon e^{-1}) =$$

$$= \mathbb{P}\left(\left|\frac{1}{N+2}\left(\sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{u}_i)} + e^{f(\mathbf{x})^T f(\mathbf{v})} + e^{f(\mathbf{x})^T f(\mathbf{x})}\right) - \mathbb{E}_{\mathbf{x}' \sim p} e^{f(\mathbf{x})^T f(\mathbf{x}')}\right| \geq \varepsilon e^{-1}\right) \leq$$

$$\leq 2\exp\left(-\frac{2(N+2)\varepsilon^2}{e^3}\right)$$

$$\left|\mathbb{E}A_3 - \mathbb{E}A_2\right| \leq \mathbb{E}\Delta_2 \leq \int_0^\infty 2\exp\left(-\frac{2(N+2)\varepsilon^2}{e^3}\right)d\varepsilon = \sqrt{\frac{\pi}{2(N+2)}}e^{3/2}.$$

3.

$$\Delta_3 := |A_3 - A_4| = \left|-\log\frac{P_{\text{est}} - \tau^- P_{\text{est}}^-}{P_{\text{emp}} + (N\tau^+ - \tau^-)P_{\text{emp}}^-} + \log\frac{P_{\text{emp}} - \tau^- P_{\text{est}}^-}{P_{\text{emp}} + (N\tau^+ - \tau^-)P_{\text{emp}}^-}\right|.$$

$$\forall \varepsilon > 0 : \mathbb{P}(\Delta_3 \geq \varepsilon) = \mathbb{P}\left(\left|\log\{P_{\text{emp}} - \tau^- P_{\text{est}}^-\} - \log\{P_{\text{est}} - \tau^- P_{\text{est}}^-\}\right| \geq \varepsilon\right)$$

9

We split the probability into a sum of two terms, and bound the first term as:

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

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

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

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

$$\mathbb{P}(\Delta_3 \geq \varepsilon) \leq \mathbb{P}(|P_{\text{emp}} - P_{\text{est}}| \geq \varepsilon\tau^+ e^{-1}) =$$

$$= \mathbb{P}\left(\left| \frac{1}{N+2}\left(\sum_{i=1}^{N} e^{f(\mathbf{x})^T f(\mathbf{u}_i)} + e^{f(\mathbf{x})^T f(\mathbf{v})} + e^{f(\mathbf{x})^T f(\mathbf{x})} \right) - \mathbb{E}_{\mathbf{x}' \sim p} e^{f(\mathbf{x})^T f(\mathbf{x}')} \right| \geq \varepsilon\tau^+ e^{-1} \right) \leq$$

$$\leq 2\exp\left(- \frac{2(N+2)\varepsilon^2(\tau^+)^2}{e^3} \right)$$

$$|\mathbb{E}A_3 - \mathbb{E}A_4| \leq \mathbb{E}\Delta_3 \leq \int_0^\infty 2\exp\left(- \frac{2(N+2)\varepsilon^2(\tau^+)^2}{e^3} \right) d\varepsilon = \frac{1}{\tau^+}\sqrt{\frac{\pi}{2(N+2)}}e^{3/2}.$$

4.

$$\Delta_4 := |A_5 - A_4| = \left| -\log \frac{P_{\text{emp}} - \tau^- P_{\text{emp}}^-}{P_{\text{emp}} + (N\tau^+ - \tau^-)P_{\text{emp}}^-} + \log \frac{P_{\text{emp}} - \tau^- P_{\text{est}}^-}{P_{\text{emp}} + (N\tau^+ - \tau^-)P_{\text{emp}}^-} \right|.$$

$$\forall \varepsilon > 0 : \mathbb{P}(\Delta_4 \geq \varepsilon) = \mathbb{P}\left(\left| \log\{P_{\text{emp}} - \tau^- P_{\text{est}}^-\} - \log\{P_{\text{emp}} - \tau^- P_{\text{emp}}^-\} \right| \geq \varepsilon \right)$$

We split the probability into a sum of two terms, and bound the first term as:

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

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

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

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

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

$$= \left(\frac{1}{\tau^-}\frac{1}{N+2} - \frac{1}{N} \right)\left(\sum_{i=1}^{N} e^{f(\mathbf{x})^T f(\mathbf{u}_i)} \right) + \frac{e^{f(\mathbf{x})^T f(\mathbf{v})} + e^{f(\mathbf{x})^T f(\mathbf{x})}}{\tau^-(N+2)} \xrightarrow[N\to\infty]{} \frac{e^{-1}}{\tau^-} - e^{-1} = \frac{\tau^+}{\tau^-}e^{-1}$$

$$\mathbb{P}(\Delta_4 \geq \varepsilon) \leq \mathbb{P}\left(|P_{\text{emp}}^- - P_{\text{est}}^-| \geq \varepsilon\left(\frac{\tau^+}{\tau^-}e^{-1} - \delta\right)\right) =$$

$$= \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}e^{f(\mathbf{x})^T f(\mathbf{u}_i)} - \mathbb{E}_{\mathbf{x}^- \sim p_x^-}e^{f(\mathbf{x})^T f(\mathbf{x}^-)}\right| \geq \varepsilon\left(\frac{\tau^+}{\tau^-}e^{-1} - \delta\right)\right) \leq$$

$$\leq 2\exp\left(-\frac{2N\varepsilon^2((\tau^+/\tau^-)e^{-1} - \delta)^2}{e}\right) =: 2\exp\left(-\frac{2N\varepsilon^2(\tau^+/\tau^- - \delta)^2}{e^3}\right)$$

$$\left|\mathbb{E}A_5 - \mathbb{E}A_4\right| \leq \mathbb{E}\Delta_4 \leq \int_0^\infty 2\exp\left(-\frac{2N\varepsilon^2(\tau^+/\tau^- - \delta)^2}{e^3}\right)d\varepsilon =$$

$$= \frac{1}{\tau^+/\tau^- - \delta}\sqrt{\frac{\pi}{2N}}e^{3/2} =: \left(\frac{\tau^-}{\tau^+} - \delta\right)\sqrt{\frac{\pi}{2N}}e^{3/2}.$$

The result follows by summing all four absolute differences:

$$\left|\tilde{L}_{\text{DebiasedPos}}^N(f) - L_{\text{DebiasedPos}}^N(f)\right| = \left|\mathbb{E}A_1 - \mathbb{E}A_5\right| \leq \left|\mathbb{E}A_1 - \mathbb{E}A_2\right| + \left|\mathbb{E}A_2 - \mathbb{E}A_3\right| + \left|\mathbb{E}A_3 - \mathbb{E}A_4\right| + \left|\mathbb{E}A_4 - \mathbb{E}A_5\right| \leq$$

$$\leq \left[\left(1 + \frac{\tau^-}{\tau^+} + \delta\right)\sqrt{\frac{\pi}{2N}} + \left(1 + \frac{1}{\tau^+}\right)\sqrt{\frac{\pi}{2(N+2)}}\right]e^{3/2}$$

∎