Feature Dropout: Revisiting the Role of Augmentations in Contrastive Learning

Alex Tamkin
Stanford University
atamkin@stanford.edu

Margalit Glasgow
Stanford University
mglasgow@stanford.edu

Xiluo He
Stanford University
xiluohe@stanford.edu

Noah Goodman
Stanford University
ngoodman@stanford.edu

Abstract

What role do augmentations play in contrastive learning? Recent work suggests that good augmentations are *label-preserving* with respect to a specific downstream task. We complicate this picture by showing that label-destroying augmentations are often crucial in the foundation model setting, where the goal is to learn diverse, general-purpose representations for *multiple* downstream tasks. We perform contrastive learning experiments on a range of image and audio datasets with multiple downstream tasks (e.g. for digits superimposed on photographs, predicting the class of one vs. the other). We find that Viewmaker Networks, a recently proposed model for learning augmentations for contrastive learning, produce label-destroying augmentations that stochastically destroy features needed for different downstream tasks. These augmentations are interpretable (e.g. altering shapes, digits, or letters added to images) and surprisingly often result in better performance compared to expert-designed augmentations, despite not preserving label information. To support our empirical results, we theoretically analyze a simple contrastive learning setting with a linear model. In this setting, label-destroying augmentations are crucial for preventing one set of features from suppressing the learning of features useful for another downstream task.

1 Introduction

Foundation models (Bommasani et al., 2021) have exhibited remarkable progress on a range of AI tasks (Devlin et al., 2019; Liu et al., 2019; Ramesh et al., 2021; Radford et al., 2021; Brown et al., 2020; Chowdhery et al., 2022; Hoffmann et al., 2022; Alayrac et al., 2022; Reed et al., 2022; Goh et al., 2021), and crucially, can be adapted for a range of downstream tasks. For example, a foundation model trained well on ImageNet should perform well not only at object classification, but should also have learned features useful for localization, segmentation, or other tasks.

One popular strategy for training foundation models involves training models to match transformed versions (known as *views* or *augmentations*) of the same input. For example, common image views include data augmentations such as cropping or color jitter, while common views for speech include pitch modulation or spectrogram masking (Kharitonov et al., 2021; Park et al., 2019).

Much work has focused on the question of what views lead to high-quality representations. The prevailing consensus, exemplified by (Tian et al., 2020), holds that views should be *label-preserving*

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

with respect to a downstream task. Here, we question whether this assumption—in particular, with its focus on a single task—is enough to explain why contrastive foundation models succeed on a *range* of downstream tasks. Indeed, the actual choice and application of **views in practice** does not align with this prevailing consensus. For example, complete class invariance to several common data augmentations (e.g. shifts in brightness or cropping) is impossible, since augmentations of inputs from different classes can collide via cropping or repeated brightness changes.

Instead, we suspect that augmentations serve as a form of **feature dropout**—preventing any one feature from becoming a shortcut feature and suppressing the learning of other features. We study this empirically in Viewmaker Networks, a recently proposed generative model for views that appears to drop out some features in the input. We apply viewmaker and expert views to datasets with two associated downstream tasks, one involving classifying the main input (e.g., an image or audio recording) and one involving a simple overlaid element (e.g., a shape or speech snippet). We observe that the viewmaker augmentations selectively obscure the overlaid features. Despite this, the viewmaker representations learn both downstream tasks better than the expert view representations.

Finally, we formalize the intuition that feature dropout can aid learning with a theoretical analysis of a simple linear contrastive setting. In this setting, we characterize how the noisiness of each feature directly determines how quickly features are learned, and uncover an **interaction between features** governing how fast they are learned. In particular, we show how learning one feature quickly can suppress the learning of other features, and show that adding noise to the "easiest" feature can *increase* the rate at which other features are learned. This further indicates that *label-destroying* may help contrastive models learn a broad range of features for downstream tasks.

2 Viewmaker Networks Succeed Despite Destroying Label Information

To provide evidence that good views need not be label-preserving, we consider the behavior of viewmaker networks (Tamkin et al., 2021), a generative model which produces augmentations for contrastive learning. Intuitively, viewmakers learn a stochastic augmentation policy that makes the contrastive task as hard as possible for the encoder. The stochastic augmentations are parameterized as additive perturbations bounded by an L_1 norm, meaning the viewmaker can alter but not completely destroy the original image. Formally, given an input $x \in \mathbb{N}$, a viewmaker network V_ψ is trained jointly with an encoder E_θ to optimize the minimax expression:

$$\max_{\psi} \min_{\theta} \mathcal{L}\left(E_\theta\left(x + \epsilon\frac{V_\psi(x,\delta_1)}{||V_\psi(x,\delta_1)||_1}\right), E_\theta\left(x + \epsilon\frac{V_\psi(x,\delta_2)}{||V_\psi(x,\delta_2)||_1}\right)\right)$$

Here $\mathcal{L}$ is a multiview loss function (e.g. (Chen et al., 2020; He et al., 2020)), x is a minibatch of inputs, ϵ is the *distortion budget* controlling the strength of the views, and $\delta_1, \delta_2 \sim N(0,1)$ are random inputs that enable the viewmaker to learn a stochastic augmentation policy.

Viewmaker networks learn to stochastically alter different parts of the input, including task-relevant features, meaning that these augmentations are not label-preserving. Nevertheless, as we will see shortly, viewmaker networks enable strong performance on multiple downstream tasks, including often better performance than expert-designed augmentations. This **feature dropout** capability of viewmaker networks may help them learn many features well rather than focusing on the easiest ones.

Datasets. We consider the behavior of viewmaker networks on four datasets, including three image and one audio dataset. Each dataset is constructed in such a way as to support two distinct downstream classification tasks, enabling us to examine how well each downstream task is learned.
► **Image datasets** The three image datasets are based on the canonical CIFAR-10 image-recognition dataset (Krizhevsky, 2009). One task is always to predict the CIFAR-10 object label (e.g. `airplane` or `bird`). The other task is dependent on an additional feature overlaid on the image: **C+Shapes:** The CIFAR-10 image is overlaid with one of three randomly-colored shapes: a square, a triangle, or a circle. The second task is to predict what shape was overlaid (N=3 classes). For the other two datasets **C+Digits:** and **C+Letters:**, the images are overlayed with digits and letters respectively.

► **Audio dataset** The audio dataset is created by overlaying the audio of a spoken digit (from the AudioMNIST dataset (Becker et al., 2018), MIT License) with a random background noise (collected from one of three possible classes: cafe, machinery, and traffic) (Saki et al., 2016; Saki and Kehtarnavaz, 2016). The tasks are to predict the digit class (N=10 classes) and to predict the noise class (N=3 classes). Inputs are presented to the network as log mel spectrograms.

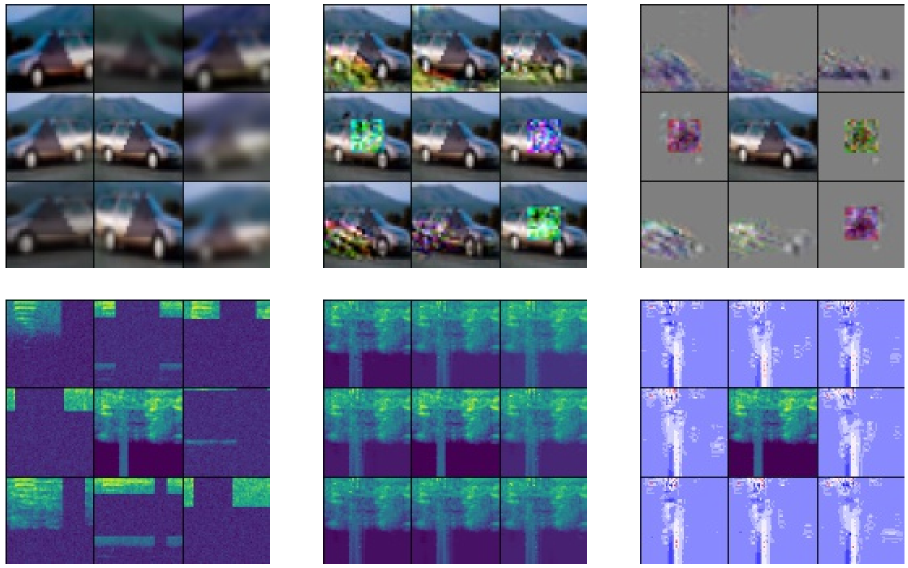


Figure 1: **Comparison of viewmaker and expert augmentations on datasets with multiple features.** The viewmaker augmentations adapt to the particular semantics of the input data, and make targeted perturbations which remove the class-relevant information of the synthetic features (e.g. occluding the digit, shape, letter, or speech). *Rows* (from top): Shapes, Audio. *Columns* (from left): Expert augmentations, viewmaker augmentations, difference between original and viewmaker augmentation. Center image in each grid is the original. Audio Expert views shown are Spectral views. Additional figures for Digits and Letter augmentations displayed in Apx D (Fig. 3).

Experiments We pretrain with the SimCLR algorithm, on a ResNet-18 model with standard modifications for smaller inputs. For the expert augmentations, we use the standard SimCLR augmentations for the image datasets (Chen et al., 2020), and the SpecAug (Park et al., 2019) and WaveAug (Kharitonov et al., 2021) augmentations for the audio datasets. We evaluate the quality of the learned representations by training a linear softmax classifier on top of the prepool representations. Full training details are given in Appendix C.

Results

► **Evidence of feature dropout.** Visually, the viewmaker augmentations seem to stochastically alter different aspects of the input. In addition to modifying the background of each input, the viewmaker also selectively modifies the additional synthetic features added to each domain. For instance, in **C+Shapes**, the viewmaker augmentations sometimes draw squares around the shape in the center, making it difficult to determine the shape class. In **Audio**, the viewmaker identifies the narrow band corresponding to the speech and applies perturbations to it. As seen in Figure 1, these label-destroying augmentations are common, occuring in a sizeable fraction of the sampled views. We measure the selectivity of feature dropout quantitatively in Appendix D and Figure 2.

► **Viewmaker succeeds despite destroying label information.** As shown in Table 1, viewmaker networks are able to achieve good accuracy on both tasks, while expert augmentations frequently achieve lower performance on one or both tasks. For example, on the image tasks, while expert views achieve slightly higher performance on the image only, they experience a large drop in accuracy when the synthetic feature is added. For the audio experiments the picture is similar—the viewmaker is able to avoid catastrophic drops in performance learning both features together, achieving the highest accuracy on both, while the expert views experience larger drops and worse overall performance.

These results provide evidence that label-preserving views are not necessary for learning good representations—the ability to perform feature dropout may even benefit learning multiple tasks.

	Viewmaker (CIFAR-10)	Expert (CIFAR-10)	Viewmaker (Object)	Expert (Object)
CIFAR-10 Only	84.5	**86.2**	-	-
C+Shape	**79.8**	76.0	**100.0**	**100.0**
C+Digit	**69.3**	58.8	**94.3**	86.7
C+Letter	71.9	**74.8**	**96.9**	94.1

	Speech Accuracy			Background Noise Accuracy		
	Viewmaker	Spectral	Waveform	Viewmaker	Spectral	Waveform
Speech Only	92.4	**97.0**	76.7	-	-	-
Bkgd. Noise Only	-	-	-	**100.0**	32.64	**100.0**
Speech + Noise	**60.8**	10.1	53.6	**97.0**	47.2	43.3

Table 1: **Transfer accuracy on different features.** Viewmaker networks achieve good performance across multiple downstream tasks, while expert views sometimes falter. Networks are pretrained on the datasets on the left, and transfer accuracy is reported for the conditions on the columns.

3 Theoretical Analysis of Feature Interactions in Simple Contrastive Setting

We theoretically analyze a simple linear model that captures the essence of how label-destroying augmentations can improve downstream accuracy. We study a setting where the data contains many underlying features that are relevant to downstream classification tasks, and where these features are preserved to varying degrees across augmentations. We will show adding noise to one feature can speed the learning of other features during gradient descent (GD) on a contrastive objective.

Data Model and Setting. We study a model which consists of data with K distinct features, each corresponding to some ground truth unit-vector directions $\mu_1, \ldots, \mu_K \in \mathbb{R}^d$. We sample each data point $u \in \mathbb{R}^{K \times d}$ and its *augmentation* (a.k.a. its *positive pair* or *view*) $v \in \mathbb{R}^{K \times d}$ as follows. For $k \in 1, \ldots, K$, the kth row of u, which we denote u_k, is drawn from the Gaussian distribution $\mathcal{N}(0, I_d)$. The kth row of the view, v_k, is drawn from the same distribution, but is correlated with u_k in the μ_k-direction (and is otherwise independent in the other directions). The strength of the correlation is governed by parameter $\alpha_k \in [0,1]$ in the following sense: $v_k^T \mu_k = \alpha_k u_k^T \mu_k + \sqrt{1-\alpha_k^2}\xi$, where $\xi \sim \mathcal{N}(0,1)$. Thus the larger α_k, the stronger the correlation in that feature across the two views.

We learn a model $\Theta \in \mathbb{R}^{K \times d}$ representing a collection of K feature extractors. The model Θ, with rows $\{\theta_k\}_{k \in [K]}$, maps a data point $w \in \mathbb{R}^{K \times d}$ to a representation $f_\Theta(w) \in \mathbb{R}^K$ by computing a score $w_k^T \theta_k$ for each element in the representation. That is, $(f_\Theta(w))_k = w_k^T \theta_k$. Our goal is that the model Θ will be useful for a downstream classification task which depends on the ground truth features. A good representation will capture ground truth features that are correlated across augmentations, such that θ_k is aligned with $\pm\mu_k$: Lemma E.1, given in Appendix E, proves that the angle $\arccos\left(\frac{|\theta_k^T \mu_k|}{\|\theta_k\|}\right)$ directly determines the test accuracy on a natural downstream classification task.

Training. We will study the evolution of Θ as we optimize a standard contrastive learning objective using GD (Dosovitskiy et al., 2014; Chen et al., 2020). At each round of GD, we sample a fresh batch of m data points and their augmentations, $(U,V) := \{(u^{(i)}, v^{(i)}\}_{i \in [m]}$. We compute the similarity scores $z_{ij} := \langle f_\Theta(u^{(i)}), f_\Theta(v^{(j)})\rangle = \sum_k (\theta_k^T u_k^{(i)})(\theta_k^T v_k^{(j)})$ using the dot product of the K-dimensional representations, and then use the cross entropy loss $\mathcal{L}(\Theta; U, V) := -\sum_i \log \frac{z_{ii}}{\sum_j z_{ij}}$.

Main Result. We consider how adding noise to one feature during training affects the learning of the other features. Formally, we say we *add noise* to some feature k' of a data point v, if for some $\beta \in [0,1)$, we let $\tilde{v}_{k'} = \beta v_{k'} + \sqrt{1-\beta^2}\xi$, where $\xi \sim \mathcal{N}(0, I_d)$, and $\tilde{v}_k = v_k$ for $k \neq k'$. Thus if (u,v) were a pair generated with the correlation coefficients $\{\alpha_k\}_{k \in [K]}$, then the distribution of $(u, \tilde{v})$ comes from the modified correlation coefficients $\{\tilde{\alpha}\}_{k \in [K]}$ with the single modification $\tilde{\alpha}_{k'} = \beta\alpha_k$. The following thoerem shows that if we add noise to the k'th feature, the next step of gradient descent learns all other features *better* than if we didn't add noise.

Theorem 3.1 (Noise improves feature learning)**.** *There exists a universal constant C, such that the following holds. Let* $\Theta^{(t+1)} = \Theta^{(t)} - \eta(\nabla\mathcal{L}(U,V;\Theta) + \lambda\Theta^{(t)})$*, and* $\tilde{\Theta}^{(t+1)} = \Theta^{(t)} - \eta(\nabla\mathcal{L}(U,\tilde{V};\Theta) +$

$\lambda\Theta^{(t)})$*, where* $\tilde{V}$ *is* V *with any amount of added noise in the* k' *feature. Then for any* $k \neq k'$*, if* $|\theta_k^T \mu_k| \leq \frac{1-\alpha_{k'}^2}{C}\|\theta_k\|$*,* $\|\theta_{k'}\|^3 \leq |\theta_{k'}^T \mu_k|$*, and* $\|\theta_k\|^2 \leq \frac{\alpha_k(1-\alpha_{k'}^2)}{C}$*, then for* η *small enough,*

$$\mathbb{E}_{U,V}\left[\arccos\left(\frac{|\mu_k^T \theta_k^{(t+1)}|}{\|\theta_k^{(t+1)}\|_2}\right)\right] > \mathbb{E}_{U,\tilde{V}}\left[\arccos\left(\frac{|\mu_k^T \tilde{\theta}_k^{(t+1)}|}{\|\tilde{\theta}_k^{(t+1)}\|_2}\right)\right]. \quad (1)$$

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198, 2022.

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

Soren Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Interpreting and explaining deep neural networks for classification of audio signals. *ArXiv*, abs/1807.03418, 2018.

Suzanna Becker and Geoffrey E. Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355:161–163, 1992.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *ArXiv*, abs/2108.07258, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020.

Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, 34, 2021.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Baindoor Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner

Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Oliveira Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311, 2022.

Elijah Cole, Xuan S. Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge J. Belongie. When does contrastive visual representation learning work? *ArXiv*, abs/2105.05837, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

Alexey Dosovitskiy, Jost Tobias Springenberg, Martin A. Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *NIPS*, 2014.

Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. How well do self-supervised models transfer? *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5410–5419, 2021.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix Wichmann. Shortcut learning in deep neural networks. *ArXiv*, abs/2004.07780, 2020.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Christopher Olah. Multimodal neurons in artificial neural networks. 2021.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2:1735–1742, 2006.

Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *NeurIPS*, 2021.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020.

Katherine L. Hermann and Andrew Kyle Lampinen. What shapes feature representations? exploring datasets, architectures, and training. *ArXiv*, abs/2006.12433, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and L. Sifre. Training compute-optimal large language models. *ArXiv*, abs/2203.15556, 2022.

A. Tarun Karthik, Mike Wu, Noah D. Goodman, and Alex Tamkin. Tradeoffs between contrastive and supervised learning: An empirical study. *ArXiv*, abs/2112.05340, 2021.

Eugene Kharitonov, Morgane Rivière, Gabriel Synnaeve, Lior Wolf, Pierre-Emmanuel Mazaré, Matthijs Douze, and Emmanuel Dupoux. Data augmenting contrastive learning of speech representations in the time domain. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 215–222. IEEE, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Tianhong Li, Lijie Fan, Yuan Yuan, Hao He, Yonglong Tian, Rogério Schmidt Feris, Piotr Indyk, and Dina Katabi. Addressing feature suppression in unsupervised visual representations. 2020.

Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. *ArXiv*, abs/1907.04595, 2019.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.

Matthias Minderer, Olivier Bachem, Neil Houlsby, and Michael Tschannen. Automatic shortcut removal for self-supervised representation learning. In *ICML*, 2020.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6706–6716, 2020.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092, 2021.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Manfred Otto Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent. *ArXiv*, abs/2205.06175, 2022.

Joshua Robinson, Li Sun, Ke Yu, K. Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34:4974–4986, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *ArXiv*, abs/1911.08731, 2019.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. *ArXiv*, abs/2005.04345, 2020.

Fatemeh Saki and Nasser Kehtarnavaz. Automatic switching between noise classification and speech enhancement for hearing aid devices. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 736–739, 2016. doi: 10.1109/EMBC.2016.7590807.

Fatemeh Saki, Abhishek Sehgal, Issa Panahi, and Nasser Kehtarnavaz. Smartphone-based real-time classification of noise signals using subband features and random forest classifier. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2204–2208, 2016. doi: 10.1109/ICASSP.2016.7472068.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *ArXiv*, abs/2006.07710, 2020.

Herbert A Simon. Spurious correlation: A causal interpretation. *Journal of the American statistical Association*, 49(267):467–479, 1954.

Megha Srivastava, Tatsunori B. Hashimoto, and Percy Liang. Robustness to spurious correlations via human annotations. In *ICML*, 2020.

Alex Tamkin, Mike Wu, and Noah D. Goodman. Viewmaker networks: Learning views for unsupervised representation learning. *ArXiv*, abs/2010.07432, 2021.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *ArXiv*, abs/2005.10243, 2020.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633, 2020.

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

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

Kai Y. Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *ArXiv*, abs/2006.09994, 2021.

Xingyi Yang, Xuehai He, Yuxiao Liang, Yue Yang, Shanghang Zhang, and Pengtao Xie. Transfer learning or self-supervised learning? a tale of two pretraining paradigms. *ArXiv*, abs/2007.04234, 2020.

Nanxuan Zhao, Zhirong Wu, Rynson W. H. Lau, and Stephen Lin. What makes instance discrimination good for transfer learning? *ArXiv*, abs/2006.06606, 2021.

A Code release

We are preparing our code and will release it on a public GitHub repo.

B Related Work

Understanding contrastive and multiview learning Many prior works have laid the foundations for current contrastive and multiview learning algorithms (Becker and Hinton, 1992; Hadsell et al., 2006; Dosovitskiy et al., 2014; Wu et al., 2018; Bachman et al., 2019; Misra and van der Maaten, 2020; He et al., 2020; Chen et al., 2020). Several works perform analysis studies of contrastive learning to identify important factors (Cole et al., 2021; Zhao et al., 2021) or how contrastive models differ from supervised learning (Yang et al., 2020; Ericsson et al., 2021; Karthik et al., 2021). HaoChen et al. (2021) study contrastive learning using the concept of an augmentation graph. This model assumes the fraction of non-label preserving augmentations is “extremely small;” interestingly, we show in practice it can quite large and still yield good performance. Wang et al. (2022) theoretically study contrastive learning under an assumption of label-preserving augmentations, though they show that such an assumption alone does not suffice to learn. Most relevant to our work, Tian et al. (2020) study how the information shared between different views impacts learning of downstream tasks. We complicate this picture by analyzing the foundation model setting of learning features relevant for multiple tasks. In this setting, we find that label-destroying perturbations, thought to be harmful by Tian et al. (2020), are useful for preventing one feature from suppressing others.

Feature suppression Our work is closely connected to the notion of *feature suppression* (Hermann and Lampinen, 2020), where the presence of one feature can crowd out or suppress the learning of other features. Several works have explored the relevance of this concept in contrastive learnings. For example, the original SimCLR paper (Chen et al., 2020) noted that color jitter augmentation was necessary to prevent the network for using only the color profile of the input to solve the contrastive task. Followup work (Chen et al., 2021) explores this phenomenon in more detail, characterizing how different hyperparameters and dataset features affect feature suppression. Other works have attempted to address feature suppression in contrastive learning, either via auxiliary losses (Li et al., 2020) or by modifying representations in the latent space (Robinson et al., 2021). Our work relates to these in two ways. First, we empirically and theoretically investigate feature suppression as an alternate rationale for the role of augmentations, as opposed to invariance. Second, we show that an existing method, viewmaker networks (Tamkin et al., 2021), can identify and potentially neutralize suppressing features in an interpretable way, resulting in better performance than expert augmentations.

Spurious correlations and shortcut features Outside the framing of feature suppression, several other works explore how classifiers can learn or make use of unwanted features. Shortcut features (Geirhos et al., 2020) describe often-simple features (e.g. the average color of an input) which are learned by networks at the expense of more salient features (e.g. the object class). This notion is connected to spurious correlations (Simon, 1954) in deep learning which have been explored extensively (Sagawa et al., 2019, 2020; Srivastava et al., 2020; Tu et al., 2020; Xiao et al., 2021), including in the context of self-supervised learning (Minderer et al., 2020). Other works have also performed theoretical analysis of how related dynamics affect learning in the supervised setting (Li et al., 2019; Shah et al., 2020). Our work suggests that viewmaker networks may be a useful tool as well here—both as an interpretability tool to visualize the different features a network relies on, and as a way to reduce reliance on particular features without completely destroying the information.

C Training Details

Pretraining We pretrain with the SimCLR algorithm for 200 epochs with a batch size of 256 and a temperature of 0.1. We use a ResNet-18 model with standard modifications for smaller inputs (including a smaller stride and no initial maxpool) as used in Tamkin et al. (2021). For the expert augmentations, we use the standard SimCLR augmentations for the image datasets (Chen et al., 2020), and the SpecAug (Park et al., 2019) augmentations for the audio datasets, which randomly mask out different frequency and time bands, as well as the WaveAug (Kharitonov et al., 2021) augmentations, which alter various properties of the waveform such as the pitch and speed. For the viewmaker augmentations, we use a budget of $\epsilon = 0.05P$ for the image datasets, and $\epsilon = 0.125P$ for the audio datasets, where P is the number of pixels in the input.

Linear Evaluation We evaluate the quality of the learned representations by training a linear softmax classifier on top of the prepool representations. We train for 100 epochs, using the same parameters as Viewmaker (Tamkin et al., 2021), training separate linear classifiers using the same pretrained network for each downstream task (Chen et al., 2020). Augmentations are applied during training but not evaluation.

D Quantifying feature dropout

We perform an exploratory analysis to testing how well different views drop out the features in an input. We augment a single example (CIFAR-10 image plus an overlaid object) 1,200 times using a given augmentation policy (either the expert or viewmaker augmentations). We then encode the model with a classifier trained off of the other augmentation policy (i.e. expert for viewmaker augmentations or the reverse) in order to test how well the augmentations drop out the features. We use a different encoder to see the effects of the augmentations prior to the encoder having a chance to adapt to them.

We observe a bimodal behavior for the viewmaker views, suggesting that the model is adapting to the semantics of the input and has learned to stochastically drop out the simple feature some fraction of the time. By contrast, the expert views display no such structure.

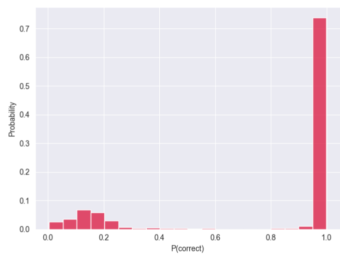
E Full proofs of propositions and theorems

Our theoretical set-up is pictured in Figure 4.

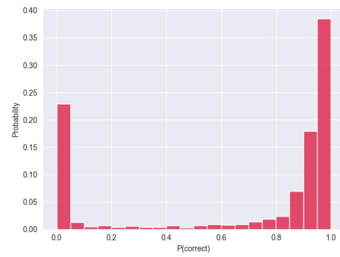
We begin by stating and proving Lemma E.1 on the downstream classification accuracy.

Lemma E.1 (Downstream classification accuracy)**.** *Suppose we draw labeled data points* $(u, y) \in \mathbb{R}^{K\times d} \times \{+1, 1\}$*, where as before,* $u_k \sim \mathcal{N}(0, I_d)$ *for* $k \in [K]$*, and the label is given by* $\mathrm{sign}(u_k^T \mu_k)$*. Then the best linear classifier* $\boldsymbol{a} \in \mathbb{R}^K$ *on the representations* $f_\Theta(u) \in \mathbb{R}^K$ *achieves an test error of* $\frac{1}{\pi}\arccos\left(\frac{|\mu_k^T\theta_k|}{\|\theta_k\|_2}\right)$*. That is*

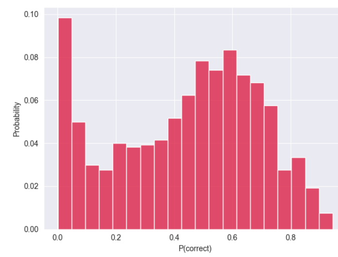
$$\min_{\boldsymbol{a}\in\mathbb{R}^K} \Pr_u[\mathrm{sign}(\boldsymbol{a}^T f_\Theta(u)) \neq \mathrm{sign}(\mu_k^T u_k)] = \frac{\arccos\left(\frac{|\mu_k^T\theta_k|}{\|\theta_k\|_2}\right)}{\pi}. \tag{2}$$



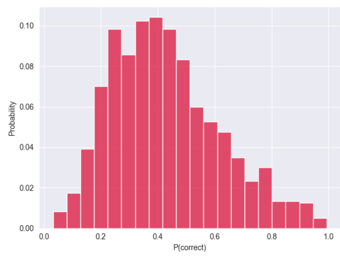
(a) Viewmaker / Shapes



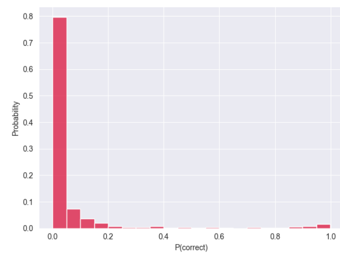
(b) Viewmaker / Letters



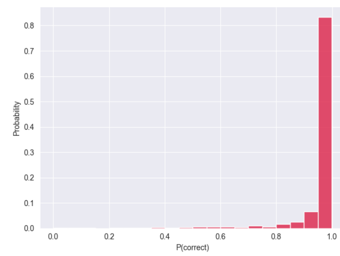
(c) Viewmaker / Digits



(d) Expert / Shapes



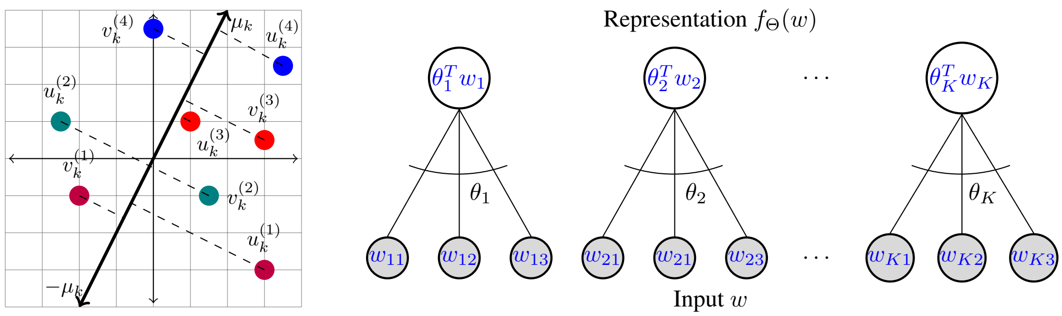
(e) Expert / Letters



(f) Expert / Digits

Figure 2: **Viewmaker augmentations stochastically drop out simple features added to the input.** Probability of the correct answer for different augmentations (Viewmaker or Expert) and different examples from different datasets (Shapes, Letters, Digits). Each histogram shows a single example from each dataset randomly augmented 1200 times, and the corresponding probabilities of the correct answer. The viewmaker augmentations display a bimodal structure, indicating that the simple feature is selectively either destroyed or preserved. The expert augmentations by contrast lack such structure, reflecting their lack of adaptation to the structure of each input.

Figure 3: **Comparison of viewmaker and expert augmentations on datasets with multiple features.** The viewmaker augmentations adapt to the particular semantics of the input data, and make targeted perturbations which remove the class-relevant information of the synthetic features (e.g. occluding the digit, shape, letter, or speech). Despite this, the encoder network is still able to learn strong representations. *Rows* (from top): Digits, Shapes. *Columns* (from left): Expert augmentations, viewmaker augmentations, difference between original and viewmaker augmentation, rescaled to [0,1]. Center image in each grid is the original. Audio Expert views shown are Spectral views.



(a) (b)

Figure 4: (a) The correlation of the kth feature of an augmentation pair, shown for $d = 2$. Each pair $u_k^{(i)}$ and $v_k^{(i)}$ have correlated projections onto the ground truth μ_k direction, representing the feature conserved across augmentations. (b) Feedforward linear network which computes the representation $f_\Theta(w)$. As each feature μ_k is learned ($\theta_k \to \mu_k$) the representations of the two views $f_\Theta(u^{(i)})$, $f_\Theta(v^{(i)})$ become more similar, decreasing the contrastive loss.

Thus if θ_k and μ_k are orthogonal, then the test error is 50%. If the angle between θ_k and the $\pm\mu_k$ is zero, then we achieve perfect classification accuracy.

Proof. It is easy to see that the best linear classifier $\boldsymbol{a}$ will (up to scaling) be equal to the vector $\operatorname{sign}(\mu_k^T\theta_k)e_k$. Such a classifier predicts the correct sign whenever $\operatorname{sign}(\boldsymbol{a}^T f_\Theta(u)) = \operatorname{sign}(\mu_k^T\theta_k)\operatorname{sign}(\theta_k^T u_k)$ equals $\operatorname{sign}(\mu_k^T u_k)$, which occurs exactly a $1 - \frac{\arccos\left(\frac{|\mu_k^T\theta_k|}{\|\theta_k\|_2}\right)}{\pi}$ fraction of the time. □

In the rest of this section, we prove our main theoretical result, Theorem 3.1, which shows that $\arccos\left(\frac{|\mu_k^T\theta_k|}{\|\theta_k\|_2}\right)$ decreases faster in expectation during gradient descent if we add noise to the k' feature.

E.1 Notation.

We let δ_{ij} denote the δ-function which equals 1 if $i = j$ and 0 otherwise. For a parameter $\Theta = \{\theta_k\}_{k\in[K]}$, we let $\theta_k^{\|} := \mu_k\mu_k^T\theta_k$ be the projection of θ_k in the μ_k direction. We let $\theta_k^{\perp} = \theta_k - \theta_k^{\|}$ be the projection of θ_k orthogonal to the feature μ_k.

Throughout this section, we consider the ground truth directions to be fixed, and we fix some initial correlation vector $\boldsymbol{\alpha}$. We let $\mathbb{P}_{\boldsymbol{\alpha}}$ denote the distribution from which the pair (u, v) is drawn from the Gaussian distribution described in Section 3 with correlation coefficients $\boldsymbol{\alpha}$. When unspecified, the variables U, V are drawn from the distribution $\mathbb{P}_{\boldsymbol{\alpha}}^m$. Since we study what happens when we vary $\alpha_{k'}$, for $x \in [0, 1]$, we use the shorthand $\mathbb{P}_x$ to denote the distribution $\mathbb{P}_{\boldsymbol{\alpha}(x)}^m$, where $\alpha(x)_{k'} = x$, and $\alpha(x)_k = \alpha_k$ for all other k.

We denote $\mathcal{L}_i(\Theta; U, V) = \operatorname{CE}(\{p_{ij}\}_{j\in[m]}, e_i) = -\log(p_{ii})$, which we abbreviate by $\mathcal{L}_i$. When it is clear that we are considering $\mathcal{L}_i$ for some fixed i, we omit the superscripts on the ith data point or its pair. That is, we denote $u_k := u_k^{(i)}$ and $v_k := v_k^{(i)}$.

E.2 Preliminaries

The following facts about of the derivative of the cross entropy loss are easy derived.

Lemma E.2.

$$\frac{\partial \mathcal{L}_i}{\partial \Theta} = \sum_j (p_{ij} - \delta_{ij})\frac{\partial z_{ij}}{\partial \Theta} = \sum_i \sum_{j\neq i} p_{ij}\left(\frac{\partial z_{ij}}{\partial \Theta} - \frac{\partial z_{ii}}{\partial \Theta}\right), \tag{3}$$

where

$$\frac{\partial z_{ij}}{\partial \theta_k} = (u_k^{(i)} {v_k^{(j)}}^T + v_k^{(j)} {u_k^{(i)}}^T)\theta_k. \tag{4}$$

We will also need the following facts on Gaussian random variables. The first, Stein's Lemma, is well known.

Lemma E.3 (Stein's Lemma)**.**

$$\mathbb{E}_{X\sim\mathcal{N}(0,\sigma^2)}[Xf(X)] = \sigma^2 \mathbb{E}_{X\sim\mathcal{N}(0,\sigma^2)}[f'(X)]. \tag{5}$$

The next two lemmas are proved in Section E.4.

Lemma E.4. *There exists some constant C such that following holds. If $\sigma \leq \frac{1}{C}$, and $0 \leq t \leq \frac{1}{\sigma}$, then for any $c \in \{0,1,2,3\}$, and $X \sim \mathcal{N}(0,\sigma^2)$ we have*

$$\mathbb{E}_X \left[|X|^c \exp(t|X|)\exp(tX^2)\right] \leq C\sigma^c. \tag{6}$$

If additionally $d \in \{0,1,2,3\}$, $\rho \leq \frac{1}{C}$ and $Y \sim \mathcal{N}(0,\rho^2)$, then

$$\mathbb{E}_X \left[|X|^c |Y|^d \exp(t|X|)\exp(|XY|)\right] \leq C\sigma^c \rho^d. \tag{7}$$

Lemma E.5. *For some universal constant C, for any $\sigma \in [0,1]$, $t \geq 0$, $c \in \{0,1,2,3,4\}$, we have*

$$\mathbb{E}_{X\sim\mathcal{N}(0,\sigma^2)}\left[(\exp(t|X|)-1)|X|^c\right] \leq Ct\sigma^c.$$

E.3 Approach and Lemmas

We outline our proof of Theorem 3.1 in this section. We prove all the lemmas below in Section E.4.

To understand $\mathbb{E}_{U,V}\left[\arccos\left(\frac{|\mu_k^T \theta_k^{(t+1)}|}{\|\theta_k^{(t+1)}\|_2}\right)\right]$ for a small enough step size, we first claim that it suffices to understand the expected projection of the gradient with respect to θ_k in the μ_k direction and in the θ_k direction. We use the notation $\nabla_k = \frac{\partial \mathcal{L}(\Theta;U,V)}{\partial \theta_k}$.

Lemma E.6. *Let $\theta_k^+ = \theta_k - \eta(\nabla_k + \lambda\theta_k)$. Then*

$$\lim_{\eta\to 0}\frac{1}{\eta}\left(\mathbb{E}_{U,V}\left[\arccos\left(\frac{|\mu_k^T\theta_k^+|}{\|\theta_k^+\|_2}\right)\right] - \arccos\left(\frac{|\mu_k^T\theta_k|}{\|\theta_k\|_2}\right)\right) = N\mathbb{E}_{U,V}\left[-(\mu_k^T\theta_k)(\mu_k^T\nabla_k) + \frac{\theta_k^T\nabla_k(\mu_k^T\theta_k)^2}{\|\theta_k\|_2^2}\right], \tag{8}$$

where N is some negative value that depends only on θ_k.

Now, since we care about the quantity $\mathbb{E}_{U,V}\left[\arccos\left(\frac{|\mu_k^T\theta_k^{(t+1)}|}{\|\theta_k^{(t+1)}\|_2}\right)\right] - \mathbb{E}_{U,\tilde{V}}\left[\arccos\left(\frac{|\mu_k^T\tilde{\theta}_k^{(t+1)}|}{\|\tilde{\theta}_k^{(t+1)}\|_2}\right)\right]$ being positive, it suffices to show that *derivative*

$$\frac{d}{dx}\mathbb{E}_{U,V\sim\mathbb{P}_x}\left[-(\mu_k^T\theta_k)(\mu_k^T\nabla_k) + \frac{\theta_k^T\nabla_k(\mu_k^T\theta_k)^2}{\|\theta_k\|_2^2}\right],$$

is negative for all $x \in [\tilde{\alpha}_{k'}, \alpha_{k'}]$. Indeed, from Lemma E.6, we have that

$$\lim_{\eta\to 0}\frac{1}{\eta}\left(\mathbb{E}_{U,V\sim\mathbb{P}_{\alpha_{k'}}}\left[\arccos\left(\frac{|\mu_k^T\theta_k^+|}{\|\theta_k^+\|_2}\right)\right] - \mathbb{E}_{U,V\sim\mathbb{P}_{\tilde{\alpha}_{k'}}}\left[\arccos\left(\frac{|\mu_k^T\theta_k|}{\|\theta_k\|_2}\right)\right]\right) \tag{9}$$

$$= N\int_{\tilde{\alpha}_{k'}}^{\alpha_{k'}}\frac{d}{dx}\mathbb{E}_{U,V\sim\mathbb{P}_x}\left[-(\mu_k^T\theta_k)(\mu_k^T\nabla_k) + \frac{\theta_k^T\nabla_k(\mu_k^T\theta_k)^2}{\|\theta_k\|_2^2}\right]dx, \tag{10}$$

so if the derivative is negative for the full range, then the difference in arccosines is positive.

In the following lemma we compute the derivative of $\mathbb{E}[\nabla_k]$ with respect to x.

Lemma E.7.

$$\frac{d}{dx}\mathbb{E}_{U,V\sim\mathbb{P}_x}[\nabla_k] = m\frac{d}{dx}\mathbb{E}_{U,V\sim\mathbb{P}_x}\left[\frac{\partial\mathcal{L}_i}{\partial\theta_k}\right]$$

$$= \frac{-m}{1-x^2}\theta_{k'}^T\mu_{k'}\sum_{j\neq i}\mathbb{E}_{U,V\sim\mathbb{P}_x}\left[p_{ij}p_{ii}\left(\theta_{k'}^T u_{k'}\right)\left(\mu_{k'}^T u_{k'}^{(i)} - x\mu_{k'}^T v_{k'}^{(i)}\right)\left(\frac{\partial(z_{ij}-z_{ii})}{\partial\theta_k}\right)\right].$$

We will analyze this quantity by explicitly taking the expectation with respect to some set of random variables. Let $S = \{U_k, V_k, U_{k'}, V_{k'}\}$ consist of the random variables $u_{k'}^{(i)}$, $u_k^{(i)}$, and $v_{k'}^{(i)}$, $v_k^{(i)}$ for all $i \in [m]$. Define q_{ij} to be the logits when all variables in S are set to 0 (Thus explicitly, $q_{ij} = \frac{\exp\left(\sum_{\tilde{k}\neq k,k'} \theta_{\tilde{k}}^T u_{\tilde{k}}^{(i)} \theta_{\tilde{k}}^T v_{\tilde{k}}^{(j)}\right)}{\sum_{j'} \exp\left(\sum_{\tilde{k}\neq k,k'} \theta_{\tilde{k}}^T u_{\tilde{k}}^{(i)} \theta_{\tilde{k}}^T v_{\tilde{k}}^{(j')}\right)}$). We will use the notation $j \sim q$ to denote the distribution on $[m]$ with mass q_{ij} on j.

Let

$$h(S) := \left(\theta_{k'}^T u_{k'}\right)\left(\mu_{k'}^T u_{k'}^{(i)} - x\mu_{k'}^T v_{k'}^{(i)}\right)\left(\frac{\partial(z_{ij} - z_{ii})}{\partial \theta_k}\right), \tag{11}$$

and

$$h_1(S) = \left(\theta_{k'}^T u_{k'}\right)\left((1-x^2)\mu_{k'}^T u_{k'}^{(i)}\right) 2\alpha_k \left((\mu_k^T u_k)(\theta_k^{\parallel} u_k)\mu_k^T\right), \tag{12}$$

which are the terms that appear in the right hand side of Lemma E.7 after $p_{ii}p_{ij}$. Observe that

$$\mathbb{E}_S[h(S) - h_1(S)] = 0.$$

The following four lemmas serve to bound $\frac{d}{dx}\mathbb{E}_S\left[\mu_k^T \nabla_k\right]$ and $\frac{d}{dx}\mathbb{E}_S\left[\theta_k^T \nabla_k\right]$. We call the terms of the form $\mathbb{E} p_{ii}p_{ij}(h(S) - h_1(S))$ "junk" terms, and our goal will be to show that these terms are small. We will control more closely the terms of the form $\mathbb{E} p_{ii}p_{ij}(h_1(S))$.

Lemma E.8 (Junk Terms for μ_k term.). *If $\|\theta_k\| \leq 1$ and $\|\theta_{k'}\| \leq 1$, then for some universal constant C*

$$\left|\mathbb{E}_S\left[p_{ii}p_{ij}\mu_k^T(h(S) - h_1(S))\right]\right| \leq C q_{ii}q_{ij}\left(\|\theta_{k'}\|^3\|\theta_k\|^3 + \|\theta_{k'}^{\parallel}\|\|\theta_k\|^3 + \alpha_k\left(\|\theta_{k'}\|^3\|\theta_k^{\parallel}\|\right)\right).$$

Lemma E.9 (Good Term for μ_k term.). *If $\|\theta_k\| \leq 1$ and $\|\theta_{k'}\| \leq 1$, then for some universal constant C*

$$\left|\mathbb{E}_S\left[p_{ii}p_{ij}\mu_k^T h_1(S)\right]\right| \geq 2\alpha_k(1-x^2)q_{ii}q_{ij}\left(\|\theta_{k'}^{\parallel}\|\|\theta_k^{\parallel}\|\right)\left(1 - C(\|\theta_{k'}\|^2 + \|\theta_k\|^2)\right).$$

Plugging these two lemmas into Lemma E.7 yields the following corollary.

Corollary E.9.1 (Total μ_k term.). *If for a sufficiently large constant C, $|\theta_k^T \mu_k| \leq \frac{1-\alpha_{k'}^2}{C}\|\theta_k\|$, $\|\theta_{k'}\|^3 \leq |\theta_{k'}^T \mu_k|$, and $\|\theta_k\|^2 \leq \frac{\alpha_k(1-\alpha_{k'}^2)}{C}$, then*

$$(\mu_k^T \theta_k)\frac{d}{dx}\mathbb{E}_{\mathbb{P}_x}\left[\mu_k^T \nabla_k\right] \geq \frac{m}{2}\mathbb{E}_{U,V\setminus S}\left[\sum_{i,j} q_{ii}q_{ij}2\alpha_k\|\theta_{k'}^{\parallel}\|^2\|\theta_k^{\parallel}\|^2\right].$$

Lemma E.10 (Junk Terms for θ_k term.). *If $\|\theta_k\| \leq 1$ and $\|\theta_{k'}\| \leq 1$, then for some universal constant C*

$$\left|\mathbb{E}_S\left[p_{ii}p_{ij}\theta_k^T(h(S) - h_1(S))\right]\right| \leq C q_{ii}q_{ij}\left(\|\theta_{k'}\|^3\|\theta_k\|^4 + \|\theta_{k'}^{\parallel}\|\|\theta_k\|^4 + \alpha_k\left(\|\theta_{k'}\|^3\|\theta_k\|\|\theta_k^{\parallel}\| + \|\theta_{k'}^{\parallel}\|\|\theta_k\|^3\|\theta_k^{\parallel}\|\right)\right).$$

Lemma E.11 (Good Term for θ_k term.). *If $\|\theta_k\| \leq 1$ and $\|\theta_{k'}\| \leq 1$, then for some universal constant C*

$$\left|\mathbb{E}_S\left[p_{ii}p_{ij}\theta_k^T h_1(S)\right]\right| \leq (1-x^2)2\alpha_k q_{ii}q_{ij}\left(\|\theta_{k'}^{\parallel}\|\|\theta_k^{\parallel}\|^2\right)\left(1 + C(\|\theta_{k'}\|^2 + \|\theta_k\|^2)\right).$$

Plugging these two lemmas into Lemma E.7 yields the following corollary.

Corollary E.11.1 (Total θ_k term.). *If for a sufficiently large constant C, $\|\theta_k^{\parallel}\| \leq \frac{1-x^2}{C}\|\theta_k\|$, $\|\theta_{k'}\|^3 \leq \|\theta_{k'}^{\parallel}\|$, $\|\theta_k\|^2 \leq \frac{\alpha_k(1-x^2)}{C}$, then*

$$\frac{(\mu_k^T \theta_k)^2}{\|\theta_k\|^2}\left|\frac{d}{dx}\mathbb{E}_{\mathbb{P}_x}\left[\theta_k^T \nabla_k\right]\right| \leq \frac{m}{2}\mathbb{E}_{U,V\setminus S}\left[\sum_{i,j} q_{ii}q_{ij}\alpha_k\|\theta_{k'}^{\parallel}\|^2\|\theta_k^{\parallel}\|^2\right].$$

Combining Corollaries E.9.1 and E.11.1, we obtain the following lemma.

Lemma E.12. *If for a sufficiently large constant C, $\|\theta_k^{\parallel}\| \leq \frac{1-x^2}{C}\|\theta_k\|$, $\|\theta_{k'}\|^3 \leq \|\theta_{k'}^{\parallel}\|$, $\|\theta_k\|^2 \leq \frac{\alpha_k(1-x^2)}{C}$, then*

$$\mathbb{E}_{U,V\sim\mathbb{P}_x}\left[-(\mu_k^T \theta_k)(\mu_k^T \nabla_k) + \frac{\theta_k^T \nabla_k(\mu_k^T \theta_k)^2}{\|\theta_k\|_2^2}\right] < 0. \tag{13}$$

Theorem 3.1 now follows.

E.4 Proofs of Lemmas

To prove the Lemmas E.4 and E.5, we will use the following well-known formula for the moment generating function (MGF) of the half-normal distribution.

Lemma E.13 (MGF of half-normal distribution)**.** *The MGF of the half-normal distribution is*

$$\mathbb{E}_{X\sim\mathcal{N}(0,1)|X>0}[e^{t|X|}] = 2e^{t^2/2}\Phi(t),$$

where $\Phi(t)$ is the cumulative distribution of a normal random variable.

Proof of Lemma E.4.

$$\begin{aligned}
\mathbb{E}_X\left[|X|^c\exp(t|X|)\exp(tX^2)\right] &= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}|x|^c\exp(t|x|)\exp(tx^2)\exp\left(-\frac{x^2}{2\sigma^2}\right)dx \\
&= \frac{\sqrt{1-2\sigma^2 t}}{\left(\frac{\sigma}{\sqrt{1-2\sigma^2 t}}\right)\sqrt{2\pi}}\int_{-\infty}^{\infty}|x|^c\exp(t|x|)\exp\left(-\frac{x^2}{2\left(\frac{\sigma}{\sqrt{1-2\sigma^2 t}}\right)^2}\right)dx \\
&= \sqrt{1-2\sigma^2 t}\,\mathbb{E}_{Z\sim\mathcal{N}(0,r)|Z\geq 0}[Z^c\exp(tZ)],
\end{aligned}$$

where $r = \frac{\sigma}{\sqrt{1-2\sigma^2 t}}$. To evaluate this, we use the MGF of the half-normal distribution in Lemma E.13. Thus for some constant C, for all $c \in \{1,2,3,4\}$,

$$\begin{aligned}
\mathbb{E}_{X\sim\mathcal{N}(0,1)|X>0}\left[c!|X|^c e^{t|X|}\right] &\leq \mathbb{E}_{X\sim\mathcal{N}(0,1)|X>0}\left[\frac{d^c}{dt^c}e^{t|X|}\right] \\
&\leq C\left(1+t^c\right)e^{t^2/2}.
\end{aligned}$$

So for some constant C (whose value changes throughout this equation), so long as $\sigma \leq \frac{1}{C}$,

$$\begin{aligned}
\sqrt{1-2\sigma^2 t}\,\mathbb{E}_{Z\sim\mathcal{N}(0,r)|Z\geq 0}[Z^c\exp(tZ)] &= \sqrt{1-2\sigma^2 t}\,\mathbb{E}_{X\sim\mathcal{N}(0,1)|Z\geq 0}[r^c Z^c\exp(rtZ)] \\
&\leq \sqrt{1-2\sigma^2 t}\,Cr^c\left(1+(tr)^c\right)e^{(tr)^2/2} \\
&\leq C\sigma^c.
\end{aligned}$$

This proves the first statement in the lemma. To prove the second, we first take the expectation over X, and using the half-Gaussian MGF as before, we obtain

$$\mathbb{E}_X\mathbb{E}_Y\left[|X|^c|Y|^d\exp(t|X|)\exp(|XY|)\right] \leq C\mathbb{E}_Y\left[|Y|^d\sigma^c(1+(t+|Y|)^c)e^{(t+|Y|)^2/2}\right]$$

Now applying the first statement to take the expectation over Y, we obtain

$$\mathbb{E}_Y\left[|Y|^d(1+(t+|Y|)^c)e^{(t+|Y|)^2/2}\right] \leq C\sigma^c\rho^d.$$

□

Proof of Lemma E.5. We prove the lemma by induction on c. Suppose $c = 0$. Then by plugging in the MGF for the half-normal distribution from Lemma E.13, for some constant C, we have

$$\begin{aligned}
\mathbb{E}_{X\sim\mathcal{N}(0,1)|X>0}[(e^{t|X|}-1)] &= 2e^{t^2/2}\Phi(t) - 1 && (14) \\
&\leq 2e^{t^2/2}\left(\frac{1+t}{2}\right) - 1 && (15) \\
&\leq \left(e^{t^2/2}-1\right) + te^{t^2/2} && (16) \\
&\leq Ct, && (17)
\end{aligned}$$

thus

$$\mathbb{E}_{X\sim\mathcal{N}(0,\sigma^2)}[(e^{t|X|}-1)] = \mathbb{E}_{X\sim\mathcal{N}(0,\sigma^2)|X>0}[(e^{\sigma t|X|}-1)] \leq Ct\sigma.$$

Now for $c \geq 1$, by Stein's Lemma, we have (for a new constant C),

$$\mathbb{E}_{X\sim\mathcal{N}(0,\sigma^2)}[|X|^c(e^{t|X|}-1)] = \mathbb{E}_{X\sim\mathcal{N}(0,\sigma^2)}[X|X|^{c-1}\operatorname{sign}(X)(e^{t|X|}-1)] \tag{18}$$

$$= \sigma^2\mathbb{E}_{X\sim\mathcal{N}(0,\sigma^2)}\left[\frac{d}{dX}\left(|X|^{c-1}\operatorname{sign}(X)(e^{t|X|}-1)\right)\right] \tag{19}$$

$$= \sigma^2\mathbb{E}_{X\sim\mathcal{N}(0,\sigma^2)}\left[(c-2)\left(|X|^{c-2}(e^{t|X|}-1)\right)+\left(|X|^{c-1}(te^{t|X|})\right)\right] \tag{20}$$

$$\leq Ct\sigma^{c+1}. \tag{21}$$

where in the last step we used the inductive hypothesis and Lemma E.4. □

Proof of Lemma E.6. First observe that

$$\begin{aligned}
&\lim_{\eta\to 0}\frac{1}{\eta}\left(\mathbb{E}_{U,V}\left[\arccos\left(\frac{|\mu_k^T\theta_k^+|}{\|\theta_k^+\|_2}\right)\right]-\arccos\left(\frac{|\mu_k^T\theta_k|}{\|\theta_k\|_2}\right)\right)\\
&\quad=\lim_{\eta\to 0}\frac{1}{\eta}\left(\mathbb{E}_{U,V}\left[\arccos\left(\frac{|\mu_k^T(\theta_k(1-\eta\lambda)-\eta\nabla_k)|}{\|\theta_k(1-\eta\lambda)-\eta\nabla_k\|_2}\right)\right]-\arccos\left(\frac{|\mu_k^T\theta_k|}{\|\theta_k\|_2}\right)\right)\\
&\quad=\lim_{\eta\to 0}\frac{1}{\eta}\left(\mathbb{E}_{U,V}\left[\arccos\left(\frac{|\mu_k^T(\theta_k-\frac{\eta}{1-\eta\lambda}\nabla_k)|}{\|\theta_k-\frac{\eta}{1-\eta\lambda}\nabla_k\|_2}\right)\right]-\arccos\left(\frac{|\mu_k^T\theta_k|}{\|\theta_k\|_2}\right)\right)\\
&\quad=\mathbb{E}_{U,V}\left[\frac{d}{d\eta}\arccos\left(\frac{|\mu_k^T(\theta_k-\eta\nabla_k)|}{\|\theta_k-\eta\nabla_k\|_2}\right)(0)\right],
\end{aligned}$$

since $\lim_{\eta\to 0}\frac{\eta}{1-\eta\lambda}=0$. Now

$$\begin{aligned}
\frac{d}{d\eta}\arccos\left(\frac{|\mu_k^T(\theta_k-\eta\nabla_k)|}{\|\theta_k-\eta\nabla_k\|_2}\right)(0) &= \arccos'\left(\frac{|\mu_k^T\theta_k|}{\|\theta_k\|_2}\right)\frac{d}{d\eta}\left(\frac{|\mu_k^T(\theta_k-\eta\nabla_k)|}{\|\theta_k-\eta\nabla_k\|_2}\right)(0)\\
&= \arccos'\left(\frac{|\mu_k^T\theta_k|}{\|\theta_k\|_2}\right)\left(\frac{-\operatorname{sign}(\mu_k^T\theta_k)\mu_k^T\nabla_k\|\theta_k\|+|\mu_k^T\theta_k|\frac{\theta_k^T\nabla_k}{\|\theta_k\|}}{\|\theta_k\|_2^2}\right)\\
&= N\left(-\mu_k^T\theta_k\mu_k^T\nabla_k+(\mu_k^T\theta_k)^2\frac{\theta_k^T\nabla_k}{\|\theta_k\|^2}\right),
\end{aligned}$$

where $N=\arccos'\left(\frac{|\mu_k^T\theta_k|}{\|\theta_k\|_2}\right)\frac{1}{\|\theta_k\||\mu_k^T\theta_k|}$. The lemma follows by taking the expectation over U,V, and observing derivative of $\arccos(x)$ is negative whenever x is positive. □

Proof of Lemma E.7. First observe that by symmetry, we have

$$\frac{d}{dx}\mathbb{E}_{U,V\sim\mathbb{P}_x}\left[\nabla_k\right]=m\frac{d}{dx}\mathbb{E}_{U,V\sim\mathbb{P}_x}\left[\frac{\partial\mathcal{L}_i}{\partial\theta_k}\right].$$

To make this expectation easier to analyze, we express the random variable $(U(x),V(x))\sim\mathbb{P}_x$ as an interpolation of Gaussians in the coordinate $\mu_{k'}^Tv_{k'}^{(i)}$. Let $\xi\sim\mathcal{N}(0,1)$, and define $(U,V)\sim\mathbb{P}_1$, such that $\mu_{k'}^Tv_{k'}^{(i)}=\mu_{k'}^Tu_{k'}^{(i)}$. For $x\in[0,1)$, define $(U(x),V(x))$ to have

$$\mu_{k'}^Tv_{k'}^{(i)}(x)=x\mu_{k'}^Tu_{k'}^{(i)}+\sqrt{1-x^2}\xi, \tag{22}$$

and otherwise be the same as (U,V). It is easy to check that $(U(x),V(x))\sim\mathbb{P}_x$.

Now

$$\frac{d}{dx}\mathbb{E}_{U,V\sim\mathbb{P}_x}\left[\frac{\partial\mathcal{L}_i(\Theta;U,V)}{\partial\theta_k}\right]=\mathbb{E}_{U,V\sim\mathbb{P}_1,\xi}\left[\frac{d}{dx}\frac{\partial\mathcal{L}_i(\Theta;U(x),V(x))}{\partial\theta_k}\right].$$

Taking the derivative of the cross-entropy loss, we have

$$\begin{aligned}\frac{d}{dx}\frac{\partial \mathcal{L}_i(\Theta; U(x), V(x))}{\partial \theta_k} &= \frac{d}{dx}\left(\sum_{j\neq i} p_{ij}\left(\frac{\partial(z_{ij}-z_{ii})}{\partial \theta_k}\right)\right)\\ &= \sum_{j\neq i}\frac{dp_{ij}}{d\mu_{k'}^T v_{k'}^{(i)}(x)}\frac{d\mu_{k'}^T v_{k'}^{(i)}(x)}{dx}\frac{\partial(z_{ij}-z_{ii})}{\partial\theta_k}\\ &= \sum_{j\neq i} -p_{ij}p_{ii}\frac{dz_{ii}}{d\mu_{k'}^T v_{k'}^{(i)}(x)}\left(\mu_{k'}^T u_{k'}^{(i)} - \frac{x}{\sqrt{1-x^2}}\xi\right)\left(\frac{\partial(z_{ij}-z_{ii})}{\partial\theta_k}\right)\end{aligned}$$

where the variables z_{ij} and p_{ij} are the similarity scores and the softmaxes from the data $(U(x), V(x))$. Here the first line is by Lemma E.2, and the second line holds by chain rule since $\frac{\partial z_{ij}}{\partial \theta_k} - \frac{\partial z_{ii}}{\partial \theta_k}$ does not depend on $v_{k'}^{(i)}$. The third line uses the proof of Claim E.14 to take the derivative of p_{ij}, and Equation 22 to take the derivative of $\mu_{k'}^T v_{k'}^{(i)}(x)$.

Now we reparameterize $\mu_{k'}^T u_{k'}^{(i)} - \frac{x}{\sqrt{1-x^2}}\xi$ as follows:

$$\mu_{k'}^T u_{k'}^{(i)} - \frac{x}{\sqrt{1-x^2}}\xi = \left(\frac{1}{1-x^2}\right)\mu_{k'}^T u_{k'}^{(i)} - \frac{x}{1-x^2}\mu_{k'}^T v_{k'}^{(i)}(x).$$

Plugging in this reparameterization and $\frac{dz_{ii}}{d\mu_{k'}^T v_{k'}^{(i)}(x)} = \theta_{k'}^T \mu_{k'} \theta_{k'}^T u_{k'}$, we obtain

$$\frac{d}{dx}\mathbb{E}_{U,V\sim\mathbb{P}_x}\left[\frac{\partial \mathcal{L}_i(\Theta; U, V)}{\partial\theta_k}\right] = \frac{-1}{1-x^2}\sum_{j\neq i}\mathbb{E}_{U,V\sim\mathbb{P}_x}\left[p_{ij}p_{ii}\left(\theta_{k'}^T\mu_{k'}\theta_{k'}^T u_{k'}\right)\left(\mu_{k'}^T u_{k'}^{(i)} - x\mu_{k'}^T v_{k'}^{(i)}\right)\left(\frac{\partial(z_{ij}-z_{ii})}{\partial\theta_k}\right)\right].$$

□

We now prove Lemmas E.8, E.9, E.10, and E.11.

Notation. Since i is fixed throughout, we drop the (i) superscripts and let $u_k = u_k^{(i)}$ and $v_k = v_k^{(i)}$. We will introduce the following random variables, which are all independent, to simplify the exposition:

- $\xi_j := \theta_k^T v_k^{(j)}$ for $j \neq i$. Thus $\xi_j \sim \mathcal{N}(0, \|\theta_k\|^2)$.
- $\xi_j' := \theta_{k'}^T v_{k'}^{(j)}$ for $j \neq i$. Thus $\xi_j' \sim \mathcal{N}(0, \|\theta_{k'}\|^2)$.
- $\xi_i := (\theta_k^\perp)^T v_k + (\theta_k^\parallel)^T(v_k - \alpha_k u_k)$. Thus $\xi_i \sim \mathcal{N}(0, \|\theta_k^\perp\|^2 + (1-\alpha_k^2)\|\theta_k^\parallel\|^2)$.
- $\xi_i' := (\theta_{k'}^\perp)^T v_{k'}$. Thus $\xi_i' \sim \mathcal{N}(0, \|\theta_{k'}^\perp\|^2 \|\theta_{k'}^\parallel\|^2)$.
- $\zeta_i' := (\theta_{k'}^\parallel)^T(v_{k'} - \alpha_{k'} u_{k'})$. Thus $\zeta_i' \sim \mathcal{N}(0, (1-\alpha_{k'}^2)\|\theta_{k'}^\parallel\|^2)$.
- $y = (\theta_k^\parallel)^T u_k$. Thus $y \sim \mathcal{N}(0, \|\theta_k^\parallel\|^2)$.
- $y' = (\theta_{k'}^\parallel)^T u_{k'}$. Thus $y' \sim \mathcal{N}(0, \|\theta_{k'}^\parallel\|^2)$.
- $\eta_i := (\theta_k^\perp)^T u_k$. Thus $\eta_i \sim \mathcal{N}(0, \|\theta_k^\perp\|^2)$.
- $\eta_i' := (\theta_{k'}^\perp)^T u_{k'}$. Thus $\eta_i' \sim \mathcal{N}(0, \|\theta_{k'}^\perp\|^2)$.

For any such random variable X, we use σ_X^2 to denote its variance. Observe that

$$\frac{p_{ii}p_{ij}}{q_{ii}q_{ij}} = \frac{\exp\left(\theta_k^T u_k \theta_k^T v_k\right)\exp\left(\theta_{k'}^T u_{k'}\theta_{k'}^T v_{k'}\right)}{\mathbb{E}_{j'\sim q}\exp\left(\theta_k^T u_k\theta_k^T v_k^{(j')}\right)\exp\left(\theta_{k'}^T u_{k'}\theta_{k'}^T v_{k'}^{(j')}\right)}\frac{\exp\left(\theta_k^T u_k\theta_k^T v_k^{(j)}\right)\exp\left(\theta_{k'}^T u_{k'}\theta_{k'}^T v_{k'}^{(j)}\right)}{\mathbb{E}_{j'\sim q}\exp\left(\theta_k^T u_k\theta_k^T v_k^{(j')}\right)\exp\left(\theta_{k'}^T u_{k'}\theta_{k'}^T v_{k'}^{(j')}\right)}.$$

We will use the following two claims in the proofs of all four lemmas.

Claim E.14. *For $\beta \in \{\xi_j, \xi_j', \xi_i, \xi_i', \zeta_i', \eta_i, \eta_i', x, x'\}$, let $\bar{\beta}_{j'} := \frac{\partial}{\partial \beta}\left(\theta_k^T u_k \theta_k^T v_k^{(j')} + \theta_{k'}^T u_{k'} \theta_{k'}^T v_{k'}^{(j')}\right)$. Then*

$$\left|\frac{\partial p_{ii} p_{ij}}{\partial \beta}\right| \leq p_{ii} p_{ij} \left(|\bar{\beta}_j| + |\bar{\beta}_i| + 2\mathbb{E}_{j' \sim q} |\bar{\beta}_{j'}|\right).$$

If additionally $\gamma \in \{\xi_j, \xi_j', \xi_i, \xi_i', \zeta_i', \eta_i, \eta_i'\}$ and $\gamma \perp \{\bar{\beta}_{j'}\}_{j' \in [m]}$, then

$$\left|\frac{\partial}{\partial \gamma} \frac{\partial p_{ii} p_{ij}}{\partial \beta}\right| \leq p_{ii} p_{ij} \left(\left(|\bar{\beta}_j| + |\bar{\beta}_i| + 2\mathbb{E}_{j' \sim q}|\bar{\beta}_{j'}|\right)\left(|\bar{\gamma}_j| + |\bar{\gamma}_i| + 2\mathbb{E}_{j' \sim q}|\bar{\gamma}_{j'}|\right) + 2\mathbb{E}_{j' \sim q}|\bar{\beta}_{j'}\bar{\gamma}_{j'}| + 2(\mathbb{E}_{j' \sim q}|\bar{\beta}_{j'}|)(\mathbb{E}_{j' \sim q}|\bar{\gamma}_{j'}|)\right)$$

Proof. By a straightforward quotient-rule computation of the derivative of $\frac{p_{ij}}{q_{ij}}$, recalling that q_{ij} is independent of S, we obtain

$$\frac{\partial p_{ij}}{\partial \beta} = p_{ij}\left(\bar{\beta}_j - \mathbb{E}_{j' \sim q} \bar{\beta}_{j'} p_{ij'}\right).$$

By applying product to the expression above, we obtain

$$\frac{\partial p_{ii} p_{ij}}{\partial \beta} = p_{ii} p_{ij}\left(\bar{\beta}_j + \bar{\beta}_i - 2\mathbb{E}_{j' \sim q} \bar{\beta}_{j'} p_{ij'}\right).$$

Taking absolute values and using the fact that $p_{ij'} \leq 1$, we obtain the first result.

Next we take the derivative of p_{ij} with respect to both β and γ. Using the expression above for $\frac{\partial p_{ij}}{\partial \beta}$, we obtain

$$\frac{\partial}{\partial \gamma} \frac{\partial p_{ij}}{\partial \beta} = p_{ij}\left(\left(\bar{\beta}_j - \mathbb{E}_{j' \sim q}\bar{\beta}_{j'} p_{ij'}\right)\left(\bar{\gamma}_j - \mathbb{E}_{j' \sim q}\bar{\gamma}_{j'} p_{ij'}\right) - \mathbb{E}_{j' \sim q}\bar{\beta}_{j'}\bar{\gamma}_{j'} p_{ij'} + (\mathbb{E}_{j' \sim q}\bar{\beta}_{j'} p_{ij'})(\mathbb{E}_{j' \sim q}\bar{\gamma}_{j'} p_{ij'})\right),$$

and

$$\frac{\partial}{\partial \gamma} \frac{\partial p_{ii} p_{ij}}{\partial \beta} = p_{ii} p_{ij}\left(\left(\bar{\beta}_j + \bar{\beta}_i - 2\mathbb{E}_{j' \sim q}\bar{\beta}_{j'} p_{ij'}\right)\left(\bar{\gamma}_j + \bar{\gamma}_i - 2\mathbb{E}_{j' \sim q}\bar{\gamma}_{j'} p_{ij'}\right) - 2\mathbb{E}_{j' \sim q}\bar{\beta}_{j'}\bar{\gamma}_{j'} p_{ij'} + 2(\mathbb{E}_{j' \sim q}\bar{\beta}_{j'} p_{ij'})(\mathbb{E}_{j' \sim q}\bar{\gamma}_{j'} p$$

The second result follows by taking absolute values and the fact that $p_{ij'} \leq 1$. ☐

Claim E.15.

$$\frac{p_{ij}}{q_{ij}} \leq \exp\left(|\theta_k^T u_k \theta_k^T v_k^{(j)}|\right) \exp\left(|\theta_{k'}^T u_{k'} \theta_{k'}^T v_{k'}^{(j)}|\right) \mathbb{E}_{j' \sim q}\left[\exp\left(|\theta_k^T u_k \theta_k^T v_k^{(j')}|\right) \exp\left(|\theta_{k'}^T u_{k'} \theta_{k'}^T v_{k'}^{(j')}|\right)\right].$$

Proof. This follows directly from using Jenson's inequality on the distribution $j' \sim q$ to show that

$$\begin{aligned}
\frac{1}{\mathbb{E}_{j' \sim q}\left[\exp\left(\theta_k^T u_k \theta_k^T v_k^{(j')}\right) \exp\left(\theta_{k'}^T u_{k'} \theta_{k'}^T v_{k'}^{(j')}\right)\right]} &\leq \mathbb{E}_{j' \sim q}\left[\exp\left(-\theta_k^T u_k \theta_k^T v_k^{(j')}\right) \exp\left(-\theta_{k'}^T u_{k'} \theta_{k'}^T v_{k'}^{(j')}\right)\right] \\
&\leq \mathbb{E}_{j' \sim q}\left[\exp\left(|\theta_k^T u_k \theta_k^T v_k^{(j')}|\right) \exp\left(|\theta_{k'}^T u_{k'} \theta_{k'}^T v_{k'}^{(j')}|\right)\right].
\end{aligned}$$

☐

Claim E.16.

$$\left|1 - \frac{p_{ij}}{q_{ij}}\right| \leq Z_j - 1,$$

where $Z_j := \exp\left(|\theta_k^T u_k \theta_k^T v_k^{(j)}|\right) \exp\left(|\theta_{k'}^T u_{k'} \theta_{k'}^T v_{k'}^{(j)}|\right) \mathbb{E}_{j' \sim q}\left[\exp\left(|\theta_k^T u_k \theta_k^T v_k^{(j')}|\right) \exp\left(|\theta_{k'}^T u_{k'} \theta_{k'}^T v_{k'}^{(j')}|\right)\right]$.

Proof. Note that for any $x \geq 0$, we have $|1 - x| \leq \max\left(x - 1, \frac{1}{x} - 1\right)$. By Claim E.15, $\frac{p_{ij}}{q_{ij}} - 1$ is at most the desired value given in this claim.

Now

$$\begin{aligned}
\frac{q_{ij}}{p_{ij}} &= \frac{\mathbb{E}_{j' \sim q}\left[\exp\left(\theta_k^T u_k \theta_k^T v_k^{(j')}\right) \exp\left(\theta_{k'}^T u_{k'} \theta_{k'}^T v_{k'}^{(j')}\right)\right]}{\exp\left(\theta_k^T u_k \theta_k^T v_k^{(j)}\right) \exp\left(\theta_{k'}^T u_{k'} \theta_{k'}^T v_{k'}^{(j)}\right)} \\
&\leq \exp\left(|\theta_k^T u_k \theta_k^T v_k^{(j)}|\right) \exp\left(|\theta_{k'}^T u_{k'} \theta_{k'}^T v_{k'}^{(j)}|\right) \mathbb{E}_{j' \sim q}\left[\exp\left(|\theta_k^T u_k \theta_k^T v_k^{(j')}|\right) \exp\left(|\theta_{k'}^T u_{k'} \theta_{k'}^T v_{k'}^{(j')}|\right)\right].
\end{aligned}$$

This yields the claim. ☐

Proof of Lemma E.8. Expanding $h(S) - h_1(S)$, we see that we need to control the following terms:

1. (a) $\left|\mathbb{E}_S\left[p_{ii}p_{ij}\left(\eta_i'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\mu_k^T u_k \xi_j\right)\right]\right|$, (b) $\left|\mathbb{E}_S\left[p_{ii}p_{ij}\left(y'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\mu_k^T u_k \xi_j\right)\right]\right|$
2. (a) $\left|\mathbb{E}_S\left[p_{ii}p_{ij}\left(\eta_i'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\mu_k^T u_k \xi_i\right)\right]\right|$, (b) $\left|\mathbb{E}_S\left[p_{ii}p_{ij}\left((y'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\mu_k^T u_k \xi_i\right)\right]\right|$
3. (a) $\left|\alpha_k\mathbb{E}_S\left[p_{ii}p_{ij}\left(\eta_i'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\mu_k^T u_k y\right)\right]\right|$, (b) $\left|\alpha_k\mathbb{E}_S\left[p_{ii}p_{ij}\left(y'(-x\xi_i')\right)\left(\mu_k^T u_k y\right)\right]\right|$
4. (a) $\left|\mathbb{E}_S\left[p_{ii}p_{ij}\left(\eta_i'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\xi_i(v_k - v_k^{(j)})^T\mu_k\right)\right]\right|$
   (b) $\left|\mathbb{E}_S\left[p_{ii}p_{ij}\left(y'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\xi_i(v_k - v_k^{(j)})^T\mu_k\right)\right]\right|$
5. (a) $\left|\alpha_k\mathbb{E}_S\left[p_{ii}p_{ij}\left(\eta_i'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(y(v_k - v_k^{(j)})^T\mu_k\right)\right]\right|$
   (b) $\left|\mathbb{E}_S\left[p_{ii}p_{ij}\left(y'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(y(v_k - \alpha_k u_k - v_k^{(j)})^T\mu_k\right)\right]\right|$

We begin by bounding the terms where the expression after $p_{ii}p_{ij}$ has two independent mean-0 terms, mainly (1a), (2a), (4a). The first step is to apply Stein's Lemma (Lemma E.3) twice to these two terms, which we will call β and γ. Let $\beta\gamma g(S \setminus \{\beta,\gamma\})$ be the terms after $p_{ii}p_{ij}$. Then we have

$$|\mathbb{E}_S[p_{ii}p_{ij}\beta\gamma g(S \setminus \{\beta,\gamma\})]| \leq \sigma_\beta^2\sigma_\gamma^2\left|\mathbb{E}_S\left[\left|\frac{\partial}{\partial\gamma}\frac{\partial p_{ii}p_{ij}}{\partial\beta}\right||g(S \setminus \{\beta,\gamma\})|\right]\right|.$$

Next we apply the final result in Claim E.14 to bound the absolute value of $\left|\frac{\partial}{\partial\gamma}\frac{\partial p_{ii}p_{ij}}{\partial\beta}\right|$. Once we do this, we achieve

$$|\mathbb{E}_S[p_{ii}p_{ij}\beta\gamma g(S \setminus \{\beta,\gamma\})]| \leq \sigma_\beta^2\sigma_\gamma^2 q_{ii}q_{ij}\mathbb{E}_S\left[Z|g(S \setminus \{\beta,\gamma\})|\sum_{j',\ell\in[m]} c_{j',\ell}|\bar{\beta_{j'}}||\bar{\gamma_\ell}|\right],$$

where $\sum_{j',\ell\in[m]} c_{j',\ell} \leq C$ for some constant C, and $Z := \frac{p_{ii}p_{ij}}{q_{ii}q_{ij}}$. Finally, we use the bound on Z from Claim E.15, and then Lemma E.4 to take the expectation over S, iteratively applying Lemma E.4 to each variable in S. Thus we have, for some (different) constant C,

1. $\left|\mathbb{E}_S\left[p_{ii}p_{ij}\left(\eta_i'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\mu_k^T u_k \xi_j\right)\right]\right| \leq Cq_{ii}q_{ij}\sigma_{\eta_i'}^2\sigma_{\xi_j}^2\|\theta_{k'}\|\|\theta_k\| = Cq_{ii}q_{ij}\|\theta_{k'}^\perp\|^2\|\theta_{k'}\|\|\theta_k\|^3 \leq Cq_{ii}q_{ij}\|\theta_{k'}\|^3\|\theta_k\|^3$.
2. $\left|\mathbb{E}_S\left[p_{ii}p_{ij}\left(\eta_i'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\mu_k^T u_k \xi_i\right)\right]\right| \leq Cq_{ii}q_{ij}\sigma_{\eta_i'}^2\sigma_{\xi_i}^2\|\theta_{k'}\|\|\theta_k\| \leq Cq_{ii}q_{ij}\|\theta_{k'}^\perp\|^2\|\theta_{k'}\|\|\theta_k\|^3 \leq Cq_{ii}q_{ij}\|\theta_{k'}\|^3\|\theta_k\|^3$.
3. $\left|\mathbb{E}_S\left[p_{ii}p_{ij}\left(\eta_i'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\xi_i(v_k - v_k^{(j)})^T\mu_k\right)\right]\right| \leq Cq_{ii}q_{ij}\sigma_{\eta_i'}^2\sigma_{\xi_i}^2\|\theta_{k'}\|\|\theta_k\| \leq Cq_{ii}q_{ij}\|\theta_{k'}^\perp\|^2\|\theta_{k'}\|\|\theta_k\|^3 \leq Cq_{ii}q_{ij}\|\theta_{k'}\|^3\|\theta_k\|^3$.

Now we consider the remaining 7 terms. Here we decompose the expression inside the expectation as $p_{ii}p_{ij}\beta g(S \setminus \beta)$, where $\beta \in S$. We proceed as before, but we only apply Stein's Lemma once, to β. Applying Steins, the expression for $\frac{\partial p_{ii}p_{ij}}{\partial\beta}$ given in the first result of Claim E.14, we obtain

$$|\mathbb{E}_S[p_{ii}p_{ij}\beta g(S \setminus \beta)]| \leq \sigma_\beta^2\left|\mathbb{E}_S\left[\left|\frac{\partial p_{ii}p_{ij}}{\partial\beta}\right||g(S \setminus \beta)|\right]\right| \leq \sigma_\beta^2 q_{ii}q_{ij}\mathbb{E}_S\left[Z|g(S \setminus \beta)|\sum_{j'\in[m]} c_{j'}|\bar{\beta_{j'}}|\right], \tag{23}$$

where $\sum_{j'\in[m]} c_{j'} \leq C$ for some constant C, and $Z := \frac{p_{ii}p_{ij}}{q_{ii}q_{ij}}$. Finally, we plug in a bound for Z in Claim E.15, an use Lemma E.4 to take the expectation over S, again iteratively over each variable.

Thus we have, for some (different) constant C,

1. $\left|\mathbb{E}_S\left[p_{ii}p_{ij}\left(y'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\mu_k^T u_k \xi_j\right)\right]\right| \leq Cq_{ii}q_{ij}\sigma_{\xi_j}^2\|\theta_k\|\|\theta_{k'}^{\|}\| = Cq_{ii}q_{ij}\|\theta_k\|^3\|\theta_{k'}^{\|}\|$.
2. $\left|\mathbb{E}_S\left[p_{ii}p_{ij}\left(y'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\mu_k^T u_k \xi_i\right)\right]\right| \leq Cq_{ii}q_{ij}\sigma_{\xi_i}^2\|\theta_k\|\|\theta_{k'}^{\|}\| \leq Cq_{ii}q_{ij}\|\theta_k\|^3\|\theta_{k'}^{\|}\|$.
3. $\left|\alpha_k\mathbb{E}_S\left[p_{ii}p_{ij}\left(\eta_i'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\mu_k^T u_k y\right)\right]\right| \leq C\alpha_k q_{ii}q_{ij}\sigma_{\eta_i'}^2\|\theta_{k'}\|\|\theta_k^{\|}\| = C\alpha_k q_{ii}q_{ij}\|\theta_{k'}^{\perp}\|^2\|\theta_{k'}\|\|\theta_k^{\|}\| \leq C\alpha_k q_{ii}q_{ij}\|\theta_{k'}\|^3\|\theta_k^{\|}\|$.
4. $\left|\alpha_k\mathbb{E}_S\left[p_{ii}p_{ij}\left(y'(-x\zeta_i')\right)\left(\mu_k^T u_k y\right)\right]\right| \leq C\alpha_k q_{ii}q_{ij}\sigma_{\zeta_i'}^2\|\theta_{k'}\|\|\theta_k^{\|}\| = C\alpha_k q_{ii}q_{ij}\|\theta_{k'}^{\|}\|^2\|\theta_{k'}\|\|\theta_k^{\|}\|$.
5. $\left|\mathbb{E}_S\left[p_{ii}p_{ij}\left(y'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\xi_i(v_k - v_k^{(j)})^T\mu_k\right)\right]\right| \leq Cq_{ii}q_{ij}\sigma_{\xi_i}^2\|\theta_k\|\|\theta_{k'}^{\|}\| \leq Cq_{ii}q_{ij}\|\theta_k\|^3\|\theta_{k'}^{\|}\|$.
6. $\left|\alpha_k\mathbb{E}_S\left[p_{ii}p_{ij}\left(\eta_i'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(x(v_k - v_k^{(j)})^T\mu_k\right)\right]\right| \leq C\alpha_k q_{ii}q_{ij}\sigma_{\eta_i'}^2\|\theta_{k'}\|\|\theta_k^{\|}\| = C\alpha_k q_{ii}q_{ij}\|\theta_{k'}^{\perp}\|^2\|\theta_{k'}\|\|\theta_k^{\|}\| \leq C\alpha_k q_{ii}q_{ij}\|\theta_{k'}\|^3\|\theta_k^{\|}\|$.
7. $\left|\mathbb{E}_S\left[p_{ii}p_{ij}\left(y'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(x(v_k - \alpha_k u_k - v_k^{(j)})^T\mu_k\right)\right]\right| \leq Cq_{ii}q_{ij}\sigma_x^2\|\theta_k\|\|\theta_{k'}^{\|}\| = Cq_{ii}q_{ij}\|\theta_k^{\|}\|^2\|\theta_k\|\|\theta_{k'}^{\|}\|$.

Combining the bounds on these 10 terms proves the lemma:

$$\left|\mathbb{E}_S\left[p_{ii}p_{ij}\mu_k^T(h(S) - h_1(S))\right]\right| \leq Cq_{ii}q_{ij}\left(\|\theta_{k'}\|^3\|\theta_k\|^3 + \|\theta_{k'}^{\|}\|\|\theta_k\|^3 + \alpha_k\left(\|\theta_{k'}\|^3\|\theta_k^{\|}\|\right)\right).$$

□

Proof of Lemma E.10. The proof of Lemma E.10 is nearly identical, besides some differences in the terms we need to bound. We list them below:

1. (a) $\left|\mathbb{E}_S\left[p_{ii}p_{ij}\left(\eta_i'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\theta_k^T u_k \xi_j\right)\right]\right|$ (b) $\left|\mathbb{E}_S\left[p_{ii}p_{ij}\left(y'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\theta_k^T u_k \xi_j\right)\right]\right|$
2. (a) $\left|\mathbb{E}_S\left[p_{ii}p_{ij}\left(\eta_i'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\theta_k^T u_k \xi_i\right)\right]\right|$ (b) $\left|\mathbb{E}_S\left[p_{ii}p_{ij}\left((y'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\theta_k^T u_k \xi_i\right)\right]\right|$
3. (a) $\left|\alpha_k\mathbb{E}_S\left[p_{ii}p_{ij}\left(\eta_i'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\theta_k^T u_k y\right)\right]\right|$ (b) $\left|\alpha_k\mathbb{E}_S\left[p_{ii}p_{ij}\left(y'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\eta_i y\right)\right]\right|$
4. $\left|\alpha_k\mathbb{E}_S\left[p_{ii}p_{ij}\left(y'\left(-x\zeta_i'\right)\right)\left(\theta_k^T u_k y\right)\right]\right|$

We use the same approach as before. For the terms (1a) and (2a) we apply Stein's Lemma to (η_i', ξ_j) and (η_i', ξ_i) respectively. For (1b), (2b), (3a) and (3b) and (4), we apply Stein's Lemma to ξ_j, ξ_i, η_i', η_i, and ξ_i' respectively. Using Claim E.15 and then Lemma E.4 as before, we obtain the following result:

1. $\left|\mathbb{E}_S\left[p_{ii}p_{ij}\left(\eta_i'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\theta_k^T u_k \xi_j\right)\right]\right| \leq Cq_{ii}q_{ij}\sigma_{\eta_i'}^2\sigma_{\xi_j}^2\|\theta_{k'}\|\|\theta_k\|\|\theta_k\| = Cq_{ii}q_{ij}\|\theta_{k'}^{\perp}\|^2\|\theta_{k'}\|\|\theta_k\|^4 \leq Cq_{ii}q_{ij}\|\theta_{k'}\|^3\|\theta_k\|^4$.
2. $\left|\mathbb{E}_S\left[p_{ii}p_{ij}\left(\eta_i'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\theta_k^T u_k \xi_i\right)\right]\right| \leq Cq_{ii}q_{ij}\sigma_{\eta_i'}^2\sigma_{\xi_i}^2\|\theta_{k'}\|\|\theta_k\|\|\theta_k\| \leq Cq_{ii}q_{ij}\|\theta_{k'}^{\perp}\|^2\|\theta_{k'}\|\|\theta_k\|^4 \leq Cq_{ii}q_{ij}\|\theta_{k'}\|^3\|\theta_k\|^4$.
3. $\left|\mathbb{E}_S\left[p_{ii}p_{ij}\left(y'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\theta_k^T u_k \xi_j\right)\right]\right| \leq Cq_{ii}q_{ij}\sigma_{\xi_j}^2\|\theta_k\|\|\theta_k\|\|\theta_{k'}^{\|}\| = Cq_{ii}q_{ij}\|\theta_k\|^4\|\theta_{k'}^{\|}\|$

4. $\left|\mathbb{E}_S\left[p_{ii}p_{ij}\left(y'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\theta_k^T u_k \xi_i\right)\right]\right| \leq Cq_{ii}q_{ij}\sigma_{\xi_i}^2\|\theta_k\|\|\theta_k\|\|\theta_{k'}^{\|}\| \leq Cq_{ii}q_{ij}\|\theta_k\|^4\|\theta_{k'}^{\|}\|$

5. $\left|\alpha_k\mathbb{E}_S\left[p_{ii}p_{ij}\left(\eta_i'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\theta_k^T u_k y\right)\right]\right| \leq C\alpha_k q_{ii}q_{ij}\sigma_{\eta_i'}^2\|\theta_{k'}\|\|\theta_k\|\|\theta_k^{\|}\| = C\alpha_k q_{ii}q_{ij}\|\theta_{k'}^{\perp}\|^2\|\theta_{k'}\|\|\theta_k\|\|\theta_k^{\|}\|$

6. $\left|\alpha_k\mathbb{E}_S\left[p_{ii}p_{ij}\left(y'\left(\mu_{k'}^T u_{k'} - x\mu_{k'}^T v_{k'}\right)\right)\left(\eta_i y\right)\right]\right| \leq C\alpha_k q_{ii}q_{ij}\sigma_{\eta_i}^2\|\theta_k\|\|\theta_{k'}^{\|}\|\|\theta_k^{\|}\| = C\alpha_k q_{ii}q_{ij}\|\theta_k^{\perp}\|^2\|\theta_k\|\|\theta_{k'}^{\|}\|\|\theta_k^{\|}\|$.

7. $\left|\alpha_k\mathbb{E}_S\left[p_{ii}p_{ij}\left(y'\left(-x\zeta_i'\right)\right)\left(\theta_k^T u_k y\right)\right]\right| \leq C\alpha_k q_{ii}q_{ij}\sigma_{\zeta_i'}^2\|\theta_{k'}\|\|\theta_k\|\|\theta_k^{\|}\| \leq C\alpha_k q_{ii}q_{ij}\|\theta_{k'}^{\|}\|^2\|\theta_{k'}\|\|\theta_k\|\|\theta_k^{\|}\|$.

Combining the bounds on these 7 terms, proves the lemma:

$$\left|\mathbb{E}_S\left[p_{ii}p_{ij}\theta_k^T(h(S)-h_1(S))\right]\right| \leq Cq_{ii}q_{ij}\left(\|\theta_{k'}\|^3\|\theta_k\|^4 + \|\theta_{k'}^{\|}\|\|\theta_k\|^4 + \alpha_k\left(\|\theta_{k'}\|^3\|\theta_k\|\|\theta_k^{\|}\| + \|\theta_{k'}^{\|}\|\|\theta_k\|^3\|\theta_k^{\|}\|\right)\right).$$

□

We now prove the lemmas on the non-junk terms.

Proof of Lemma E.9.

$$\begin{aligned}
&\mathbb{E}_S\left[p_{ii}p_{ij}\left((\theta_{k'}^{\|})^T u_{k'}u_{k'}^T\mu_{k'}\right)\left(2\mu_k^T u_k\alpha_k(\theta_k^{\|})^T u_k\right)\right] \\
&\quad= \mathbb{E}_S\left[q_{ii}q_{ij}\left((\theta_{k'}^{\|})^T u_{k'}u_{k'}^T\mu_{k'}\right)\left(2\mu_k^T u_k\alpha_k(\theta_k^{\|})^T u_k\right)\right] + \mathbb{E}_S\left[(p_{ii}p_{ij}-q_{ii}q_{ij})\left((\theta_{k'}^{\|})^T u_{k'}u_{k'}^T\mu_{k'}\right)\left(2\mu_k^T u_k\alpha_k(\theta_k^{\|})^T u_k\right)\right] \\
&\quad= 2\alpha_k q_{ii}q_{ij}\theta_{k'}^T\mu_{k'}\theta_k^T\mu_k + 2\alpha_k q_{ii}q_{ij}\mathbb{E}_S\left[\left(\frac{p_{ii}p_{ij}}{q_{ii}q_{ij}}-1\right)\left((\theta_{k'}^{\|})^T u_{k'}u_{k'}^T\mu_{k'}\right)\left(\mu_k^T u_k(\theta_k^{\|})^T u_k\right)\right].
\end{aligned}$$

Now by Claim E.16, we have $\left|\frac{p_{ii}p_{ij}}{q_{ii}q_{ij}}-1\right| \leq Z_iZ_j - 1$ (where the variable's Z_i, Z_j are defined in the Claim E.16) so

$$\begin{aligned}
\left|\mathbb{E}_S\left[\left(\frac{p_{ii}p_{ij}}{q_{ii}q_{ij}}-1\right)\left((\theta_{k'}^{\|})^T u_{k'}u_{k'}^T\mu_{k'}\right)\left(\mu_k^T u_k(\theta_k^{\|})^T u_k\right)\right]\right| &\leq \mathbb{E}_S\left[(Z_iZ_j-1)\left|(\theta_{k'}^{\|})^T u_{k'}u_{k'}^T\mu_{k'}\right|\left|\mu_k^T u_k(\theta_k^{\|})^T u_k\right|\right] \\
&\leq C\left(\|\theta_k\|^2 + \|\theta_{k'}\|^2\right)\|\theta_{k'}^{\|}\|\|\theta_k^{\|}\|.
\end{aligned}$$

Here the second inequality follows from applying Lemma E.5 first, and then Lemma E.4 repeatedly for the remainder of the variables in S. This proves the lemma. Note that we need to apply Lemma E.5 several times to a single variable $X \in S$. Indeed we can write

$$\begin{aligned}
(Z_iZ_j-1)\left|(\theta_{k'}^{\|})^T u_{k'}u_{k'}^T\mu_{k'}\right|\left|\mu_k^T u_k(\theta_k^{\|})^T u_k\right| &= \left(\mathbb{E}_\ell\exp(|t_\ell X|)S_\ell - 1\right)B|X|^c \\
&= \left(\mathbb{E}_\ell S_\ell(\exp(|t_\ell X|)-1)\right)B|X|^c + \left(\mathbb{E}_\ell S_\ell - 1\right)B|X|^c
\end{aligned}$$

for some distribution on ℓ, and for some terms S_ℓ, t_ℓ, and B that are independent of X, and $c \in \{0,1,2\}$. Then to take the expectation of this term over X, we first apply Lemma E.5 to on X to the first term, and iteratively apply Lemma E.5 to the random variables appearing in the next terms. □

Proof of Lemma E.11.

$$\begin{aligned}
\frac{1}{1-x^2}\mathbb{E}_S\left[p_{ii}p_{ij}\theta_k^T h_1(S)\right] &= \mathbb{E}_S\left[p_{ii}p_{ij}\left((\theta_{k'}^{\|})^T u_{k'}u_{k'}^T\mu_{k'}\right)\left(2(\theta_k^{\|})^T u_k\alpha_k(\theta_k^{\|})^T u_k\right)\right] \\
&= \mathbb{E}_S\left[q_{ii}q_{ij}\left((\theta_{k'}^{\|})^T u_{k'}u_{k'}^T\mu_{k'}\right)\left(2\alpha_k((\theta_k^{\|})^T u_k)^2\right)\right] \\
&\qquad+ \mathbb{E}_S\left[(p_{ii}p_{ij}-q_{ii}q_{ij})\left((\theta_{k'}^{\|})^T u_{k'}u_{k'}^T\mu_{k'}\right)\left(2\alpha_k((\theta_k^{\|})^T u_k)^2\right)\right] \\
&= 2\alpha_k q_{ii}q_{ij}\theta_{k'}^T\mu_{k'}\|\theta_k^{\|}\|^2 + 2\alpha_k q_{ii}q_{ij}\mathbb{E}_S\left[\left(\frac{p_{ii}p_{ij}}{q_{ii}q_{ij}}-1\right)\left((\theta_{k'}^{\|})^T u_{k'}u_{k'}^T\mu_{k'}\right)\left((\theta_k^{\|})^T u_k\right)^2\right].
\end{aligned}$$

Now by Claim E.16, we have $\left|\frac{p_{ii}p_{ij}}{q_{ii}q_{ij}} - 1\right| \leq Z_i Z_j - 1$, so

$$\left|\mathbb{E}_S\left[\left(\frac{p_{ii}p_{ij}}{q_{ii}q_{ij}} - 1\right)\left((\theta_{k'}^{\|})^T u_{k'} u_{k'}^T \mu_{k'}\right)\left((\theta_k^{\|})^T u_k\right)^2\right]\right| \leq \mathbb{E}_S\left[(Z_i Z_j - 1)\left|(\theta_{k'}^{\|})^T u_{k'} u_{k'}^T \mu_{k'}\right|\left((\theta_k^{\|})^T u_k\right)^2\right]$$
$$\leq C\left(\|\theta_k\|^2 + \|\theta_{k'}\|^2\right)\|\theta_{k'}^{\|}\|\theta_k^{\|}\|^2,$$

Again the second inequality follows from applying Lemma E.5 first (several times as described in the previous lemma), and then Lemma E.4 repeatedly for the remainder of the variables in S. Taking absolute values proves the lemma.

□