Contrastive Learning with Hard Negative Samples

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, Stefanie Jegelka
CSAIL & LIDS, Massachusetts Institute of Technology
Cambridge, MA, USA
{joshrob, cychuang, suvrit, stefje}@mit.edu

Abstract

We consider the question: how can you sample good negative examples for contrastive learning? We argue that, as with metric learning, learning contrastive representations benefits from hard negative samples (i.e., points that are difficult to distinguish from an anchor point). The key challenge toward using hard negatives is that contrastive methods must remain unsupervised, making it infeasible to adopt existing negative sampling strategies that use label information. In response, we develop a new class of unsupervised methods for selecting hard negative samples where the user can control the amount of hardness. A limiting case of this sampling results in a representation that tightly clusters each class, and pushes different classes as far apart as possible. The proposed method improves the generalization of visual representations, requires only few additional lines of code to implement, and introduces no computational overhead.

1 Introduction

Owing to their empirical success, contrastive learning methods [7, 16] have become one of the most popular self-supervised approaches for learning representations [34, 46, 4]. For instance, on certain computer vision tasks unsupervised contrastive learning methods have even outperformed supervised pre-training [30, 20]. Contrastive learning relies on two key ingredients: notions of similar (positive) (x, x^+) and dissimilar (negative) (x, x^-) pairs of data points. The training objective, typically *noise-contrastive estimation* [15], guides the learned representation f to map positive pairs nearby, and negative pairs farther apart. The success of such methods depends critically on the informativeness of the positive and negative pairs, whose selection is challenging without true label supervision.

Much research effort has addressed sampling strategies for positive pairs [2, 53, 47]. For image data, positive sampling strategies often apply transformations that preserve semantic content, e.g., jittering, random cropping, separating color channels, etc. [4, 6, 46]. Such transformations have also been effective in learning control policies from raw pixel data [42], and other positive sampling techniques have been proposed for sentence, audio, and video data [28, 34, 37, 40].

Surprisingly, the choice of negative pairs has drawn much less attention in contrastive learning. Often, given an "anchor" point x, a "negative" x^- is simply sampled uniformly from the training data, independent of how informative it may be. In supervised and metric learning settings, "hard" (negative) examples - intuitively those point that are mapped nearby to an anchor but should be far apart - have been shown to help an embedding learn to correct mistakes more quickly [38, 41].

With this motivation, we address the challenge of selecting informative negatives for contrastive representation learning. We propose a solution that builds a tunable sampling distribution that prefers negative pairs whose representations are currently very similar. This solution faces two challenges: (1) we do not have access to any true label or dissimilarity information; and (2) we need an efficient sampling strategy for this tunable distribution. We overcome (1) by building on ideas from positive-unlabeled learning [13, 12], and (2) by designing an efficient, easy to implement importance sampling technique that incurs no computational overhead.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Contributions. In summary, we make the following contributions:

1. We formulate a method for sampling hard negative pairs, and an efficient sampling strategy that takes into account the lack of true dissimilarity information;
2. We theoretically analyze the properties of the objective and optimal representations, showing that they capture desired goals for representations;
3. Empirical study showing improved generalization of visual representations.

2 Hard Negative Sampling

2.1 Contrastive Learning Setup

We wish to learn an embedding $f : \mathcal{X} \to \mathbb{S}^{d-1}/t$ that maps an observation x to a point on a hypersphere $\mathbb{S}^{d-1}/t$ in $\mathbb{R}^d$ of radius $1/t$, where t is the "temperature" scaling hyperparameter. Following the setup of [1], we assume an underlying set of discrete latent classes $\mathcal{C}$ that represent semantic content, so that similar pairs (x, x^+) have the same latent class. Denoting the distribution over latent classes by $\rho(c)$ for $c \in \mathcal{C}$, we define the joint distribution $p_{x,c}(x, c) = p(x|c)\rho(c)$ whose marginal $p(x)$ we refer to simply as p, and assume $\text{supp}(p) = \mathcal{X}$. For simplicity, we assume $\rho(c) = \tau^+$ is uniform, and let $\tau^- = 1 - \tau^+$ be the probability of another class. Since the class-prior τ^+ is unknown in practice, it must either be treated as a hyperparameter, or estimated [8, 23].

Let $h : \mathcal{X} \to \mathcal{C}$ be the true underlying hypothesis that assigns class labels to inputs. We write $x \sim x'$ to denote the label equivalence relation $h(x) = h(x')$. We denote by $p_x^+(x') = p(x'|h(x') = h(x))$, the distribution over points with same label as x, and by $p_x^-(x') = p(x'|h(x') \neq h(x))$, the distribution over points with labels different from x. We drop the subscript x when the context is clear. For each data point $x \sim p$, the noise-contrastive estimation (NCE) objective [15] for learning the representation f uses a *positive* example x^+ with the same label as x, and *negative* examples $\{x_i^-\}_{i=1}^N$ sampled from q (which is frequently chosen to be the marginal distribution p):

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

2.2 Proposed Hard Sampling Method

Our first goal is to design a distribution q over $\mathcal{X}$ that is allowed to depend on the embedding f and the anchor x from which to sample a batch of negatives $\{x_i^-\}_{i=1}^N$. We propose sampling negatives from the distribution q_β^- defined as

$$q_\beta^-(x^-) := q_\beta(x^-|h(x) \neq h(x^-)), \quad \text{where} \quad q_\beta(x^-) \propto e^{\beta f(x)^\top f(x^-)} \cdot p(x^-),$$

for $\beta \geq 0$. There are two key components to q_β^-, corresponding to two fundamenal principles: 1) conditioning on the event $\{h(x) \neq h(x^-)\}$ which guarantees that (x, x^-) correspond to different latent classes; 2) the concentration parameter β term controls the degree by which q_β up-weights points x^- that have large inner product (similarity) to the anchor x. Since f lies on the surface of a hypersphere of radius $1/t$, we have $\|f(x) - f(x')\|^2 = 2/t^2 - 2f(x)^\top f(x')$ so preferring points with large inner product is equivalent to preferring points with small squared Euclidean distance.

Although we have designed q_β^- to have two desirable properties, it is not clear how to sample efficiently from it. To work towards a practical method, note that we can rewrite this distribution by adopting a PU-learning viewpoint [13, 12, 9]. That is, by conditioning on the event $\{h(x) = h(x^-)\}$ we can split $q_\beta(x^-)$ as $q_\beta(x^-) = \tau^- q_\beta^-(x^-) + \tau^+ q_\beta^+(x^-)$, where $q_\beta^+(x^-) = q_\beta(x^-|h(x) = h(x^-)) \propto e^{\beta f(x)^\top f(x^-)} \cdot p^+(x^-)$. Rearranging yields a formula $q_\beta^-(x^-) = \big(q_\beta(x^-) - \tau^+ q_\beta^+(x^-)\big)/\tau^-$ for the negative sampling distribution q_β^- in terms of two distributions that are tractable since we have samples from p and can approximate samples from p^+ using a set of semantics-preserving transformations, as is typical in contrastive learning methods.

It is possible to generate samples from q_β and (approximately from) q_β^+ using rejection sampling. However, rejection sampling involves adding an additional step to sampling batches, which would involve additional forward passes through f and could be slow. Instead note that fixing the number

Q, taking the limit $N \to \infty$, and inserting $q = q_\beta^- = (q_\beta - \tau^+ q_\beta^+)/\tau^-$ into the objective (1) yields the following hardness-biased objective:

$$\mathop{\mathbb{E}}_{\substack{x\sim p \\ x^+\sim p_x^+}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + \frac{Q}{\tau^-}(\mathbb{E}_{x^-\sim q_\beta}[e^{f(x)^T f(x^-)}] - \tau^+ \mathbb{E}_{v\sim q_\beta^+}[e^{f(x)^T f(v)}])} \right]. \tag{2}$$

This objective suggests that we need only to approximate *expectations* $\mathbb{E}_{x^-\sim q_\beta}[e^{f(x)^T f(x^-)}]$ and $\mathbb{E}_{v\sim q_\beta^+}[e^{f(x)^T f(v)}]$ over q_β and q_β^+ (rather than explicily sampling). This can be achieved using classical Monte-Carlo importance sampling techniques using samples from p and p^+ as follows:

$$\mathbb{E}_{x^-\sim q_\beta}[e^{f(x)^T f(x^-)}] = \mathbb{E}_{x^-\sim p}[e^{f(x)^T f(x^-)} q_\beta / p] = \mathbb{E}_{x^-\sim p}[e^{(\beta+1) f(x)^T f(x^-)}/Z_\beta],$$
$$\mathbb{E}_{v\sim q_\beta^+}[e^{f(x)^T f(v)}] = \mathbb{E}_{v\sim p^+}[e^{f(x)^T f(v)} q_\beta^+ / p^+] = \mathbb{E}_{v\sim p^+}[e^{(\beta+1) f(x)^T f(v)}/Z_\beta^+],$$

where Z_β, Z_β^+ are the partition functions of q_β and q_β^+ respectively. The right hand terms readily admit empirical approximations by replacing p and p^+ with empirical versions using samples $\{x_i^-\}_{i=1}^N$ and $\{x_i^+\}_{i=1}^M$ respectively. The only unknowns left are the partition functions, Z_β, Z_β^+, which themselves are expectations over p and p^+, so also admit empirical estimates using $\{x_i^-\}_{i=1}^N$ and $\{x_i^+\}_{i=1}^M$.

3 Analysis of Hard Negative Sampling

3.1 Hard Sampling Interpolates Between Marginal and Worst-Case Negatives

Intuitively, the concentration parameter β controls the level of "hardness" of the negative samples. It is clear that taking $\beta = 0$ (the "least hard" negatives) recovers the population distribution $q_0 = p$ (and q_0^- recovers the debiasing idea of [9]). But what interpretation does large β admit? Specifically, what does the distribution q_β^- converge to in the limit $\beta \to \infty$, if anything? It turns out that in the limit q_β^- approximates an inner solution to the following zero-sum two player game.

$$\inf_f \sup_{q\in\Pi} \left\{ \mathcal{L}(f,q) = \mathop{\mathbb{E}}_{\substack{x\sim p \\ x^+\sim p_x^+}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + Q\mathbb{E}_{x^-\sim q}[e^{f(x)^T f(x^-)}]} \right] \right\}. \tag{3}$$

where $\Pi = \{q = q(\cdot; x, f) : \operatorname{supp}\big(q(\cdot; x, f)\big) \subseteq \{x' \in \mathcal{X} : x' \nsim x\}, \forall x \in \mathcal{X}\}$ is the set of distributions with support that is disjoint from points with the same class as x (without loss of generality we assume $\{x' \in \mathcal{X} : x' \nsim x\}$ is non-empty). Since $q = q(\cdot; x, f)$ depends on x and f it can be thought of as a family of distributions. The formal statement is as follows.

Proposition 1. *Let $\mathcal{L}^*(f) = \sup_{q\in\Pi} \mathcal{L}(f,q)$. Then for any $t > 0$ and $f : \mathcal{X} \to \mathbb{S}^{d-1}/t$ we observe the convergence $\mathcal{L}(f, q_\beta^-) \longrightarrow \mathcal{L}^*(f)$ as $\beta \to \infty$.*

3.2 Optimal Embeddings on the Hypersphere for Worst-Case Negative Samples

What desirable properties does an optimal contrastive embedding (global minimizer of $\mathcal{L}$) possess that make the representation generalizable? To study this question, we analyze the distribution of an optimal embedding f^* on the hypersphere for adversarial worst-case negatives. Following the formulation of [50] we consider the following alternate viewpoint of objective (1): take $Q = N$ in (1), and subtract $\log N$. This changes neither the set of minimizers, nor the geometry of the loss surface. Taking the number of negative samples $N \to \infty$ yields the limiting objective,

$$\mathcal{L}_\infty(f,q) = \mathop{\mathbb{E}}_{\substack{x\sim p \\ x^+\sim p_x^+}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{\mathbb{E}_{x^-\sim q}[e^{f(x)^T f(x^-)}]} \right]. \tag{4}$$

Theorem 2. *Suppose $\mathcal{C}$ is finite (classification task), and let $\mathcal{L}_\infty^*(f) = \sup_{q\in\Pi} \mathcal{L}_\infty(f,q)$. Then $\inf_{f:\text{ measurable}} \mathcal{L}_\infty^*(f)$ is attained, and any global minimizer f^* is such that $f^*(x) = f^*(x^+)$ almost surely. Furthermore, letting $\mathbf{v}_c = f^*(x)$ for any x such that $h(x) = c$ (so $\mathbf{v}_c$ is well defined up to a set of x of measure zero), f^* is characterized as any solution to the ball-packing problem,*

$$\max_{\{\mathbf{v}_c \in \mathbb{S}^{d-1}/t\}_{c\in\mathcal{C}}} \sum_{c\in\mathcal{C}} \rho(c) \cdot \min_{c'\neq c} \|\mathbf{v}_c - \mathbf{v}_{c'}\|^2. \tag{5}$$

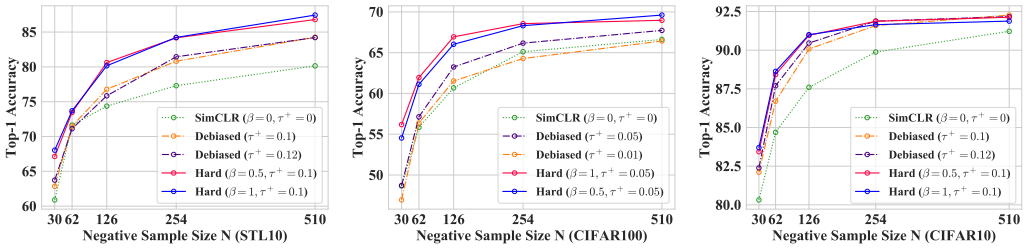

Figure 1: **Classification accuracy on downstream tasks.** Embeddings trained using hard ($\beta > 0$), debiased ($\beta = 0$), and standard ($\beta = 0, \tau^+ = 0$) versions of SimCLR, and evaluated using linear readout accuracy.

Interpretation. The first component of the result states that an optimal embedding f^* must be invariant across pairs of similar inputs x, x^+. The second component characterizes solutions as solutions to the classical Tammes Ball-Packing Problem from geometry [45, 39, 32, 45] (Eq. 5). When the distribution ρ over classes is uniform this problem is solved by a set of $|\mathcal{C}|$ points on the hypersphere such that the average squared-ℓ_2 distance from a point to the nearest other point is as large as possible. Non-uniform ρ adds importance weights to each fixed ball. This interpretation shows that solutions of the problem $\min_f \mathcal{L}^*_\infty(f)$ are a type of maximum margin clustering.

4 Empirical Study

We test the hard sampling method on vision tasks using STL10, CIFAR100 and CIFAR10. We use SimCLR [4] as the baseline method, use one positive example x^+ per data point and treat the class-prior τ^+ as a hyperparameter, and keep the number of positive samples $M = 1$. The results in Figure 1 show consistent improvement over SimCLR ($\beta = 0, \tau^+ = 0$) and "debiased" ($\beta = 0$) baselines [9] on STL10 and CIFAR100. For $N = 510$ negative examples per data point we observe absolute improvements over the best debiased baseline of 1.9% and 3.2%, respectively, on CIFAR100 and STL10, and absolute improvements of 3% and 7.3% over SimCLR. On CIFAR10 there is a slight improvement for smaller N, which disappears at larger N. See Appendix C for full setup details.

Are harder samples necessarily better? By setting β to large values, one can focus on only the hardest samples in a training batch. But is this desirable? Figure 2 (left, middle) shows that for vision problems, taking larger β does not necessarily lead to better representations. In contrast, when one uses true positive pairs during training (green curve, uses label information for positive but not negative pairs), the downstream performance monotonically increases with β until convergence (Figure 2 , middle).

Does debiasing improve hard sampling? Our proposed hard negative sampling method uses debiasing (i.e., conditioning on the event $\{h(x) \neq h(x^-)\}$ [9]). But does debiasing help? To test this, we train four embeddings: hard sampling with and without debiasing, and uniform sampling ($\beta = 0$) with and without debiasing. The results in Figure 2 (right) show that hard sampling with debiasing obtains the highest linear readout accuracy on STL10, only using hard sampling or only debiasing yields (in this case) similar accuracy. All improve over the SimCLR baseline. See Appendix B for further ablation studies.

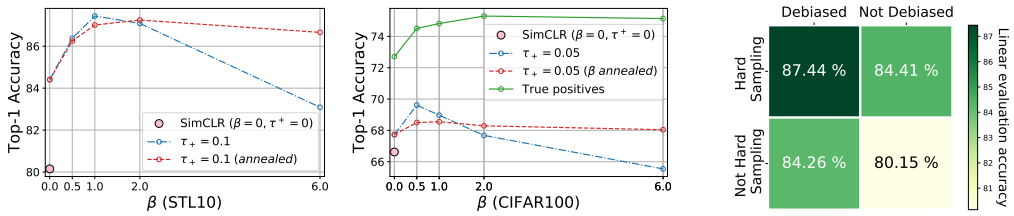

Figure 2: Embeddings evaluated with linear readout accuracy. Left & middle: the effect of varying concentration parameter β for: contrastive learning (fully unsupervised), using true positive samples (uses label information), and an annealing method that improves robustness to the choice of β (see Appendix C for details). Right: STL10 for hard sampling with and without debiasing, and non-hard sampling ($\beta = 0$) with and without debiasing. Best results come from using both simultaneously.

References

[1] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. In *Int. Conference on Machine Learning (ICML)*, pages 5628–5637, 2019.

[2] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

[3] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Int. Conference on Machine Learning (ICML)*, pages 10709–10719, 2020.

[5] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv:2006.10029*, 2020.

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv:2003.04297*, 2020.

[7] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 539–546, 2005.

[8] Marthinus Christoffel, Gang Niu, and Masashi Sugiyama. Class-prior estimation for learning from positive and unlabeled data. In *Asian Conference on Machine Learning*, pages 221–236, 2016.

[9] Ching-Yao Chuang, Joshua Robinson, Lin Yen-Chen, Antonio Torralba, and Stefanie Jegelka. Debiased Contrastive Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[10] Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine learning*, 20(3): 273–297, 1995.

[11] Bill Dolan, Chris Quirk, and Chris Brockett. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*, page 350. Association for Computational Linguistics, 2004.

[12] Marthinus C Du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 703–711, 2014.

[13] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–220, 2008.

[14] Weifeng Ge. Deep metric learning with hierarchical triplet loss. In *Europ. Conference on Computer Vision (ECCV)*, pages 269–285, 2018.

[15] Michael Gutmann and Aapo Hyvärinen. Noise-Contrastive Estimation: A new estimation principle for unnormalized statistical models. In *Proc. Int. Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 297–304, 2010.

[16] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1735–1742, 2006.

[17] Ben Harwood, Vijay Kumar BG, Gustavo Carneiro, Ian Reid, and Tom Drummond. Smart mining for deep metric learning. In *Int. Conference on Computer Vision (ICCV)*, pages 2821–2829, 2017.

[18] Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *Int. Conference on Machine Learning (ICML)*, pages 3451–3461, 2020.

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

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020.

[21] Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. In *Int. Conference on Machine Learning (ICML)*, pages 6661–6671, 2020.

[22] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004.

[23] Shantanu Jain, Martha White, and Predrag Radivojac. Estimating the class prior and posterior from noisy positives and unlabeled data. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2693–2701, 2016.

[24] SouYoung Jin, Aruni RoyChowdhury, Huaizu Jiang, Ashish Singh, Aditya Prasad, Deep Chakraborty, and Erik Learned-Miller. Unsupervised hard example mining from videos for improved object detection. In *Europ. Conference on Computer Vision (ECCV)*, pages 307–324, 2018.

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Int. Conf. on Learning Representations (ICLR)*, 2015.

[26] Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-Thought Vectors. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3294–3302, 2015.

[27] Yujia Li, Chenjie Gu, Thomas Dullien, Oriol Vinyals, and Pushmeet Kohli. Graph matching networks for learning the similarity of graph structured objects. In *Int. Conference on Machine Learning (ICML)*, pages 3835–3845, 2019.

[28] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. In *Int. Conf. on Learning Representations (ICLR)*, 2018.

[29] K. V. Mardia and P. Jupp. *Directional Statistics*. John Wiley and Sons Ltd., second edition, 2000.

[30] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6707–6717, 2020.

[31] Christopher Morris, Nils M Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. In *Graph Representation Learning and Beyond, ICML Workshop*, 2020.

[32] Oleg R Musin and Alexey S Tarasov. The tammes problem for n= 14. *Experimental Mathematics*, 24(4):460–468, 2015.

[33] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 271–279, 2016.

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

[35] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

[36] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 115–124. Association for Computational Linguistics, 2005.

[37] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *arXiv:2007.13916*, 2020.

[38] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[39] K Schütte and BL Van der Waerden. Auf welcher kugel haben 5, 6, 7, 8 oder 9 punkte mit mindestabstand eins platz? *Mathematische Annalen*, 123(1):96–124, 1951.

[40] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1134–1141, 2018.

[41] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4004–4012, 2016.

[42] Aravind Srinivas, Michael Laskin, and Pieter Abbeel. CURL: Contrastive unsupervised representations for reinforcement learning. In *Int. Conference on Machine Learning (ICML)*, pages 10360–10371, 2020.

[43] Yumin Suh, Bohyung Han, Wonsik Kim, and Kyoung Mu Lee. Stochastic class-based hard example mining for deep metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7251–7259, 2019.

[44] Fan-Yun Sun, Jordan Hoffmann, Vikas Verma, and Jian Tang. InfoGraph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *Int. Conf. on Learning Representations (ICLR)*, 2020.

[45] Pieter Merkus Lambertus Tammes. On the origin of number and arrangement of the places of exit on the surface of pollen-grains. *Recueil des travaux botaniques néerlandais*, 27(1):1–84, 1930.

[46] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive Multiview Coding. In *Europ. Conference on Computer Vision (ECCV)*, pages 770–786, 2019.

[47] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *arXiv:2005.10243*, 2020.

[48] Petar Velickovic, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep Graph Infomax. In *Int. Conf. on Learning Representations (ICLR)*, 2019.

[49] Ellen M Voorhees and Donna Harman. Overview of trec 2002. 2002.

[50] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Int. Conference on Machine Learning (ICML)*, pages 9574–9584, 2020.

[51] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210, 2005.

[52] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Int. Conference on Computer Vision (ICCV)*, pages 2840–2848, 2017.

[53] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv:1304.5634*, 2013.

[54] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *Int. Conf. on Learning Representations (ICLR)*, 2019.

A Analysis of Hard Sampling

A.1 Hard Sampling Interpolates Between Marginal and Worst-Case Negatives

To develop a better intuitive understanding of the worst case negative distribution objective $\mathcal{L}^*(f) = \sup_{q\in\Pi} \mathcal{L}(f,q)$, we note that the supremum can be characterized analytically. Indeed,

$$\begin{aligned}\sup_{q\in\Pi} \mathcal{L}(f,q) &= -\mathbb{E}_{\substack{x\sim p\\ x^+\sim p_x^+}} f(x)^T f(x^+) + \sup_{q\in\Pi} \mathbb{E}_{\substack{x\sim p\\ x^+\sim p_x^+}} \log\left\{e^{f(x)^T f(x^+)} + Q\mathbb{E}_{x^-\sim q}[e^{f(x)^T f(x^-)}]\right\}\\ &= -\mathbb{E}_{\substack{x\sim p\\ x^+\sim p_x^+}} f(x)^T f(x^+) + \mathbb{E}_{\substack{x\sim p\\ x^+\sim p_x^+}} \log\left\{e^{f(x)^T f(x^+)} + Q\cdot\sup_{q\in\Pi}\mathbb{E}_{x^-\sim q}[e^{f(x)^T f(x^-)}]\right\}.\end{aligned}$$

The supremum over q can be pushed inside the expectation since q is a family of distribution indexed by x, reducing the problem to maximizing $\mathbb{E}_{x^-\sim q}[e^{f(x)^T f(x^-)}]$, which is solved by any q^* whose support is a subset of $\arg\sup_{x^-:x^-\nsim x} e^{f(x)^T f(x^-)}$ if the supremum is attained. If it is not attained, distributions supported on approximations of the supremum yield approximate solutions to $\sup_{q\in\Pi} \mathcal{L}(f,q)$. However, computing such points involves maximizing a neural network. Instead of taking this challenging route, q_β represents a tractable approximation for large β (Proposition 1).

Next we prove Proposition 1. Recall that the proposition stated the following.

Proposition 3. *Let $\mathcal{L}^*(f) = \sup_{q\in\Pi} \mathcal{L}(f,q)$. Then for any $t > 0$ and measurable $f : \mathcal{X} \to \mathbb{S}^{d-1}/t$ we observe the convergence $\mathcal{L}(f, q_\beta^-) \longrightarrow \mathcal{L}^*(f)$ as $\beta \to \infty$.*

Proof. Consider the following essential supremum,

$$M(x) = \operatorname*{ess\,sup}_{x^-\in\mathcal{X}:x^-\nsim x} f(x)^T f(x^-) = \sup\{m > 0 : m \geq f(x)^T f(x^-) \text{ a.s. for } x^- \sim p^-\}.$$

The second inequality holds since $\operatorname{supp}(p) = \mathcal{X}$. We may rewrite

$$\begin{aligned}\mathcal{L}^*(f) &= \mathbb{E}_{\substack{x\sim p\\ x^+\sim p_x^+}}\left[-\log\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + Qe^{M(x)}}\right],\\ \mathcal{L}(f, q_\beta^-) &= \mathbb{E}_{\substack{x\sim p\\ x^+\sim p_x^+}}\left[-\log\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + Q\mathbb{E}_{x^-\sim q_\beta^-}[e^{f(x)^T f(x^-)}]}\right].\end{aligned}$$

The difference between these two terms can be bounded as follows,

$$\begin{aligned}\left|\mathcal{L}^*(f) - \mathcal{L}(f, q_\beta^-)\right| &\leq \mathbb{E}_{\substack{x\sim p\\ x^+\sim p_x^+}}\left|-\log\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + Qe^{M(x)}} + \log\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + Q\mathbb{E}_{x^-\sim q_\beta^-}[e^{f(x)^T f(x^-)}]}\right|\\ &= \mathbb{E}_{\substack{x\sim p\\ x^+\sim p_x^+}}\left|\log\left(e^{f(x)^T f(x^+)} + Q\mathbb{E}_{x^-\sim q_\beta^-}[e^{f(x)^T f(x^-)}]\right) - \log\left(e^{f(x)^T f(x^+)} + Qe^{M(x)}\right)\right|\\ &\leq \frac{e^{1/t}}{Q+1}\cdot\mathbb{E}_{\substack{x\sim p\\ x^+\sim p_x^+}}\left|e^{f(x)^T f(x^+)} + Q\mathbb{E}_{x^-\sim q_\beta^-}[e^{f(x)^T f(x^-)}] - e^{f(x)^T f(x^+)} - Qe^{M(x)}\right|\\ &= \frac{e^{1/t}Q}{Q+1}\cdot\mathbb{E}_{x\sim p}\left|\mathbb{E}_{x^-\sim q_\beta^-}[e^{f(x)^T f(x^-)}] - e^{M(x)}\right|\\ &\leq e^{1/t}\cdot\mathbb{E}_{x\sim p}\mathbb{E}_{x^-\sim q_\beta^-}\left|e^{M(x)} - e^{f(x)^T f(x^-)}\right|\end{aligned}$$

where for the second inequality we have used the fact that f lies on the hypersphere of radius $1/t$ to restrict the domain of the logarithm to values greater than $(Q+1)e^{-1/t}$. Because of this the logarithm is Lipschitz with parameter $e^{1/t}/(Q+1)$. Using again the fact that f lies on the hypersphere we know that $\left|f(x)^T f(x^-)\right| \leq 1/t^2$ and hence have the following inequality,

$$\mathbb{E}_{x\sim p}\mathbb{E}_{q_\beta^-}\left|e^{M(x)} - e^{f(x)^T f(x^-)}\right| \leq e^{1/t^2}\mathbb{E}_{x\sim p}\mathbb{E}_{q_\beta^-}\left|M(x) - f(x)^T f(x^-)\right|$$

Let us consider the inner expectation $E_\beta(x) = \mathbb{E}_{q_\beta^-}\left|M(x) - f(x)^T f(x^-)\right|$. Note that since f is bounded, $E_\beta(x)$ is uniformly bounded in x. Therefore, in order to show the convergence $\mathcal{L}(f, q_\beta^-) \to \mathcal{L}^*(f)$ as $\beta \to \infty$, it suffices by the dominated convergence theorem to show that $E_\beta(x) \to 0$ pointwise as $\beta \to \infty$ for arbitrary fixed $x \in \mathcal{X}$.

From now on we denote $M = M(x)$ for brevity, and consider a fixed $x \in \mathcal{X}$. From the definition of q_β^- it is clear that $q_\beta^- \ll p^-$. That is, since $q_\beta^- = c \cdot p^-$ for some (non-constant) c, it is absolutely continuous with respect to p^-. So $M(x) \geq f(x)^T f(x^-)$ almost surely for $x^- \sim q_\beta^-$, and we may therefore drop the absolute value signs from our expectation. Define the following event $\mathcal{G}_\varepsilon = \{x^- : f(x)^\top f(x^-) \geq M - \varepsilon\}$ where $\mathcal{G}$ is refers to a "good" event. Define its complement $\mathcal{B}_\varepsilon = \mathcal{G}_\varepsilon^c$ where $\mathcal{B}$ is for "bad". For a fixed $x \in \mathcal{X}$ and $\varepsilon > 0$ consider,

$$\begin{aligned}
E_\beta(x) &= \mathbb{E}_{x^-\sim q_\beta^-}\left|M(x) - f(x)^T f(x^-)\right| \\
&= \mathbb{P}_{x^-\sim q_\beta^-}(\mathcal{G}_\varepsilon)\cdot \mathbb{E}_{x^-\sim q_\beta^-}\left[\left|M(x) - f(x)^T f(x^-)\right| |\mathcal{G}_\varepsilon\right] + \mathbb{P}_{x^-\sim q_\beta^-}(\mathcal{B}_\varepsilon)\cdot \mathbb{E}_{x^-\sim q_\beta^-}\left[\left|M(x) - f(x)^T f(x^-)\right| |\mathcal{B}_\varepsilon\right] \\
&\leq \mathbb{P}_{x^-\sim q_\beta^-}(\mathcal{G}_\varepsilon)\cdot \varepsilon + 2\mathbb{P}_{x^-\sim q_\beta^-}(\mathcal{B}_\varepsilon) \\
&\leq \varepsilon + 2\mathbb{P}_{x^-\sim q_\beta^-}(\mathcal{B}_\varepsilon).
\end{aligned}$$

We need to control $\mathbb{P}_{x^-\sim q_\beta^-}(\mathcal{B}_\varepsilon)$. Expanding,

$$\mathbb{P}_{x^-\sim q_\beta^-}(\mathcal{B}_\varepsilon) = \int_{\mathcal{X}} \mathbf{1}\left\{f(x)^T f(x^-) < M(x) - \varepsilon\right\} \frac{e^{\beta f(x)^T f(x^-)}\cdot p^-(x^-)}{Z_\beta}\mathrm{d}x^-$$

where $Z_\beta = \int_{\mathcal{X}} e^{\beta f(x)^T f(x^-)} p^-(x^-)\mathrm{d}x^-$ is the partition function of q_β^-. We may bound this expression by,

$$\begin{aligned}
\int_{\mathcal{X}} \mathbf{1}\left\{f(x)^T f(x^-) < M - \varepsilon\right\} \frac{e^{\beta(M-\varepsilon)}\cdot p^-(x^-)}{Z_\beta}\mathrm{d}x^- &\leq \frac{e^{\beta(M-\varepsilon)}}{Z_\beta}\int_{\mathcal{X}} \mathbf{1}\left\{f(x)^T f(x^-) < M - \varepsilon\right\} p^-(x^-)\mathrm{d}x^- \\
&= \frac{e^{\beta(M-\varepsilon)}}{Z_\beta}\mathbb{P}_{x^-\sim p^-}(\mathcal{B}_\varepsilon) \\
&\leq \frac{e^{\beta(M-\varepsilon)}}{Z_\beta}
\end{aligned}$$

Note that

$$Z_\beta = \int_{\mathcal{X}} e^{\beta f(x)^T f(x^-)} p^-(x^-)\mathrm{d}x^- \geq e^{\beta(M-\varepsilon/2)}\mathbb{P}_{x^-\sim p^-}(f(x)^T f(x^-) \geq M - \varepsilon/2).$$

By the definition of $M = M(x)$ the probability $\rho_\varepsilon = \mathbb{P}_{x^-\sim p^-}(f(x)^T f(x^-) \geq M - \varepsilon/2) > 0$, and we may therefore bound,

$$\begin{aligned}
\mathbb{P}_{x^-\sim q_\beta^-}(\mathcal{B}_\varepsilon) &= \frac{e^{\beta(M-\varepsilon)}}{e^{\beta(M-\varepsilon/2)}\rho_\varepsilon} \\
&= e^{-\beta\varepsilon/2}/\rho_\varepsilon \\
&\longrightarrow 0 \ \text{ as } \ \beta \to \infty.
\end{aligned}$$

We may therefore take β to be sufficiently big so as to make $\mathbb{P}_{x^-\sim q_\beta^-}(\mathcal{B}_\varepsilon) \leq \varepsilon$ and therefore $E_\beta(x) \leq 3\varepsilon$. In other words, $E_\beta(x) \longrightarrow 0$ as $\beta \to \infty$. □

A.2 Optimal Embeddings on the Hypersphere for Worst-Case Negative Samples

In order to study properties of global optima of the contrastive objective using the adversarial worst case hard sampling distribution recall that we have the following limiting objective,

$$\mathcal{L}_\infty(f,q) = \mathbb{E}_{\substack{x\sim p \\ x^+\sim p_x^+}}\left[-\log \frac{e^{f(x)^T f(x^+)}}{\mathbb{E}_{x^-\sim q_\beta}[e^{f(x)^T f(x^-)}]}\right]. \tag{6}$$

We may separate the logarithm of a quotient into the sum of two terms plus a constant,

$$\mathcal{L}_\infty(f,q) = \mathcal{L}_{\text{align}}(f) + \mathcal{L}_{\text{unif}}(f,q) - 1/t^2$$

where $\mathcal{L}_{\text{align}}(f) = \mathbb{E}_{x,x^+}\|f(x) - f(x^+)\|^2/2$ and $\mathcal{L}_{\text{unif}}(f,q) = \mathbb{E}_{x\sim p}\log \mathbb{E}_{x^-\sim q} e^{f(x)^\top f(x^-)}$. Here we have used the fact that f lies on the boundary of the hypersphere of radius $1/t$, which gives us the following equivalence between inner products and squared Euclidean norm,

$$2/t^2 - 2f(x)^\top f(x^+) = \|f(x)\|^2 + \|f(x^+)\|^2 - 2f(x)^\top f(x^+) = \|f(x) - f(x^+)\|^2. \tag{7}$$

Taking supremum to obtain $\mathcal{L}^*_\infty(f) = \sup_{q\in\Pi}\mathcal{L}_\infty(f,q)$ we find that the second expression simplifies to,

$$\mathcal{L}^*_{\text{unif}}(f) = \sup_{q\in\Pi}\mathcal{L}_{\text{unif}}(f,q) = \mathbb{E}_{x\sim p}\log \sup_{x^-\nsim x} e^{f(x)^\top f(x^-)} = \mathbb{E}_{x\sim p}\sup_{x^-\nsim x} f(x)^\top f(x^-).$$

Using Eqn. (7), this can be re-expressed as,

$$\mathbb{E}_{x\sim p}\sup_{x^-\nsim x} f(x)^\top f(x^-) = -\mathbb{E}_{x\sim p}\inf_{x^-\nsim x}\|f(x) - f(x^+)\|^2/2 + 1/t^2. \tag{8}$$

The forthcoming theorem exactly characterizes the global optima of $\min_f \mathcal{L}^*_\infty(f)$

Theorem 4. *Suppose the downstream task is classification (i.e. $\mathcal{C}$ is finite), and let $\mathcal{L}^*_\infty(f) = \sup_{q\in\Pi}\mathcal{L}_\infty(f,q)$. The infimum $\inf_{f:\text{ measurable}}\mathcal{L}^*_\infty(f)$ is attained, and any f^* achieving the global minimum is such that $f^*(x) = f^*(x^+)$ almost surely. Furthermore, letting $\mathbf{v}_c = f^*(x)$ for any x such that $h(x) = c$ (so $\mathbf{v}_c$ is well defined up to a set of x of measure zero), f^* is characterized as being any solution to the following ball-packing problem,*

$$\max_{\{\mathbf{v}_c\in\mathbb{S}^{d-1}/t\}_{c\in\mathcal{C}}}\sum_{c\in\mathcal{C}}\rho(c)\cdot\min_{c'\neq c}\|\mathbf{v}_c - \mathbf{v}_{c'}\|^2. \tag{9}$$

Proof. Any minimizer of $\mathcal{L}_{\text{align}}(f)$ has the property that $f(x) = f(x^+)$ almost surely. So, in order to prove the first claim, it suffices to show that there exist functions $f\in\arg\inf_f\mathcal{L}^*_{\text{unif}}(f)$ for which $f(x) = f(x^+)$ almost surely. This is because, at that point, we have shown that $\arg\min_f\mathcal{L}_{\text{align}}(f)$ and $\arg\min_f\mathcal{L}^*_{\text{unif}}(f)$ intersect, and therefore any solution of $\mathcal{L}^*_\infty(f) = \mathcal{L}_{\text{align}}(f) + \mathcal{L}^*_{\text{unif}}(f)$ must lie in this intersection.

To this end, suppose that $f\in\arg\min_f\mathcal{L}^*_{\text{unif}}(f)$ but that $f(x)\neq f(x^+)$ with non-zero probability. We shall show that we can construct a new embedding $\hat{f}$ such that $f(x) = f(x^+)$ almost surely, and $\mathcal{L}^*_{\text{unif}}(\hat{f}) \leq \mathcal{L}^*_{\text{unif}}(f)$. Due to Eqn. (8) this last condition is equivalent to showing,

$$\mathbb{E}_{x\sim p}\inf_{x^-\nsim x}\|\hat{f}(x) - \hat{f}(x^-)\|^2 \geq \mathbb{E}_{x\sim p}\inf_{x^-\nsim x}\|f(x) - f(x^-)\|^2. \tag{10}$$

Fix a $c\in\mathcal{C}$, and let $x_c\in\arg\max_{x:h(x)=c}\inf_{x^-\nsim x}\|f(x) - f(x^-)\|^2$. Then, define $\hat{f}(x) = f(x_c)$ for any x such that $h(x) = c$ and $\hat{f}(x) = f(x)$ otherwise. Let us first aim to show that Eqn. (10) holds for this $\hat{f}$. Let us begin to expand the left hand side of Eqn. (10),

$$
\begin{aligned}
&\mathbb{E}_{x\sim p} \inf_{x^- \nsim x} \|\hat{f}(x) - \hat{f}(x^-)\|^2 \\
&\qquad = \mathbb{E}_{\hat{c}\sim\rho}\mathbb{E}_{x\sim p(\cdot|\hat{c})} \inf_{x^- \nsim x} \|\hat{f}(x) - \hat{f}(x^-)\|^2 \\
&\qquad = \rho(c)\mathbb{E}_{x\sim p(\cdot|c)} \inf_{x^- \nsim x} \|\hat{f}(x) - \hat{f}(x^-)\|^2 \\
&\qquad\qquad + (1-\rho(c))\mathbb{E}_{\hat{c}\sim\rho(\cdot|\hat{c}\neq c)}\mathbb{E}_{x\sim p(\cdot|\hat{c})} \inf_{x^- \nsim x} \|\hat{f}(x) - \hat{f}(x^-)\|^2 \\
&\qquad = \rho(c)\mathbb{E}_{x\sim p(\cdot|c)} \inf_{x^- \nsim x} \|f(x_c) - f(x^-)\|^2 \\
&\qquad\qquad + (1-\rho(c))\mathbb{E}_{\hat{c}\sim\rho(\cdot|\hat{c}\neq c)}\mathbb{E}_{x\sim p(\cdot|\hat{c})} \inf_{x^- \nsim x} \|\hat{f}(x) - \hat{f}(x^-)\|^2 \\
&\qquad = \rho(c) \inf_{x^- \nsim x_c} \|f(x_c) - f(x^-)\|^2 \\
&\qquad\qquad + (1-\rho(c))\mathbb{E}_{\hat{c}\sim\rho(\cdot|\hat{c}\neq c)}\mathbb{E}_{x\sim p(\cdot|\hat{c})} \inf_{h(x^-)\neq\hat{c}} \|\hat{f}(x) - \hat{f}(x^-)\|^2 \qquad (11)
\end{aligned}
$$

By construction, the first term can be lower bounded by $\inf_{x^- \nsim x_c} \|f(x_c) - f(x^-)\|^2 \geq \inf_{x^- \nsim x} \|f(x) - f(x^-)\|^2$ for any x such that $h(x) = c$. To lower bound the second term, consider any fixed $\hat{c} \neq c$ and $x \sim p(\cdot|\hat{c})$ (so $h(x) = \hat{c}$). Define the following two subsets of the input space $\mathcal{X}$

$$
\mathcal{A} = \{f(x^-) : f(x^-) \neq \hat{c} \text{ for } x^- \in \mathcal{X}\} \qquad \widehat{\mathcal{A}} = \{f(x^-) \in \mathcal{X} : \hat{f}(x^-) \neq \hat{c} \text{ for } x^- \in \mathcal{X}\}.
$$

Since by construction the range of $\hat{f}$ is a subset of the range of f, we know that $\widehat{\mathcal{A}} \subseteq \mathcal{A}$. Combining this with the fact that $\hat{f}(x) = f(x)$ whenever $h(x) = \hat{c} \neq c$ we see,

$$
\begin{aligned}
\inf_{h(x^-)\neq\hat{c}} \|\hat{f}(x) - \hat{f}(x^-)\|^2 &= \inf_{h(x^-)\neq\hat{c}} \|f(x) - \hat{f}(x^-)\|^2 \\
&= \inf_{u\in\widehat{\mathcal{A}}} \|f(x) - u\|^2 \\
&\geq \inf_{u\in\mathcal{A}} \|f(x) - u\|^2 \\
&= \inf_{h(x^-)\neq\hat{c}} \|f(x) - f(x^-)\|^2
\end{aligned}
$$

Using these two lower bounds we may conclude that Eqn. (11) can be lower bounded by,

$$
\rho(c) \inf_{x^- \nsim x_c} \|f(x_c) - f(x^-)\|^2 + (1-\rho(c))\mathbb{E}_{\hat{c}\sim\rho(\cdot|\hat{c}\neq c)}\mathbb{E}_{x\sim p(\cdot|\hat{c})} \inf_{h(x^-)\neq\hat{c}} \|f(x) - f(x^-)\|^2
$$

which equals $\mathbb{E}_{x\sim p} \inf_{x^- \nsim x} \|f(x) - f(x^-)\|^2$. We have therefore proved Eqn. (10). To summarize the current progress; given an embedding f we have constructed a new embedding $\hat{f}$ that attains lower $\mathcal{L}_{\text{unif}}$ loss and which is constant on x such that $\hat{f}$ is constant on $\{x : h(x) = c\}$. Enumerating $\mathcal{C} = \{c_1, c_2 \ldots, c_{|\mathcal{C}|}\}$, we may repeatedly apply the same argument to construct a sequence of embeddings $f_1, f_2, \ldots, f_{|\mathcal{C}|}$ such that f_i is constant on each of the following sets $\{x : h(x) = c_j\}$ for $j \leq i$. The final embedding in the sequence $f^* = f_{|\mathcal{C}|}$ is such that $\mathcal{L}^*_{\text{unif}}(f^*) \leq \mathcal{L}^*_{\text{unif}}(f)$ and therefore f^* is a minimizer. This embedding is constant on each of $\{x : h(x) = c_j\}$ for $j = 1, 2, \ldots, |\mathcal{C}|$. In other words, $f^*(x) = f^*(x^+)$ almost surely. We have proved the first claim.

Obtaining the second claim is a matter of manipulating $\mathcal{L}^*_\infty(f^*)$. Indeed, we know that $\mathcal{L}^*_\infty(f^*) = \mathcal{L}^*_{\text{unif}}(f^*) - 1/t^2$ and defining $\mathbf{v}_c = f^*(x)$ for any x such that $h(x) = c$, this expression is minimized if and only if f^* attains,

$$\max_f \mathbb{E}_{x\sim p} \inf_{x^- \nsim x} \|f(x) - f(x^-)\|^2 = \max_f \mathbb{E}_{c\sim\rho}\mathbb{E}_{x\sim p(\cdot|c)} \inf_{h(x^-)\neq c} \|f(x) - f(x^-)\|^2$$
$$= \max_f \sum_{c\in\mathcal{C}} \rho(c) \cdot \inf_{h(x^-)\neq c} \|f(x) - f(x^-)\|^2$$
$$= \max_{\{\mathbf{v}_c \in \mathbb{S}^{d-1}/t\}_{c\in\mathcal{C}}} \sum_{c\in\mathcal{C}} \rho(c) \cdot \min_{c'\neq c} \|\mathbf{v}_c - \mathbf{v}_{c'}\|^2$$

□

B Additional Ablations

Figure 3 compares the histograms of cosine similarities of positive and negative pairs for the four learned representations. The representation trained with hard negatives and debiasing assigns much lower similarity score to a pair of negative samples than other methods. On the other hand, the SimCLR baseline assigns higher cosine similarity scores to pairs of positive samples. However, to discriminate positive and negative pairs, a key property is the amount of *overlap* of positive and negative histograms. Our hard sampling method appears to obtain less overlap than SimCLR, by better trading off higher dissimilarity of negative pairs with less similarity of positive pairs. Similar tradeoffs are observed for the debiased objective, and hard sampling without debiasing.

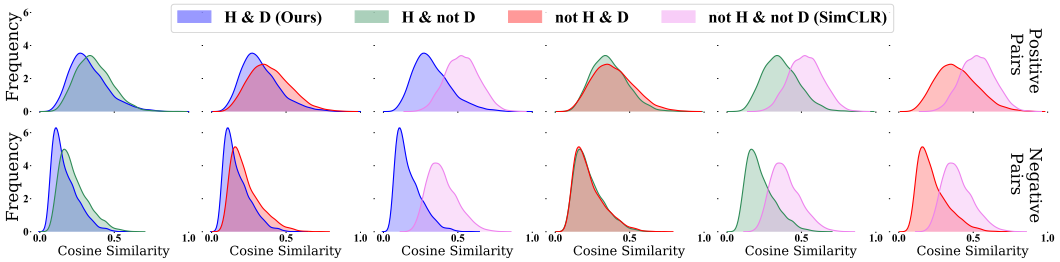

Figure 3: Histograms of cosine similarity of pairs of points with the same label (top) and different labels (bottom) for embeddings trained on STL10 with four different objectives. H=Hard Sampling, D=Debiasing. Histograms overlaid pairwise to allow for convenient comparison.

To study the affect of varying the concentration parameter β on the learned embeddings Figure 4 plots cosine similarity histograms of pairs of similar and dissimilar points. The results show that for β moving from 0 through 0.5 to 2 causes both the positive and negative similarities to gradually skew left. In terms of downstream classification, an important property is the *relative* difference in similarity between positive and negative pairs. In this case $\beta = 0.5$ find the best balance (since it achieves the highest downstream accuracy). When β is taken very large ($\beta = 6$), we see a change in conditions. Both positive and negative pairs are assigned higher similarities in general. Visually it seems that the positive and negative histograms for $\beta = 6$ overlap a lot more than for smaller values, which helps explain why the linear readout accuracy is lower for $\beta = 6$.

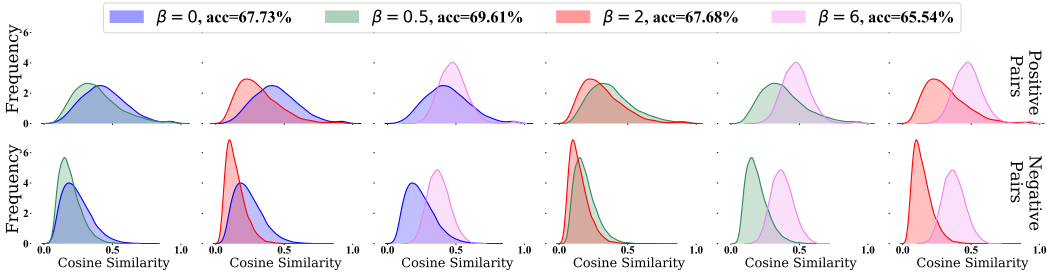

Figure 4: Histograms of cosine similarity of pairs of points with different label (bottom) and same label (top) for embeddings trained on CIFAR100 with different values of β. Histograms overlaid pairwise to allow for easy comparison.

```
# pos      : exp of inner products for positive examples
# neg      : exp of inner products for negative examples
# N        : number of negative examples
# t        : temperature scaling
# tau_plus: class probability
# beta     : concentration parameter

#Original objective
standard_loss = -log(pos.sum() / (pos.sum() + neg.sum()))

#Debiased objective
Neg = max((-N*tau_plus*pos + neg).sum() / (1-tau_plus), e**(-1/t))
debiased_loss = -log(pos.sum() / (pos.sum() + Neg))

#Hard sampling objective (Ours)
reweight = (beta*neg) / neg.mean()
Neg = max((-N*tau_plus*pos + reweight*neg).sum() / (1-tau_plus), e**(-1/t))
hard_loss = -log( pos.sum() / (pos.sum() + Neg))
```

Figure 5: Pseudocode for our proposed new hard sample objective, as well as the original NCE contrastive objective, and debiased contrastive objective. In each case we take the number of positive samples to be $M = 1$. The implementation of our hard sampling method only requires two additional lines of code compared to the standard objective.

C Experimental Details

Figure 5 shows PyTorch-style pseudocode for the standard objective, the debiased objective, and the hard sampling objective. The proposed hard-sample loss is very simple to implement, requiring only two extra lines of code compared to the standard objective.

We implement SimCLR in PyTorch. We use a ResNet-50 [19] as the backbone with embedding dimension 2048 (the representation used for linear readout), and projection head into the lower 128-dimensional space (the embedding used in the contrastive objective). We use the Adam optimizer [25] with learning rate 0.001 and weight decay 10^{-6}. Official code will be released. Since we adopt the SimCLR framework, the number of negative samples $N = 2(\text{batch size} - 1)$. Since we always take the batch size to be a power of 2 $(16, 32, 64, 128, 256)$ the negative batch sizes are $30, 62, 126, 254, 510$ respectively.

Annealing β Method: We detail the annealing method whose results are given in Figure 2. The idea is to reduce the concentration parameter down to zero as training progresses. Specifically, suppose we have e number of total training epochs. We also specify a number ℓ of "changes" to the concentration parameter we shall make. We initialize the concentration parameter $\beta_1 = \beta$ (where this β is the numbere reported in Figure 2), then once every e/ℓ epochs we reduce β_i by β/ℓ. In other words, if we are currently on β_i, then $\beta_{i+1} = \beta_i - \beta/\ell$, and we switch from β_i to β_{i+1} in epoch number $i \cdot e/\ell$. The idea of this method is to select particularly difficult negative samples early on order to obtain useful gradient information early on, but later (once the embedding is already quite good) we reduce the "hardness" level so as to reduce the harmful effect of only approximately correcting for false negatives (negatives with the same labels as the anchor).