THE UNIVERSITY OF CHICAGO

STATISTICAL LEARNING AND INFERENCE IN WEAKLY SPECIFIED SETTINGS: SHIFTED DISTRIBUTIONS AND UNLABELED DATA

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES DIVISION
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
YU GUI

CHICAGO, ILLINOIS
APRIL 2025

Copyright © 2025 by Yu Gui

All Rights Reserved

To my family, and in memory of my grandfather, Qingguo.

CONTENTS

LIST OF FIGURES viii

ACKNOWLEDGMENTS xi

ABSTRACT xiv

1 INTRODUCTION 1
1.1 Distribution-free statistical inference 2
1.2 Statistical learning and inference with distribution shift 3
1.3 Leveraging unlabeled data to enhance statistical tasks 4
1.4 Organization of thesis 5

2 DISTRIBUTION-FREE UNCERTAINTY QUANTIFICATION FOR MATRIX COMPLETION 6
2.1 Introduction 6
2.1.1 Conformal prediction for distribution-free uncertainty quantification . 7
2.1.2 Main contributions 8
2.2 Problem setup 9
2.3 Conformalized matrix completion 11
2.3.1 Exchangeability and weighted exchangeability 12
2.3.2 Theoretical guarantee 15
2.3.3 Examples for missingness 16
2.4 Simulation studies 18
2.4.1 Synthetic dataset 18
2.4.2 Real data application 23
2.5 Discussion 24

3 DISTRIBUTIONALLY ROBUST INFERENCE WITH SHAPE CONSTRAINTS 26
3.1 Introduction 26
3.1.1 Problem formulation 27
3.1.2 Prior work: distributionally robust learning 30
3.1.3 Our proposal: iso-DRL 30
3.1.4 Organization of paper 34
3.2 The distributional robustness framework 35
3.2.1 Worst-case excess risk with DRL 37
3.3 Worst-case excess risk with an isotonic constraint 40
3.3.1 Equivalent formulation 41
3.3.2 Setting: iso-DRL with estimated density ratio 42
3.4 Estimation of worst-case excess risk with isotonic constraint 43
3.4.1 Plug-in estimators 43
3.4.2 Computation: estimation after projection 45
3.4.3 Performance guarantees for plug-in estimators 46

3.4.4 The role of the isotonic constraint . . . 49
3.5 Numerical experiments . . . 49
3.5.1 Synthetic dataset . . . 52
3.5.2 Real data: wine quality dataset . . . 54
3.6 Additional related work . . . 56
3.7 Discussion . . . 58

4 UNRAVELING PROJECTION HEADS IN CONTRASTIVE LEARNING: INSIGHTS FROM EXPANSION AND SHRINKAGE . . . 60
4.1 Introduction . . . 61
4.1.1 Classical unsupervised learning based on encoders and decoders . . . 61
4.1.2 An encoder-projector framework for contrastive learning . . . 62
4.1.3 Puzzling effect of projectors on the representations . . . 65
4.1.4 Our contributions . . . 66
4.1.5 Paper organization . . . 67
4.1.6 Notations . . . 67
4.2 Empirical discovery: expansion and shrinkage . . . 68
4.2.1 Insights from spectral decomposition . . . 68
4.2.2 Expansion/shrinkage affects generalization . . . 69
4.2.3 A simple simulation and heuristic explanations . . . 70
4.3 Feature-level modeling via GMM . . . 72
4.3.1 Model setup . . . 72
4.3.2 Why fixing a well-trained encoder? . . . 74
4.4 Contrastive loss drives expansion/shrinkage . . . 75
4.4.1 Quantitative results under simplification . . . 75
4.4.2 Extensions to the general case . . . 79
4.5 Effect of projectors on generalization . . . 82
4.5.1 One-parameter projector model . . . 82
4.5.2 Confirming folklore: invariance of test errors in low-dimension regime 83
4.5.3 High-dimensional regime: inductive bias matters . . . 84
4.5.4 Precise asymptotic characterization in high dimension . . . 85
4.6 Extensions to inhomogeneous feature augmentation . . . 87
4.6.1 Feature augmentation with spiked covariance . . . 87
4.6.2 Simultaneous expansion/shrinkage . . . 92
4.7 Discussion and related work . . . 93
4.7.1 Connections to emerging empirical phenomena . . . 93
4.7.2 Related work . . . 95

A APPENDIX FOR CHAPTER 2 . . . 119
A.1 Proofs . . . 119
A.1.1 Proof of Lemma 2.3.1 . . . 119
A.1.2 Proof of Theorem 2.3.2 . . . 121
A.1.3 Proofs for examples in Section 2.3.3 . . . 124
A.2 Additional numerical experiments . . . 127

A.2.1 Comparison between one-shot and exact weighted approaches 128
A.2.2 Local coverage of cmc 128
A.2.3 Additional results for homogeneous misingness 129
A.2.4 Estimation error for one-bit matrix estimation 130
A.2.5 Additional results for heterogeneous missingness 130
A.2.6 Misspecified sampling probability 131
A.2.7 Additional results for the sales dataset 132
A.3 Additional details of algorithms and extensions 135
A.3.1 Extension to likelihood-based scores and categorical matrices 135
A.3.2 Full conformalized matrix completion 137
A.3.3 Exact split conformalized matrix completion 137

B APPENDIX FOR CHAPTER 3 140
B.1 Proofs of results in Section 3.2 140
B.1.1 Proof of Proposition 3.2.2 140
B.2 Proofs of results in Section 3.3 142
B.2.1 Preliminaries 142
B.2.2 Proof of Theorem 3.3.1 144
B.2.3 Proof of Proposition 3.3.2 148
B.2.4 A misspecified isotonic constraint 150
B.2.5 Proof of Proposition B.2.1 151
B.3 Proofs of results in Section 3.4 153
B.3.1 Proof of Proposition 3.4.2 153
B.3.2 Proof of Theorem 3.4.4 154
B.3.3 Proof of Lemma 3.4.5 155
B.3.4 Proof of Lemma 3.4.6 156
B.3.5 Proof of Lemma 3.4.7 158
B.3.6 The role of the isotonic constraint 160
B.4 Additional simulation results 162
B.4.1 Simulations for iso-DRL under componentwise order 162
B.4.2 Details for the wine quality data set: a proxy of the oracle KL-divergence 163

C APPENDIX FOR CHAPTER 4 165
C.1 Experiments: details and extensions 165
C.1.1 Experiment setup and details 165
C.1.2 Additional experiments 168
C.2 Proofs for Section 4.4 172
C.2.1 Proof of Proposition 4.4.1 172
C.2.2 Justification of approximate loss 176
C.2.3 Proof of Theorem 4.4.3 177
C.2.4 Proof of Proposition 4.4.4 179
C.2.5 Proof of Proposition 4.4.5 187
C.2.6 An example of expansion/shrinkage transition 191
C.3 Proofs for Section 4.5 192

C.3.1 Proof of Theorem 4.5.1 192
C.3.2 High-dimensional regime: max-margin classifier 204
C.4 Proof for Section 4.6 236
C.4.1 Proof of Proposition 4.6.1 236
C.4.2 Proof of Theorem 4.6.2 238
C.4.3 Proof of Corollary 4.6.3 244

LIST OF FIGURES

2.1 Comparison between $cmc - als$ and als. 20
2.2 Histogram of standardized scores for als and prediction lower and upper bounds for 50 distinct unobserved entries. 21
2.3 Comparison between cmc with true and estimated weights under heterogeneous missingness. 23
2.4 Comparison between conformalized and model-based matrix completion with sales dataset. 24

3.1 Coverage rate and average width of intervals for the wine quality dataset. The red dashed line (in the left-hand plot) marks the nominal coverage level, $1 - \alpha = 90\%$. 33
3.2 Results in the well-specified setting. 53
3.3 Results in the misspecified setting. 53
3.4 Results with varying ρ. 54
3.5 Results for wine quality dataset. 55

4.1 Encoder-decoder framework vs. encoder-projector framework. 64
4.2 Positive and negative pairs after image augmentation. 65
4.3 Results with the pretrained encoder and a one-layer linear projector $\mathbf{W} \in \mathbb{R}^{512\times512}$ under the standard SimCLR loss (ResNet-18 and 10-class STL-10 dataset). **Left**: singular values of $\mathbf{W}$ with varying temperature τ. **Middle**: cumulative score plot $\text{score}_i = \sum_{j\le i}\langle \boldsymbol{v}_j, \boldsymbol{\mu}_{c_1,c_2}\rangle^2/\|\boldsymbol{\mu}_{c_1,c_2}\|^2$ where $(c_1, c_2) = (7, 2)$. **Right**: downstream task accuracy using features $\boldsymbol{h}$ versus using embeddings $\boldsymbol{z}$ on the test set containing all 10 classes. 69
4.4 Contrastive loss drives expansion or shrinkage. **Left:** we generate clean 2-component GMM data. **Middle and right:** we use SimCLR loss for training 100-dim data, and then visualize embeddings on a 2D plane under large/small augmentation and large/small temperature (middle/right plot). Also, we show the projection of embeddings onto a circle with varying radii, which is $\boldsymbol{z}_i/\|\boldsymbol{z}_i\|$ used for calculating cosine similarities. **Bottom text:** expansion decreases test error, while shrinkage reduces signals and increases test error. 71
4.5 Pretrained STL-10 features from the standard SimCLR model exhibit a strong cluster structure. **Left**: positive pairwise margins of linear classifier imply pairwise linear separability. **Right**: 2D visualization of 3-class subsets of features after 30 epochs. The cluster structure emerges quickly. 75
4.6 Expansion/shrinkage phase transition for the configuration $n = 2000$, $p = 50$, $\|\boldsymbol{\mu}\| = 5$. **Left**: heatmap of expansion measure $T^* := T(\mathbf{W}^*)$; **green curve**: theoretical phase transition. **Right**: T^* vs. τ with different fixed σ_{aug}. 79
4.7 Test error is monotone in η under the one-parameter projector model. We apply a linear transform $\mathbf{W} \in \mathcal{W}$ on train data from 2-GMM, and calculate the test errors of the max-margin classifiers. Note $\eta = 0$ (vertical dashed) is equivalent to no transform. 87

4.8 GMM with inhomogeneous augmentation. (a) **Left**: $p = 100$ singular values of $\mathbf{W}$ in descending order with varying τ indicated by different colors. (a) **Middle**: Cumulative score for $\boldsymbol{\mu}$, i.e., $\text{score}_i(\boldsymbol{\mu}) = \sum_{j \le i} \langle \boldsymbol{\mu}/\|\boldsymbol{\mu}\|, \boldsymbol{v}_j \rangle^2$. (a) **Right**: Cumulative score for $\boldsymbol{v}_{\text{aug}}$, i.e., $\text{score}_i(\boldsymbol{v}_{\text{aug}}) = \sum_{j \le i} \langle \boldsymbol{v}_{\text{aug}}, \boldsymbol{v}_j \rangle^2$. (b) Phase transition showing T^* against τ: theoretical prediction (dashed curve) versus empirical values (circles). 91
4.9 An illustration showing the phase changes in the one-spike inhomogeneous model as we increase τ. In the figure 'E' is a shorthand for expansion and 'S' for shrinkage; see Corollary 4.6.3. 92

A.2 Comparison between conformalized and model-based matrix completion approaches. 129
A.3 Heatmaps for $\mathbf{P}$ and $\widehat{\mathbf{P}}$. 130
A.4 Comparison under heterogenous missingness. 131
A.6 Descriptive plots for the underlying matrix $\mathbf{M}$. 133
A.7 Sales dataset with heterogeneous and misspecified missingness. 133
A.8 Comparison between conformalized and model-based matrix completion with sales dataset. 134

B.1 Results with varying ρ in the well-specified setting. The solid vertical line denotes an estimate $\widehat{\rho}$ of the KL divergence, $D_{\text{KL}}(P_{\text{target}} \| P)$ (See Appendix B.4.2 for details). The solid horizontal line (in the left-hand plot) marks the nominal coverage level, $1 - \alpha = 90\%$. 163
B.2 Results with varying ρ in the misspecified setting. The solid vertical line denotes an estimate $\widehat{\rho}$ of the KL divergence, $D_{\text{KL}}(P_{\text{target}} \| P)$ (See Appendix B.4.2 for details). The solid horizontal line (in the left-hand plot) marks the nominal coverage level, $1 - \alpha = 90\%$. 163
B.3 Histogram of $\widehat{\rho}$. (See Appendix B.4.2 for details.) 164

C.1 Cumulative sums of alignment scores $\kappa_{j,c_1,c_2} = \langle \boldsymbol{v}_j, \boldsymbol{\mu}_{c_1,c_2} \rangle^2 / \|\boldsymbol{\mu}_{c_1,c_2}\|^2$ for different pairs with the pretrained encoder and a one-layer linear projector $\mathbf{W} \in \mathbb{R}^{512 \times 512}$ under the standard SimCLR loss (ResNet-18 and 10-class `STL-10` dataset). 167
C.2 Results with the pretrained encoder and a one-layer linear projector $\mathbf{W} \in \mathbb{R}^{2048 \times 2048}$ under the standard SimCLR loss (ResNet-50 and 10-class `STL-10` dataset). **(a) Left**: singular values of $\mathbf{W}$ with varying temperature. **(a) Middle**: cumulative sums of alignment scores $\kappa_{j,c_1,c_2} = \langle \boldsymbol{v}_j, \boldsymbol{\mu}_{c_1,c_2} \rangle^2 / \|\boldsymbol{\mu}_{c_1,c_2}\|^2$ for the pair $(c_1, c_2) = (7, 2)$. **(a) Right**: comparison between downstream task accuracy with features and embeddings for 10-class evaluation. **(b)** Cumulative sums of alignment scores $\kappa_{j,c_1,c_2} = \langle \boldsymbol{v}_j, \boldsymbol{\mu}_{c_1,c_2} \rangle^2 / \|\boldsymbol{\mu}_{c_1,c_2}\|^2$ for different pairs. 246
C.3 Results for SimCLR models trained from scratch. **Left**: comparison between classification accuracy with features (before projection) and embeddings (after projection). **Middle**: singular values of feature matrix $\mathbf{H}_0 = [\boldsymbol{h}_{1,0}, \ldots, \boldsymbol{h}_{n,0}]$ with varying temperatures. **Right**: singular values of embedding matrix $\mathbf{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n]$ with varying temperatures. 247

C.4 Plots for train-from-scratch experiment with projector $\mathbf{W}^{(r)} = \mathbf{W}\mathbf{V}_r(\mathbf{H}_0) \in \mathbb{R}^{p\times r}$ and $\boldsymbol{\mu}^{(r)} = \mathbf{V}_r^\top(\mathbf{H}_0)\boldsymbol{\mu} \in \mathbb{R}^r$ which are both projected onto feature matrix $\mathbf{H}_0$'s top-r right singular subspace with $r = \text{argmax}k : s_k(\mathbf{H}_0) \geq 10^{-3}\}$. **Left**: unsquared alignment scores $\langle \boldsymbol{v}_j^{(r)}, \boldsymbol{\mu}_{c_1,c_2}^{(r)}\rangle / \|\boldsymbol{\mu}_{c_1,c_2}^{(r)}\|$ for the pair $(c_1, c_2) = (7, 2)$. **Middle**: cumulative sum of alignment scores. **Right**: singular values of $\mathbf{W}^{(r)}$. . 247

C.5 **Left** the relative change of features and embeddings: $\|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}\|/\|\boldsymbol{\theta}^{(t)}\|$ and $\|\mathbf{W}^{(t+1)} - \mathbf{W}^{(t)}\|/\|\mathbf{W}^{(t)}\|$; **Right** the standardized distance between $\mathbf{W}^{(t)}$ and the minimizer $\widetilde{\mathbf{W}}^{(t)}$: $\|\widetilde{\mathbf{W}}^{(t)} - \mathbf{W}^{(t)}\|/\|\mathbf{W}^{(t)}\|$. 248

C.6 Results with the pretrained encoder and a one-layer linear projector $\mathbf{W} \in \mathbb{R}^{2048\times 2048}$ under the standard SimCLR loss (ResNet-50, 10-class `STL`-10 dataset and feature-level perturbation with $\sigma_{\text{aug}} = 1.0$). **Left**: singular values of $\mathbf{W}$ with varying temperature. **Middle**: cumulative sums of alignment scores $\kappa_{j,c_1,c_2} = \langle \boldsymbol{v}_j, \boldsymbol{\mu}_{c_1,c_2}\rangle^2/\|\boldsymbol{\mu}_{c_1,c_2}\|^2$ for the pair $(c_1, c_2) = (7, 2)$. **Right**: comparison between downstream task accuracy with features and embeddings for 10-class evaluation. 248

ACKNOWLEDGMENTS

First and foremost, I would like to dedicate my acknowledgments to my advisors, Rina Foygel Barber and Cong Ma. My gratitude for their support and guidance throughout my PhD journey is beyond words. Rina's wisdom, professionalism, humbleness, and generosity have been a lighthouse for me, which have set and will always be a role model for me in my future career. Each of our meetings and brainstorming discussions keeps impressing me with her sharp and deep intuition for statistical problems. Most importantly, Rina's care for students and her consistently meticulous organization of group meetings over the years provided crucial support to me, especially as a junior PhD student during the pandemic. I was so fortunate to meet Cong at the end of my second year, driven by our shared interest in the topic of transfer learning, and at that time, I would never have expected our journey since then in the field of self-supervised learning. In our projects on self-supervised learning with collaborators, those aha moments, and endless discussions on machine learning theory and research tastes have all been my precious memories. During our discussions, Cong can always explain complicated concepts in an extremely concise way and can always think differently. Thank you, Rina and Cong! You have set very high standards for a researcher and an advisor, which will be guiding me in my future career!

I would like to thank Professor Rebecca Willett for serving on my thesis defense committee and for her inspiring suggestions and insightful questions during both my proposal and defense, which have significantly helped me improve my projects and presentation skills. I have also benefited greatly from various research internships and collaborations. Professor Zemin Zheng at the University of Science and Technology of China (USTC) was my first advisor in statistics and introduced me to multiple testing and the Knockoff methods. Professor Lingzhou Xue at Penn State kindly offered me the opportunity to visit Penn State in my Sophomore year, allowing me a fulfilling summer learning about high-dimensional statistics and exploring a new town. My five-month internship with Professor Jun S Liu's group

at Harvard was one of the most rewarding summers of my academic journey, providing a clear picture of PhD life and exposing me to diverse topics in statistics and machine learning. Working with Professors Zongming Ma and Yiqiao Zhong enriched my understanding of both statistical theory and its role in modern data science. I am also grateful to Bahar Taşkesen, who has been both an excellent teacher in distributionally robust optimization and a supportive friend during challenging times. I always enjoy the collaboration with Zhimei Ren and Ying Jin, with all the brainstorming meetings, and discussions broadly on statistics as well as random chats and jokes. I cannot thank Zhimei enough for her guidance and help since my first PhD project, and her encouragement during my early and ambitious-but-lost years.

Memories in Jones Laboratory, Hyde Park, and downtown Chicago will never fade away. I would like to thank all my friends, group members—Aabesh Bhattacharyya, Rohan Hore, Yoonhaeng Hur, Rong Jiang, Jeonghwan Lee, Yonghoon Lee, Ruiting Liang, Yuetian Luo, Zhimei Ren, Raphael Rossellini, Jake Soloff, Lekun Wang, Qiaosen Wang, Yuepeng Yang, Boxin Zhao, and Wanrong Zhu—and cohorts, Melissa Adrian, Ziming Gan, Lin Gui, Wentao Hu, Wei Kuang, Soumyabrata Kundu, Yating Liu, Zihao Wang, Ziyang Wei, Jinwen Yang, and Zihan Zhou—for the engaging discussions, shared laughter, and the many meals we enjoyed together. I also wish to thank my roommate, Zhouyu Shen, for sharing the challenges of the Covid-19 pandemic, our travels to different cities, and the countless discussions on mathematical problems over four years. I am more than grateful to my former roommate and long-time friend since USTC, Ming Yuan, for the cherished memories from Hefei to Chicago since old days when we were 17, our late-night chats with friends at USTC, and his constant inspiration through his pure and dedicated passion for theoretical physics.

Last but not least, I wish to express my deepest gratitude to my family for their unconditional support. The love, care, and patience of my parents, Mei and Jun, and my wife, Zaishi, have been a constant source of strength throughout my journey. I also dedicate this

thesis to the memory of my late grandfather. He taught me to read and write and, from my earliest childhood through my PhD years, stood by my side with unwavering encouragement. His remarkable and often turbulent life always inspired me during my darkest moments. He was always proud of me, and I hope that completing this PhD would have brought him joy. Their firm belief in me has made every achievement possible, and I am forever grateful.

ABSTRACT

This dissertation investigates statistical learning and inference in weakly specified settings, focusing on three key challenges that deviate from ideal statistical settings in canonical studies: distribution-free statistical inference, inference with distribution shifts, and representation learning with unlabeled data.

As an unsupervised task, matrix completion has a wide range of applications across domains, for which it is of interest to quantify the uncertainty in filled entries. Leveraging the framework of weighted conformal prediction, we develop a method that constructs valid predictive intervals for missing entries in a matrix, ensuring robust inference without requiring assumptions on the matrix structure or noise distribution. Our approach provides theoretical guarantees on coverage while remaining agnostic to the choice of completion algorithm, making it broadly applicable in practical scenarios.

As a generalization of matrix completion, which exhibits a shift from the sampled subpopulation to the unsampled subpopulation of interest, distribution shifts present fundamental challenges in statistical inference, as real-world data often exhibit discrepancies between training and test distributions. Standard methods based on sample reweighting can be sensitive to misspecifications, while traditional distributionally robust learning (DRL) techniques may be overly conservative. To address this, we propose a new framework that integrates shape constraints into DRL, leveraging structural properties of data to refine robustness guarantees. By imposing isotonic constraints on estimated density ratios, our approach mitigates the trade-off between reweighting errors and worst-case risk control, enabling more practical and theoretically sound solutions for learning under distribution shifts.

Finally, motivated by the growing prevalence of unlabeled data in modern machine learning, we investigate the theoretical foundations of contrastive self-supervised learning (SSL). While SSL has demonstrated remarkable empirical success in extracting high-quality representations from high-dimensional data, its theoretical understanding remains incomplete.

We provide a rigorous analysis of the statistical properties of representations learned through contrastive objectives, establishing guarantees on their generalization and utility in downstream tasks. Our results reveal key mechanisms underlying the effectiveness of contrastive learning, offering new insights into its advantages in high-dimensional settings.

Together, the contributions of this dissertation advance the methodological developments and theoretical foundations of statistical learning in settings where traditional assumptions are weakened or absent, offering principled approaches for inference and theoretical guarantees for representation learning in modern data science applications.

CHAPTER 1

INTRODUCTION

The rapid expansion of data availability and the development of large-scale models have significantly advanced modern data science and enriched the context of statistical learning, enabling progress across a wide range of applications. At the same time, these advancements have introduced new challenges, including distribution shifts, data heterogeneity, an abundance of unlabeled multimodal data, and model overfitting. Such phenomena deviate from traditional statistical settings, which often assume independent and identically distributed (i.i.d.) observations drawn from well-specified or testable distributions with strong parametric or structural assumptions. To address these emerging complexities, new theoretical frameworks and methodological approaches have been developed in recent years.

This thesis focuses on three key aspects of statistical learning and inference in weakly specified settings:

1. **Distribution-free statistical inference**. Unlike classical approaches that rely on parametric distributional assumptions, we develop methods that provide finite-sample valid inference under minimal distributional constraints. Chapter 2 particularly focuses on distribution-free inference for missing data problems.

2. **Statistical learning and inference with shifted distributions**. We study settings where the test distribution differs from the training distribution, exploring principled approaches to ensure both robust and powerful inference despite distributional discrepancies.

3. **Representation learning with unlabeled data**. In the absence of labels, self-supervised and contrastive learning has been the state-of-the-art technique to extract meaningful and informative representations from data. We analyze the behaviors of learned representations from contrastive learning from a theoretical perspective.

By addressing these challenges, this thesis contributes to the theoretical foundations of statistical learning and inference in some modern scenarios, offering insights into inference, robustness, and representation learning beyond the constraints of classical statistical assumptions.

1.1 Distribution-free statistical inference

In canonical regression problems, the underlying distribution of (X, Y), where Y is the response of interest and X is a covariate, is usually assumed to follow a parametric model. Concretely, we usually assume

$$Y = f_{\boldsymbol{\theta}}(X) + \varepsilon,$$

where ε is typically assumed to be sub-Gaussian and the form of $f_{\boldsymbol{\theta}}$ is given, e.g. the linear model with $f_{\boldsymbol{\theta}}(X) = \boldsymbol{\theta}^\top X$. Traditional estimation techniques, such as maximum likelihood, yield a consistent estimate $\widehat{\boldsymbol{\theta}}$ from data under well-specified conditions. Given the distribution of $f_{\widehat{\boldsymbol{\theta}}}(X) + \varepsilon$, we can construct a distribution-based confidence interval $\widehat{C}_{1-\alpha}(X)$ for Y such that

$$\mathbb{P}\left(Y \in \widehat{C}_{1-\alpha}(X)\right) \approx 1 - \alpha.$$

However, the (approximate) validity of the aforementioned confidence intervals heavily relies on the correctly specified distribution of $Y \mid X$, but as we may not have access to the ground truth in practice, the validity of this distribution-based confidence interval will not be guaranteed.

Hence, with observations (X_i, Y_i), $i \in [n]$, and (X_{n+1}, Y_{n+1}) where Y_{n+1} is not observed, it is of crucial interest to determine the minimal distributional assumptions on this finite sample, with which we are able to construct a valid confidence interval for Y_{n+1}. More specifically, in Chapter 2, conformal prediction (Vovk et al., 2005; Angelopoulos et al., 2024)

will be introduced in detail.

Moreover, as the inferential target in regression tasks is clear, it is also meaningful to extend the inferential procedure to unsupervised settings, e.g., matrix completion, clustering, and matrix factorization. For example, in matrix completion where a (deterministic) matrix $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$ is only partially observed, the task is to fill in the missing entries. Although sophisticated algorithms for matrix completion have been studied in the field for over two decades and assumption-based optimality has also been established, the valid uncertainty quantification for matrix completion algorithms has long been an open question until Chen et al. (2019). In addition, as existing inferential procedures heavily rely on distributional assumptions on the underlying matrix $\mathbf{M}$, it is of crucial importance to question the existence of a distribution-free inferential procedure for a deterministic matrix. Chapter 2 provides a positive answer.

1.2 Statistical learning and inference with distribution shift

Traditional statistical analysis is mostly built upon the i.i.d. assumption of data. For example, in a regression problem with

$$(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P,\ i \in [n], \qquad (X_{n+1}, Y_{n+1}) \sim P_{\text{test}},$$

where Y_{n+1} is unobserved, it is typically assumed that $P = P_{\text{test}}$. However, data can be highly heterogeneous and deviates from the ideal i.i.d. setting in various ways. A well-recognized challenge in this context is *distribution shift*, which occurs when the target distribution of interest differs from the distribution from which the observed data is sampled, i.e. $P_{\text{test}} \neq P$.

In addition, sampling bias is quite common in practice, which is implicitly related to distribution shift problems. For example, data can be imperfect due to missingness or trun-

cation. Although it is intuitive to utilize available data for statistical learning or inference, the observed data may not be representative for the underlying distribution, or not exchangeable with the unsampled data, due to sampling bias. Under certain sampling mechanisms (e.g. missing at random), the sampling bias can be modeled as distribution shifts between sampled and unsampled subpopulations. Consequently, these issues can be addressed using reweighting approaches, which are applicable to scenarios such as matrix completion (Gui et al., 2024a) and survival analysis (Gui et al., 2023b).

However, as a mainstream technique in addressing distribution shifts, sample reweighting, which relies on the density ratio between target and data distributions, can be sensitive to the misspecification of weights, especially when available data is scarce. Alternatively, distributionally robust learning (DRL) aims to control out-of-distribution performance uniformly across a set of candidate target distributions, which may result in excessive conservativeness. Chapter 3 will introduce an approach with the goal of balancing the misspecification of sample reweighting and the over-pessimism of DRL by incorporating the shape constraints (e.g., the isotonic constraints with respect to an estimated density ratio) as side information. More broadly, this work aims to find more practical solutions to distribution shifts under certain structures, which is motivated by the blessing of structural assumptions (e.g., sparsity) in high-dimensional statistics.

1.3 Leveraging unlabeled data to enhance statistical tasks

The increasing availability of unsupervised data is an emergent phenomenon in modern data science. As the sample size plays a crucial role in the performance of statistical models, it motivates efficient practices in leveraging the unlabeled data for enhancing supervised tasks. Although this topic is proposed and studied in the area of unsupervised and semi-supervised learning, the modern self-supervised learning (SSL) technique has started a new chapter and has been widely adopted in training large vision-language models.

One of the key components is the infoNCE contrastive loss. Taking image classification as an example, with augmented copies of each image (by color distortion or rotation), the contrastive loss learns a map f that simultaneously repels copies of distinct images and maximizes the similarity between copies of the same image. Accordingly, the learned representation $f(X)$ can be used as the input for downstream (supervised) tasks. Chapter 4 will investigate the advantage of SSL and validate that contrastive learning benefitting the downstream accuracy is specifically a high-dimensional phenomenon. More concretely, in the overparametrized regime, for classification tasks with data from a mixture model, even if the representation map f learned via the contrastive loss lies in the linear class, our theoretical analysis shows that the precise asymptotic accuracy can be improved with $f(X)$ as the input for classification.

1.4 Organization of thesis

In light of the three key challenges mentioned in previous paragraphs, this thesis contributes one chapter to each topic: Chapter 2 will focus on the distribution-free uncertainty quantification for matrix completion; Chapter 3 considers a more general setting with distribution shifts and proposes a constrained distributionally robust procedure for valid inference on the target distribution; Chapter 4 contributes to the theoretical foundations of contrastive self-supervised learning with unlabeled data. Technical proofs and additional experimental details are deferred to Chapters A, B, C for three chapters, respectively.

CHAPTER 2

DISTRIBUTION-FREE UNCERTAINTY QUANTIFICATION FOR MATRIX COMPLETION

Matrix completion aims to estimate missing entries in a data matrix, using the assumption of a low-complexity structure (e.g., low rank) so that imputation is possible. While many effective estimation algorithms exist in the literature, uncertainty quantification for this problem has proved to be challenging, and existing methods are extremely sensitive to model misspecification. In this work, we propose a distribution-free method for predictive inference in the matrix completion problem. Our method adapts the framework of conformal prediction, which provides confidence intervals with guaranteed distribution-free validity in the setting of regression, to the problem of matrix completion. Our resulting method, conformalized matrix completion (cmc), offers provable predictive coverage regardless of the accuracy of the low-rank model. Empirical results on simulated and real data demonstrate that cmc is robust to model misspecification while matching the performance of existing model-based methods when the model is correct.[1]

2.1 Introduction

Matrix completion, the task of filling in missing entries in a large data matrix, has a wide range of applications such as gene-disease association analysis in bioinformatics (Natarajan and Dhillon, 2014), collaborative filtering in recommender systems (Rennie and Srebro, 2005), and panel data prediction and inference in econometrics (Amjad et al., 2018; Bai and Ng, 2021; Athey et al., 2021). If the underlying signal is assumed to be low-rank, a range of estimation algorithms have been proposed in the literature, including approaches based on convex relaxations of rank (Candes and Plan, 2010; Candès and Tao, 2010; Koltchinskii

1. This chapter is based on Gui et al. (2023a).

et al., 2011; Foygel and Srebro, 2011; Negahban and Wainwright, 2012; Yang and Ma, 2022) and nonconvex optimization over the space of low-rank matrices (Keshavan et al., 2010; Jain et al., 2013; Sun and Luo, 2016; Ma et al., 2020).

While the problem of estimation has been explored through many different approaches in the literature, the question of uncertainty quantity for matrix completion remains challenging and under-explored. Cai et al. (2016) and Carpentier et al. (2018) construct confidence intervals for the missing entries using order-wise concentration inequalities, which could be extremely loose. Chen et al. (2019) proposes a de-biased estimator for matrix completion and characterizes its asymptotic distribution, which then yields entrywise confidence intervals are then derived for the underlying incoherent matrix with i.i.d. Gaussian noise. See also Xia and Yuan (2021); Farias et al. (2022) for related approaches.

The effectiveness of the aforementioned uncertainty quantification methods rely heavily on the exact low-rank structure as well as assumptions on the tail of the noise. In practice, the low-rank assumption can only be met approximately, and heavy-tailed noise is quite common in areas such as macroeconomics and genomics (Nair et al., 2022). Without exact model specifications, the solution to the matrix completion algorithm is hard to interpret and the asymptotic distribution of the obtained estimator is no longer valid. The dependence of model-based approaches on exact model assumptions motivates us to answer the question:

can we construct valid and short confidence intervals for the missing entries that are free of both model assumptions and of constraints on which estimation algorithms we may use?

2.1.1 Conformal prediction for distribution-free uncertainty quantification

Our proposed method is built on the idea of conformal prediction. The conformal prediction framework, developed by Vovk et al. (2005) and Shafer and Vovk (2008), has drawn significant attention in recent years in that it enables the construction of distribution-free confidence intervals that are valid with exchangeable data from any underlying distribution

and with any "black-box" algorithm. To reduce computation in full conformal prediction, variants exist based on data splitting (Vovk et al., 2005; Lei et al., 2018), leave-one-out (Barber et al., 2021), cross-validation (Vovk, 2015; Barber et al., 2021), etc.

When distribution shift is present, with covariate shift as a special case, the exchangeability is violated, and Tibshirani et al. (2019) proposes a more general procedure called weighted conformal prediction that guarantees the validity under the weighted exchangeability of data. Barber et al. (2022) further relaxes the exchangeability assumption on the data and characterizes the robustness of a generalized weighted approach via an upper bound for the coverage gap. Other related works (Lei and Candès, 2020; Candès et al., 2021; Jin et al., 2023) adapt the conformal framework to applications in causal inference, survival analysis, etc., and study the robustness of weighted conformal prediction with estimated weights.

2.1.2 Main contributions

In this paper, we adapt the framework of conformal prediction for the problem of matrix completion, and make the following contributions:

1. We construct distribution-free confidence intervals for the missing entries in matrix completion via conformal prediction. The validity is free of any assumption on the underlying matrix and holds regardless of the choice of estimation algorithms practitioners use. To achieve this, we prove the (weighted) exchangeability of unobserved and observed units when they are sampled (possibly nonuniformly) without replacement from a finite population.

2. When the sampling mechanism is unknown, we develop a provable lower bound for the coverage rate which degrades gracefully as the estimation error of the sampling probability increases.

3. In addition, to improve computational efficiency when faced with a large number of missing entries, we propose a one-shot conformalized matrix completion approach that only computes the weighted quantile once for all missing entries.

2.2 Problem setup

In this section, we formulate the matrix completion problem and contrast our distribution-free setting with the model-based settings that are more common in the existing literature. For a partially-observed $d_1 \times d_2$ matrix, the subset $\mathcal{S} \subseteq [d_1] \times [d_2]$ denotes the set of indices where data is observed—that is, our observations consist of $(M_{ij})_{(i,j)\in\mathcal{S}}$. In much of the matrix completion literature, it is common to assume a signal-plus-noise model for the observations,

$$M_{ij} = M^*_{ij} + E_{ij},$$

where $\mathbf{M}^* = (M^*_{ij})_{(i,j)\in[d_1]\times[d_2]}$ is the "true" underlying signal while $(E_{ij})_{(i,j)\in\mathcal{S}}$ denotes the (typically zero-mean) noise. Frequently, it is assumed that $\mathbf{M}^*$ is low-rank, or follows some other latent low-dimensional structure, so that recovering $\mathbf{M}^*$ from the observed data is an identifiable problem. In works of this type, the goal is to construct an estimator $\widehat{\mathbf{M}}$ that accurately recovers the underlying $\mathbf{M}^*$. In contrast, in a more model-free setting, we may no longer wish to hypothesize a signal-plus-noise model, and can instead assume that we are observing a random subset of entries of a deterministic matrix $\mathbf{M}$; in this setting, estimating $\mathbf{M}$ itself becomes the goal.

For both frameworks, many existing results focus on the problem of estimation (either of $\mathbf{M}^*$ or of $\mathbf{M}$), with results establishing bounds on estimation errors under various assumptions. For instance, see Candes and Plan (2010); Candès and Tao (2010); Negahban and Wainwright (2012); Chen et al. (2020c) for results on estimating $\mathbf{M}^*$ (under stronger conditions, e.g., a low-rank and incoherent signal, and zero-mean sub-Gaussian noise), or Srebro

and Shraibman (2005); Foygel and Srebro (2011) for results on estimating $\mathbf{M}$ (under milder conditions); note that in the latter case, it is common to instead refer to predicting the entries of $\mathbf{M}$, e.g., temperature or stock market returns, since we often think of these as noisy observations of some underlying signal. On the other hand, relatively little is known about the problem of uncertainty quantification for these estimates: given some estimator $\widehat{\mathbf{M}}$, can we produce a confidence interval around each $\widehat{M}_{ij}$ that is (asymptotically) guaranteed to contain the target with some minimum probability?

The results of Chen et al. (2019) address this question under strong assumptions, namely, a low-rank and incoherent signal $\mathbf{M}^*$, plus i.i.d. Gaussian noise $(E_{ij})_{(i,j)\in[d_1]\times[d_2]}$. Zhao and Udell (2020) proposes another inferential procedure under the Gaussian copula assumption for the data. In this work, we provide a complementary answer that instead addresses the model-free setting: without relying on assumptions, we aim to produce a provably valid confidence interval for the entries M_{ij} of $\mathbf{M}$ (not of $\mathbf{M}^*$, because without assuming a low-rank model, there is no meaningful notion of an "underlying signal").

To do so, we will treat the matrix $\mathbf{M}$ as deterministic, while the randomness arises purely from the random subset of entries that are observed. More specifically, we assume

$$Z_{ij} = \mathbb{1}\left\{(i,j) \text{ is observed}\right\} \sim \text{Bern}(p_{ij}), \quad \text{independently for all } (i,j) \in [d_1] \times [d_2], \tag{2.1}$$

and let $\mathcal{S} = \{(i,j) \in [d_1] \times [d_2] : Z_{ij} = 1\}$ be the subset of observed locations. We consider the setting where the sample probabilities p_{ij}'s are nonzero. After observing $\mathbf{M}_{\mathcal{S}} = (M_{ij})_{(i,j)\in\mathcal{S}}$, our goal is to provide confidence intervals $\{\widehat{C}(i,j)\}_{(i,j)\in\mathcal{S}^c}$ for all unobserved entries, with $1-\alpha$ coverage over the unobserved portion of the matrix—that is, we would like our confidence intervals to satisfy

$$\mathbb{E}\left[AvgCov(\widehat{C}; \mathbf{M}, \mathcal{S})\right] \geq 1-\alpha, \tag{2.2}$$

where

$$AvgCov(\widehat{C}; \mathbf{M}, \mathcal{S}) = \frac{1}{|\mathcal{S}^c|} \sum\nolimits_{(i,j)\in\mathcal{S}^c} \mathbb{1}\Big\{ M_{ij} \in \widehat{C}(i,j) \Big\} \tag{2.3}$$

measures the "average coverage rate" of the confidence intervals.[2] The goal of this work is to provide an algorithm for constructing confidence intervals $\{\widehat{C}(i,j)\}_{(i,j)\in\mathcal{S}^c}$ based on the observed entries $\mathbf{M}_{\mathcal{S}}$, satisfying (2.2) with no assumptions on $\mathbf{M}$. In particular, given the strong success of various matrix completion algorithms for the problem of estimation, we would like to be able to provide uncertainty quantification around any base estimator—i.e., given any algorithm for producing an estimate $\widehat{\mathbf{M}}$, our method constructs confidence intervals $\{\widehat{C}(i,j)\}_{(i,j)\in\mathcal{S}^c}$ around this initial estimate.

Notation. For a matrix $\mathbf{A} = (A_{ij}) \in \mathbb{R}^{m_1 \times m_2}$, we define $\|\mathbf{A}\|_\infty \coloneqq \max_{(i,j)\in[m_1]\times[m_2]} |A_{ij}|$, and $\|\mathbf{A}\|_1 \coloneqq \sum_{(i,j)\in[m_1]\times[m_2]} |A_{ij}|$. We use $\mathbf{A}_{i,\cdot}$ and $\mathbf{A}_{\cdot,j}$ to refer to the ith row and jth column of $\mathbf{A}$, respectively. For any $t \in \mathbb{R}$, δ_t denotes the distribution given by a point mass at t.

2.3 Conformalized matrix completion

We next present our method, conformalized matrix completion (cmc). Our procedure adapts the split conformal prediction method (Vovk et al., 2005) to the problem at hand. As is standard in the conformal prediction framework, the goal of cmc is to provide uncertainty quantification (i.e., confidence intervals) around the output of any existing estimation procedure that the analyst chooses to use—this is a core strength of the conformal methodology, as it allows the analyst to use state-of-the-art estimation procedures without compromising on the validity of inference. At a high level, the cmc procedure will output a confidence

2. To handle the degenerate case where $\mathcal{S} = [d_1] \times [d_2]$ (i.e., we happen to have observed the entire matrix) and thus $\mathcal{S}^c = \varnothing$, we simply define $AvgCov(\widehat{C}; \mathbf{M}, \mathcal{S}) \equiv 1$.

interval of the form

$$\widehat{C}(i,j) = \widehat{M}_{ij} \pm \widehat{q} \cdot \widehat{s}_{ij}$$

for the matrix value M_{ij} at each unobserved entry (i,j). Here, after splitting the observed entries into a training set and a calibration set, the training set is used to produce $\widehat{M}_{ij}$ (a point estimate for the target value M_{ij}) and $\widehat{s}_{ij}$ (a scaling parameter that estimates our uncertainty for this point estimate); then, the calibration set is used to tune the scalar $\widehat{q}$, to adjust the width of these confidence intervals and ensure coverage at the desired level $1-\alpha$.[3]

2.3.1 Exchangeability and weighted exchangeability

We split the set of observed indices $\mathcal{S}$ into a training and a calibration set $\mathcal{S} = \mathcal{S}_{\text{train}} \cup \mathcal{S}_{\text{cal}}$ as is shown in Algorithm 1. We should notice that the sampling without replacement may introduce implicit dependence as well as a distribution shift from the i.i.d. sampling. Thus the two sets are dependent, which is different from the split conformal methods in regression problems. Before we present the coverage guarantees for our method, we first examine the role of (weighted) exchangeability in this method, to build intuition for how the method is constructed.

Intuition: the uniform sampling case

First, for intuition, consider the simple case where the entries are sampled with constant probability, $p_{ij} \equiv p$. If this is the case, then the set of calibration samples and the test point are exchangeable—that is, for (i_*, j_*) denoting the test point that is drawn uniformly from $\mathcal{S}^c$, if we are told that the combined set $\mathcal{S}_{\text{cal}} \cup \{(i_*, j_*)\}$ is equal to $\{(i_1, j_1), \ldots, (i_{n_{\text{cal}}+1}, j_{n_{\text{cal}}+1})\}$ in no particular order (where $n_{\text{cal}} = |\mathcal{S}_{\text{cal}}|$), then the test point location (i_*, j_*) is equally

3. More generally, cmc can be implemented with any choice of a nonconformity score that determines the shape of the resulting confidence intervals—and indeed, this generalization allows for categorical rather than real-valued matrix entries M_{ij}. We will address this extension in the Supplementary Material, as well as a version of cmc based on full conformal rather than split conformal, in order to avoid data splitting.

Algorithm 1 Conformalized matrix completion (cmc)

1: **Input**: target coverage level $1-\alpha$; data splitting proportion $q \in (0,1)$; observed entries $\mathbf{M}_{\mathcal{S}}$.

2: Split the data: draw $W_{ij} \overset{\text{i.i.d.}}{\sim} \text{Bern}(q)$, and define training and calibration sets

$$\mathcal{S}_{\text{train}} = \{(i,j) \in \mathcal{S} : W_{ij} = 1\}, \quad \text{and} \quad \mathcal{S}_{\text{cal}} = \{(i,j) \in \mathcal{S} : W_{ij} = 0\}.$$

3: Using the training data $\mathbf{M}_{\mathcal{S}_{\text{train}}}$ indexed by $\mathcal{S}_{\text{train}} \subseteq [d_1] \times [d_2]$, compute:

- An initial estimate $\widehat{\mathbf{M}}$ using any matrix completion algorithm (with $\widehat{M}_{ij}$ estimating the target M_{ij});
- Optionally, a local uncertainty estimate $\widehat{\mathbf{s}}$ (with $\widehat{s}_{ij}$ estimating our relative uncertainty in the estimate $\widehat{M}_{ij}$), or otherwise set $\widehat{s}_{ij} \equiv 1$;
- An estimate $\widehat{\mathbf{P}}$ of the observation probabilities (with $\widehat{p}_{ij}$ estimating p_{ij}, the probability of entry (i,j) being observed).

4: Compute normalized residuals on the calibration set, $R_{ij} = \frac{|M_{ij} - \widehat{M}_{ij}|}{\widehat{s}_{ij}},\ (i,j) \in \mathcal{S}_{\text{cal}}$.

5: Compute estimated odds ratios for the calibration set and test set, $\widehat{h}_{ij} = \frac{1-\widehat{p}_{ij}}{\widehat{p}_{ij}},\ (i,j) \in \mathcal{S}_{\text{cal}} \cup \mathcal{S}^c$, and then compute weights for the calibration set and test point,

$$\widehat{w}_{ij} = \frac{\widehat{h}_{ij}}{\sum_{(i',j') \in \mathcal{S}_{\text{cal}}} \widehat{h}_{i'j'} + \max_{(i',j') \in \mathcal{S}^c} \widehat{h}_{i'j'}},\ (i,j) \in \mathcal{S}_{\text{cal}}, \quad \widehat{w}_{\text{test}} = \frac{\max_{(i,j) \in \mathcal{S}^c} \widehat{h}_{ij}}{\sum_{(i',j') \in \mathcal{S}_{\text{cal}}} \widehat{h}_{i'j'} + \max_{(i',j') \in \mathcal{S}^c} \widehat{h}_{i'j'}}. \tag{2.4}$$

6: Compute threshold $\widehat{q} = \text{Quantile}_{1-\alpha}\left(\sum_{(i,j) \in \mathcal{S}_{\text{cal}}} \widehat{w}_{ij} \cdot \delta_{R_{ij}} + \widehat{w}_{\text{test}} \cdot \delta_{+\infty}\right)$.

7: **Output**: confidence intervals $\widehat{C}(i,j) = \widehat{M}_{ij} \pm \widehat{q} \cdot \widehat{s}_{ij}$ for each unobserved entry $(i,j) \in \mathcal{S}^c$.

likely to be at any one of these $n_{\text{cal}} + 1$ locations. Consequently, by exchangeability, the calibration residuals $\{R_{ij} : (i,j) \in \mathcal{S}_{\text{cal}}\}$, defined in Algorithm 1 above, are exchangeable with the test residual $R_{i_*j_*} = |M_{i_*j_*} - \widehat{M}_{i_*j_*}|/\widehat{s}_{i_*j_*}$; as a result, we can construct our confidence interval $\widehat{C}(i_*, j_*)$ with $\widehat{q}$ determined by the $(1-\alpha)$-quantile of the calibration residuals $\{R_{ij} : (i,j) \in \mathcal{S}_{\text{cal}}\}$.

Indeed, this is exactly what cmc does in this case: if we are aware that sampling is uniform, it would naturally follow that we estimate $\widehat{p}_{ij} \equiv \widehat{p}_0$ by some (potentially data-dependent) scalar $\widehat{p}_0$; consequently, all the weights are given by $\widehat{w}_{ij} \equiv \frac{1}{n_{\text{cal}}+1}$ and $\widehat{w}_{\text{test}} =$

$\frac{1}{n_{\text{cal}}+1}$, and thus $\widehat{q}$ is the (unweighted) $(1-\alpha)$-quantile of the calibration residuals (with a small correction term, i.e., the term $+\widehat{w}_{\text{test}} \cdot \delta_{+\infty}$ appearing in the definition of $\widehat{q}$, to ensure the correct probability of coverage at finite sample sizes).

Weighted exchangeability for the matrix completion problem

Next, we move to the general case, where now the sampling may be highly nonuniform. First suppose the sampling probabilities p_{ij} are known, and suppose again that we are told that the combined set $\mathcal{S}_{\text{cal}} \cup \{(i_*, j_*)\}$ is equal to $\{(i_1, j_1), \ldots, (i_{n_{\text{cal}}+1}, j_{n_{\text{cal}}+1})\}$ in no particular order. In this case, the test point (i_*, j_*) is not equally likely to be at any one of these $n_{\text{cal}} + 1$ locations. Instead, we have the following result:

Lemma 2.3.1. *Following the notation defined above, if* $(i_*, j_*) \mid \mathcal{S} \sim \text{Unif}(\mathcal{S}^c)$*, it holds that*

$$\mathbb{P}\big\{(i_*, j_*) = (i_k, j_k) \mid \mathcal{S}_{\text{cal}} \cup \{(i_*, j_*)\} = \{(i_1, j_1), \ldots, (i_{n_{\text{cal}}+1}, j_{n_{\text{cal}}+1})\}, \mathcal{S}_{\text{train}}\big\} = w_{i_k j_k},$$

where we define the weights $w_{i_k j_k} = \dfrac{h_{i_k j_k}}{\sum_{k'=1}^{n_{\text{cal}}+1} h_{i_{k'} j_{k'}}}$ *for odds ratios given by* $h_{ij} = \frac{1-p_{ij}}{p_{ij}}$.

Consequently, while the test point's residual $R_{i_* j_*} = |M_{i_* j_*} - \widehat{M}_{i_* j_*}| \,/\, \widehat{s}_{i_* j_*}$ is not exchangeable with the calibration set residuals $\{R_{ij} : (i,j) \in \mathcal{S}_{\text{cal}}\}$, the framework of weighted exchangeability (Tibshirani et al., 2019) tells us that we can view $R_{i_* j_*}$ as a draw from the weighted distribution that places weight $w_{i_k j_k}$ on each residual $R_{i_k j_k}$—and in particular,

$$\mathbb{P}\big\{M_{i_* j_*} \in \widehat{M}_{i_* j_*} \pm q^*_{i_*, j_*} \cdot \widehat{s}_{i_* j_*}\big\} = \mathbb{P}\left\{R_{i_* j_*} \le q^*_{i_*, j_*}\right\} \ge 1 - \alpha,$$

where $q^*_{i_*, j_*} = \text{Quantile}_{1-\alpha}\big(\sum_{k=1}^{n_{\text{cal}}+1} w_{i_k j_k} \cdot \delta_{R_{i_k j_k}}\big)$. Noting that this threshold $q^*_{i_*, j_*}$ may depend on the test location (i_*, j_*), to accelerate the computation of prediction sets for all missing entries, we instead propose a one-shot weighted conformal approach by upper

bounding the threshold uniformly over all test points:

$$q^*_{i_*,j_*} \le q^* := \text{Quantile}_{1-\alpha}\left(\sum_{(i,j)\in\mathcal{S}_{\text{cal}}} w^*_{ij}\cdot\delta_{R_{ij}} + w^*_{\text{test}}\cdot\delta_{+\infty}\right),$$

for $(i,j)\in\mathcal{S}_{\text{cal}}$,

$$w^*_{ij} = \frac{h_{ij}}{\sum_{(i',j')\in\mathcal{S}_{\text{cal}}} h_{i'j'} + \max_{(i',j')\in\mathcal{S}^c} h_{i'j'}}, \quad w^*_{\text{test}} = \frac{\max_{(i,j)\in\mathcal{S}^c} h_{ij}}{\sum_{(i',j')\in\mathcal{S}_{\text{cal}}} h_{i'j'} + \max_{(i',j')\in\mathcal{S}^c} h_{i'j'}}. \tag{2.5}$$

Now the threshold q^* no longer depends on the location (i_*, j_*) of the test point[4]. If the true probabilities p_{ij} were known, then, we could define "oracle" conformal confidence intervals

$$\widehat{C}^*(i,j) = \widehat{M}_{ij} \pm q^*\cdot\widehat{s}_{ij}, \quad (i,j)\in\mathcal{S}^c.$$

As we will see in the proofs below, weighted exchangeability ensures that, if we do have oracle knowledge of the p_{ij}'s, then these confidence intervals satisfy the goal (2.2) of $1-\alpha$ average coverage. Since the p_{ij}'s are not known in general, our algorithm simply replaces the oracle weights $\boldsymbol{w}^*$ in (2.5) with their estimates $\widehat{\boldsymbol{w}}$ defined in (2.4), using $\widehat{p}_{ij}$ as a plug-in estimate of the unknown p_{ij}.

2.3.2 Theoretical guarantee

In the setting where the p_{ij}'s are not known but instead are estimated, our cmc procedure simply repeats the procedure described above but with weights $\widehat{\boldsymbol{w}}$ calculated using $\widehat{p}_{ij}$'s in place of p_{ij}'s. Our theoretical results therefore need to account for the errors in our estimates $\widehat{p}_{ij}$, to quantify the coverage gap of conformal prediction with the presence of

4. Comparison between the one-shot weighted approach and the exact weighted approach (Algorithm 3) is conducted in Section A.2.1

these estimation errors. To do this, we define an estimation gap term

$$\Delta = \frac{1}{2} \sum_{(i,j) \in \mathcal{S}_{\text{cal}} \cup \{(i_*, j_*)\}} \left| \frac{\widehat{h}_{ij}}{\sum_{(i',j') \in \mathcal{S}_{\text{cal}} \cup \{(i_*, j_*)\}} \widehat{h}_{i'j'}} - \frac{h_{ij}}{\sum_{(i',j') \in \mathcal{S}_{\text{cal}} \cup \{(i_*, j_*)\}} h_{i'j'}} \right|, \quad (2.6)$$

where (i_*, j_*) denotes the test point. Effectively, Δ is quantifying the difference between the "oracle" weights, w^*_{ij}, defined in (2.5), and the estimated weights, $\widehat{w}_{ij}$, defined in (2.4), except that Δ is defined relative to a specific test point, while the weights $\widehat{\boldsymbol{w}}$ (and $\boldsymbol{w}^*$, for the oracle version) provide a "one-shot" procedure that is universal across all possible test points, and is thus slightly more conservative.

Theorem 2.3.2. *Let $\widehat{\mathbf{M}}$, $\widehat{\mathbf{s}}$, and $\widehat{\mathbf{P}}$ be estimates constructed using any algorithms, which depend on the data only through the training samples $\mathbf{M}_{\mathcal{S}_{\text{train}}}$ at locations $\mathcal{S}_{\text{train}}$. Then, under the notation and definitions above, conformalized matrix completion (cmc) satisfies*

$$\mathbb{E}\big[AvgCov(\widehat{C}; \mathbf{M}, \mathcal{S})\big] \geq 1 - \alpha - \mathbb{E}[\Delta].$$

In the homogeneous case, where we are aware that sampling is uniform (i.e., $p_{ij} \equiv p$ for some potentially unknown $p \in (0, 1)$), our error in estimating the weights is given by $\Delta \equiv 0$ (since h_{ij} is constant across all (i, j), and same for $\widehat{h}_{ij}$—and therefore, the true and estimated weights are all equal to $\frac{1}{n_{\text{cal}}+1}$). In this setting, therefore, we would achieve the exact coverage goal (2.2), with $\mathbb{E}\big[AvgCov(\widehat{C}; \mathbf{M}, \mathcal{S})\big]$ guaranteed to be at least $1 - \alpha$.

2.3.3 Examples for missingness

To characterize the coverage gap explicitly, we present the following concrete examples for $\mathbf{P}$ and show how the coverage gap Δ can be controlled. Technical details are shown in the Supplementary Material.

Logistic missingness

Suppose that the missingness follows a logistic model, with $\log(p_{ij}/(1-p_{ij})) = -\log\big(h_{ij}\big) = u_i + v_j$ for some $\boldsymbol{u} \in \mathbb{R}^{d_1}$ and $\boldsymbol{v} \in \mathbb{R}^{d_2}$, where we assume $\boldsymbol{u}^\top \mathbf{1} = 0$ for identifiability. This model is closely related to logistic regression with a diverging number of covariates (Portnoy, 1988; He and Shao, 2000; Wang, 2011). Following Chen et al. (2023), we estimate $\boldsymbol{u}$, $\boldsymbol{v}$ via maximum likelihood,

$$\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}} = \arg\max_{\boldsymbol{u},\boldsymbol{v}} \ \mathcal{L}(\boldsymbol{u}, \boldsymbol{v}) \qquad \text{subject to} \quad \mathbf{1}^\top \boldsymbol{u} = 0.$$

Here the log-likelihood is defined by

$$\mathcal{L}(\boldsymbol{u}, \boldsymbol{v}) = \sum_{(i,j)\in[d_1]\times[d_2]} \{-\log\big(1+\exp(-u_i - v_j)\big) + \mathbf{1}_{(i,j)\in\mathcal{S}_{\text{train}}^c} \log\big(1 - q + \exp(-u_i - v_j)\big)\}.$$

Consequently, we define the estimate of p_{ij} as $\widehat{p}_{ij} = (1+\exp(-\widehat{u}_i - \widehat{v}_j))^{-1}$, for which one can show that $\mathbb{E}[\Delta] \lesssim \sqrt{\frac{\log(\max\{d_1,d_2\})}{\min\{d_1,d_2\}}}$. The proof of this upper bound is shown in Section A.1.3.

Missingness with a general link function

More broadly, we can consider a general link function, where we assume $\log(p_{ij}/(1-p_{ij})) = -\log\big(h_{ij}\big) = \phi(A_{ij})$, where ϕ is a link function and $\mathbf{A} \in \mathbb{R}^{d_1\times d_2}$ is the model parameter. The logistic model we introduced above corresponds to the case when ϕ is the identity function and $\mathbf{A} = \boldsymbol{u}\mathbf{1}^\top + \mathbf{1}\boldsymbol{v}^\top$ is a rank-2 matrix. In this general setup, if $\text{rank}(\mathbf{A}) = k^*$ and $\|\mathbf{A}\|_\infty \le \tau$, we can apply one-bit matrix completion (Davenport et al., 2014) to estimate $\{p_{ij}\}$. More precisely, we define the log-likelihood

$$\mathcal{L}_{\mathcal{S}_{\text{train}}}(\mathbf{B}) = \sum_{(i,j)\in\mathcal{S}_{\text{train}}} \log(\psi(B_{ij})) + \sum_{(i,j)\in\mathcal{S}_{\text{train}}^c} \log(1-\psi(B_{ij})),$$

where $\psi(t) = q(1 + e^{-\phi(t)})^{-1}$ (here rescaling by the constant q is necessary to account for subsampling the training set $\mathcal{S}_{\text{train}}$ from the observed data indices $\mathcal{S}$), and solve the following optimization problem

$$\widehat{\mathbf{A}} = \arg\max_{\mathbf{B}} \ \mathcal{L}_{\mathcal{S}_{\text{train}}}(\mathbf{B}) \qquad \text{subject to} \quad \|\mathbf{B}\|_* \leq \tau\sqrt{k^* d_1 d_2}, \ \ \|\mathbf{B}\|_\infty \leq \tau,$$

where $\|\mathbf{B}\|_*$ is the nuclear norm of $\mathbf{B}$. Consequently, we let $\widehat{h}_{ij} = \exp(-\phi(\widehat{A}_{ij}))$, which leads to $\mathbb{E}[\Delta] \lesssim \min\{d_1, d_2\}^{-1/4}$. The proof of this upper bound is shown in Section A.1.3.

2.4 Simulation studies

In this section, we conduct numerical experiments to verify the coverage guarantee of conformalized matrix completion (cmc) using both synthetic and real datasets. As the validity of cmc is independent of the choice of base algorithm, we choose alternating least squares (als) (Jain et al., 2013) as the base algorithm (results using a convex relaxation (Fazel et al., 2004; Candes and Recht, 2012; Chen et al., 2020c) are shown in the Supplementary Material). We use $cmc - als$ to refer to this combination. We use $AvgCov(\widehat{C})$ (the average coverage, defined in (2.3)), and average confidence interval length, $AvgLength(\widehat{C}) = \frac{1}{|\mathcal{S}^c|}\sum_{(i,j)\in\mathcal{S}^c} \text{length}\big(\widehat{C}(i,j)\big)$, to evaluate the performance of different methods. All the results in this section can be replicated with the code available at `https://github.com/yugjerry/conf-mc`.

2.4.1 Synthetic dataset

Throughout the experiments, we set the true rank $r^* = 8$, the desired coverage rate $1 - \alpha = 0.90$, and report the average results over 100 random trials.

Data generation. In the synthetic setting, we generate the data matrix by $\mathbf{M} = \mathbf{M}^* + \mathbf{E}$, where $\mathbf{M}^*$ is a rank r^* matrix, and $\mathbf{E}$ is a noise matrix (with distribution specified below).

The low-rank component $\mathbf{M}^*$ is generated by $\mathbf{M}^* = \kappa \mathbf{U}^* \mathbf{V}^{*\top}$, where $\mathbf{U}^* \in \mathbb{R}^{d_1 \times r^*}$ is the orthonormal basis of a random $d_1 \times r^*$ matrix consisting of i.i.d. entries from a certain distribution $\mathcal{P}_{u,v}$ (specified below) and $\mathbf{V}^* \in \mathbb{R}^{d_2 \times r^*}$ is independently generated in the same manner. In addition, the constant κ is chosen such that the average magnitude of the entries $|M^*_{ij}|$ is 2. Following the observation model (2.1), we only observe the entries in the random set $\mathcal{S}$.

Implementation. First, when implementing *als*, we adopt r as the hypothesized rank of the underlying matrix. For model-based methods (Chen et al., 2019), by the asymptotic distributional characterization $\widehat{M}_{ij} - M^*_{ij} \xrightarrow{d} \mathcal{N}(0, \theta^2_{ij})$ and $E_{ij} \perp (\widehat{M}_{ij} - M^*_{ij})$, one has $\widehat{M}_{ij} - M_{ij} \xrightarrow{d} \mathcal{N}(0, \theta^2_{ij} + \sigma^2)$. Consequently, the confidence interval for the model-based methods can be constructed as

$$\widehat{M}_{ij} \pm q_{1-\alpha/2} \sqrt{\widehat{\theta}^2_{ij} + \widehat{\sigma}^2},$$

where q_β is the βth quantile of $\mathcal{N}(0,1)$. The estimates are obtained via

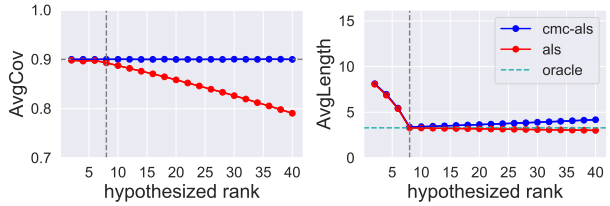
$$\widehat{\sigma}^2 = \|\mathbf{M}_{\mathcal{S}_{\text{train}}} - \widehat{\mathbf{M}}_{\mathcal{S}_{\text{train}}}\|_{\text{F}}^2 / |\mathcal{S}_{\text{train}}| \qquad \text{and} \qquad \widehat{\theta}^2_{ij} = (\widehat{\sigma}^2/\widehat{p})(\|\widehat{\mathbf{U}}_{i,\cdot}\|^2 + \|\widehat{\mathbf{V}}_{j,\cdot}\|^2)$$

with the SVD $\widehat{\mathbf{M}} = \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^\top$. For cmc-als, we obtain $\widehat{\mathbf{M}}$, $\widehat{\mathbf{s}}$ from $\mathbf{M}_{\mathcal{S}_{\text{train}}}$ by running als on this training set (and taking $\widehat{s}_{ij} = (\widehat{\theta}^2_{ij} + \widehat{\sigma}^2)^{1/2}$). The probabilities $\widehat{p}_{ij}$'s are estimated via $\widehat{p}_{ij} \equiv (d_1 d_2 q)^{-1} |\mathcal{S}_{\text{train}}|$ for this setting where the p_{ij}'s are homogeneous.

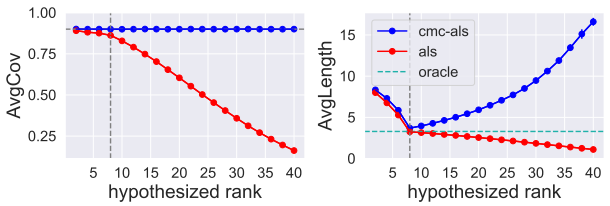
Homogeneous missingness

In the homogeneous case, we consider the following four settings with $d_1 = d_2 = 500$:

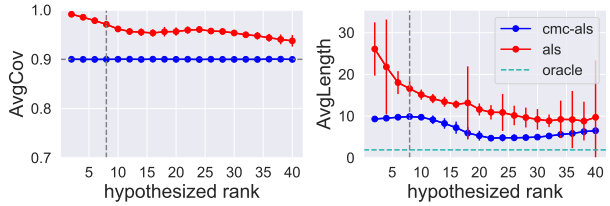
- **Setting 1: large sample size + Gaussian noise**. $p = 0.8$, $\mathcal{P}_{u,v} = \mathcal{N}(0,1)$, and $E_{ij} \sim \mathcal{N}(0,1)$.



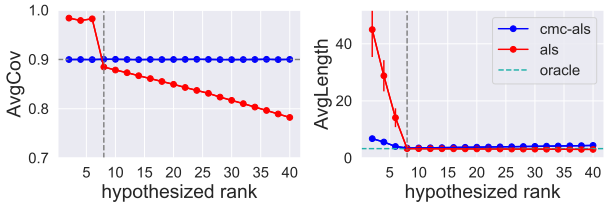
(a) Setting 1: large sample size + Gaussian noise



(b) Setting 2: small sample size + Gaussian noise



(c) Setting 3: large sample size + heavy-tailed noise



(d) Setting 4: violation of incoherence

Figure 2.1: Comparison between $cmc - als$ and als.

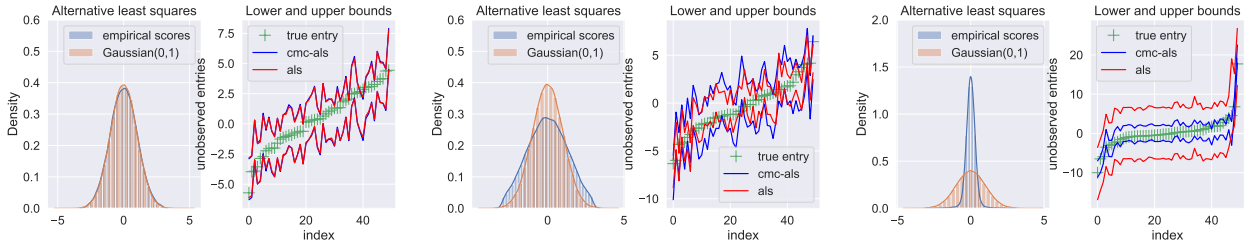
- **Setting 2: small sample size + Gaussian noise**. $p = 0.2$, $\mathcal{P}_{u,v} = \mathcal{N}(0,1)$, and $E_{ij} \sim \mathcal{N}(0,1)$.

- **Setting 3: large sample size + heavy-tailed noise**. $p = 0.8$, $\mathcal{P}_{u,v} = \mathcal{N}(0,1)$, and $E_{ij} \sim 0.2\ \mathrm{t}_{1.2}$, where t_ν denotes the t distribution with ν degrees of freedom.

- **Setting 4: violation of incoherence**. $p = 0.8$, $\mathcal{P}_{u,v} = t_{1.2}$, and $E_{ij} \sim \mathcal{N}(0,1)$.

Figure 2.1 displays the results (i.e., the coverage rate and the interval length) for both $cmc - als$ and als when we vary the hypothesized rank r from 2 to 40. The oracle length is the difference between the $(1-\alpha/2)$-th and $(\alpha/2)$-th quantiles of the underlying distribution for E_{ij}. As can be seen, cmc-als achieves nearly exact coverage regardless of the choice of r across all four settings. When the hypothesized rank r is chosen well, cmc-als is often able to achieve an average length similar to that of the oracle.

For the model-based method als, when we underestimate the true rank ($r < r^*$), we may still see adequate coverage since $\widehat{\sigma}$ tends to over-estimate σ by including the remaining $r^* - r$ factors as noise. However, if we overestimate the rank ($r > r^*$), then als tends to



(a) Large sample size + Gaussian noise: $r = r^* = 8$

(b) Large sample size + Gaussian noise: $r^* = 8 < 50 = r$

(c) Violation of incoherence: $r = 6 < 8 = r^*$

Figure 2.2: Histogram of standardized scores for als and prediction lower and upper bounds for 50 distinct unobserved entries.

overfit the noise with additional factors and under-estimate σ regardless of the choice of r, leading to significant undercoverage in Settings 1, 2 and 4. In Setting 3 with heavy-tailed noise, als tends to be conservative due to overestimating σ. Overall, Figure 2.1 confirms our expectation that the validity of the model-based estimates relies crucially on a well specified model, and it fails to hold when sample size is small (cf. Figure 2.1b), when noise is heavy tailed (cf. Figure 2.1c), and when the underlying matrix is not incoherent (cf. Figure 2.1d).

In Figure 2.2, we present the histogram of standardized scores $(\widehat{M}_{ij} - M_{ij})/\sqrt{\widehat{\theta}_{ij}^2 + \widehat{\sigma}^2}$ and the plot of the upper and lower bounds for three settings. In Figure 2.2a, when the model assumptions are met and $r = r^*$, the scores match well with the standard Gaussian and the prediction bounds produced by als and cmc-als are similar. With the same data generating process, when the rank is overparametrized, the distribution of scores cannot be captured by the standard Gaussian, thus the quantiles are misspecified. As we can see from the confidence intervals, als tends to have smaller intervals which lead to the undercoverage. In the last setting, the underlying matrix is no longer incoherent. When the rank is underestimated, the $r^* - r$ factors will be captured by the noise term and the high heterogeneity in the entries will further lead to overestimated noise level. As a result, the intervals by als are much larger while the conformalized intervals are more adaptive to the magnitude of entries.

Heterogeneous missingness

Now we move on to the case with heterogeneous missingness, where the observed entries are no longer sampled uniformly. To simulate this setting, we draw $\{a_{il} : i \leq d_1, l \leq k^*\}$ i.i.d. from $\mathrm{Unif}(0,1)$ and $\{b_{lj} : l \leq k^*, j \leq d_2\}$ i.i.d. from $\mathrm{Unif}(-0.5, 0.5)$, and define sampling probabilities by

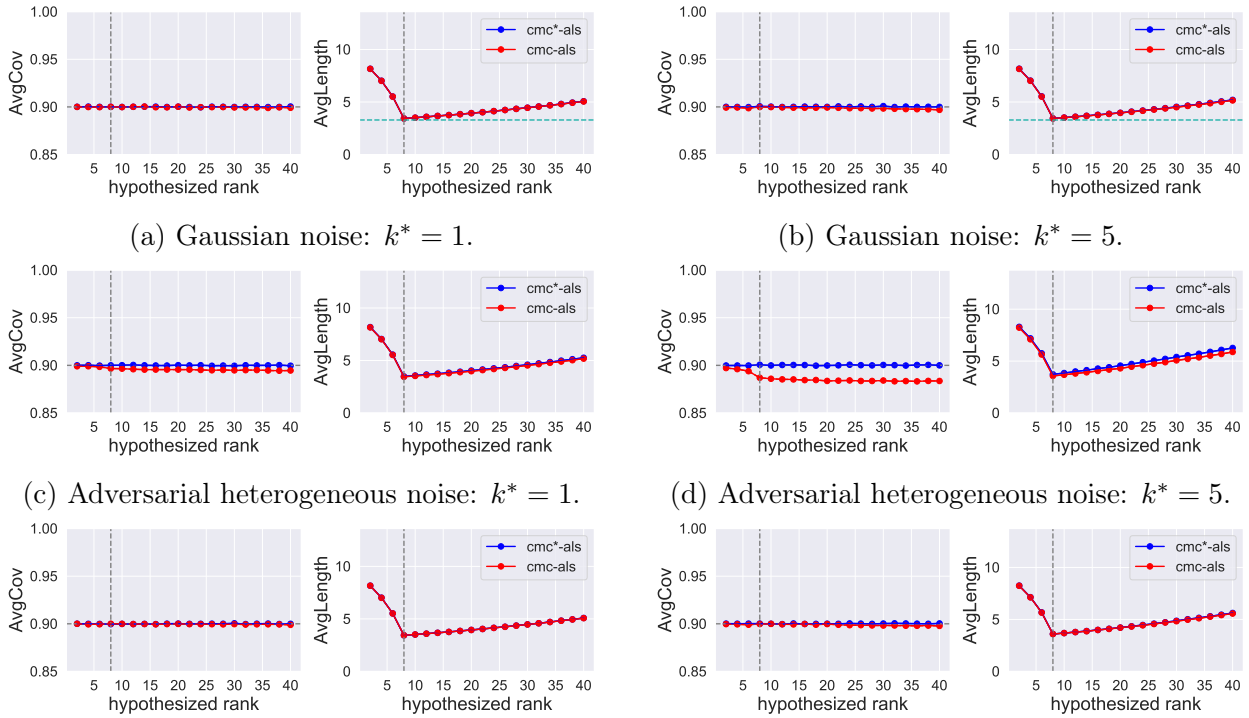
$$\log(p_{ij}/(1-p_{ij})) = \sum_{l=1}^{k^*} a_{il} b_{lj}.$$

We consider two settings with $k^* = 1$ and $k^* = 5$, respectively. In both settings, $\widehat{p}_{ij}$ is constructed by estimating $\widehat{a}_{il}$ and $\widehat{b}_{lj}$ via constrained maximum likelihood estimation as is shown in Example 2.3.3.

To analyze the effect of estimation error predicted by Theorem 2.3.2, we consider three matrix generation processes for each choice of k^*, where $\mathbf{M}^*$ is generated in the same way as in Section 2.4.1 with $d_1 = d_2 = d = 500$, $\mathcal{P}_{u,v} = \mathcal{N}(0,1)$:

- **Gaussian homogeneous noise**: $E_{ij} \sim \mathcal{N}(0,1)$ independently.
- **Adversarial heterogeneous noise**: $E_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2)$ independently, where we take $\sigma_{ij} = 1/2p_{ij}$. For this setting, note that the high-noise entries (i.e., those entries that are hardest to predict) occur in locations (i,j) that are least likely to be observed during training.
- **Random heterogeneous noise**: $E_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2)$ independently, where $(\sigma_{ij})_{(i,j)\in[d_1]\times[d_2]}$ is drawn from the same distribution as $(1/2p_{ij})_{(i,j)\in[d_1]\times[d_2]}$.

We write cmc*-als to denote the conformalized method (with als as the base algorithm) where we use oracle knowledge of the true p_{ij}'s for weighting, while cmc-als uses estimates $\widehat{p}_{ij}$. The results are shown in Figures 2.3, for the settings $k^* = 1$ and $k^* = 5$ (i.e., the rank of $\mathbf{P}$). Under both homogeneous noise and random heterogeneous noise, both cmc*-als and cmc-als achieve the correct coverage level, with nearly identical interval length. For adversarial heterogeneous noise, on the other hand, cmc*-als achieves the correct coverage



(a) Gaussian noise: $k^* = 1$.

(b) Gaussian noise: $k^* = 5$.

(c) Adversarial heterogeneous noise: $k^* = 1$.

(d) Adversarial heterogeneous noise: $k^* = 5$.

(e) Random heterogeneous noise: $k^* = 1$.

(f) Random heterogeneous noise: $k^* = 5$.

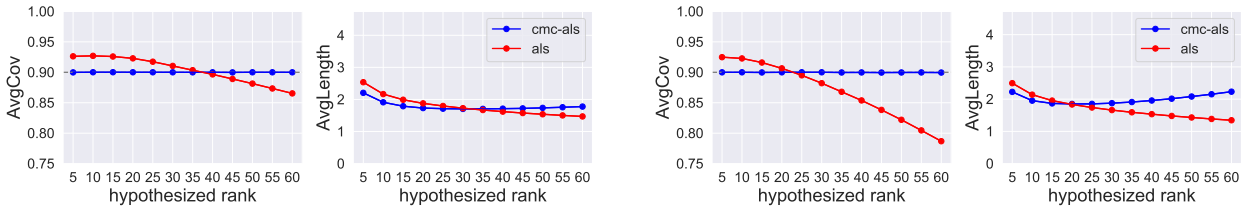
Figure 2.3: Comparison between cmc with true and estimated weights under heterogeneous missingness.

level as guaranteed by the theory, but cmc-als shows some undercoverage due to the errors in the estimates $\widehat{p}_{ij}$ (since now the variance of the noise is adversarially aligned with these errors); nonetheless, the undercoverage is mild. In Section A.2.2 in the appendix, the local coverage of cmc-als conditioning $p_{ij} = p_0$ for varying p_0 is presented.

2.4.2 Real data application

We also compare conformalized approaches with model-based approaches using the Rossmann sales dataset[5] (Farias et al., 2022). This real data provides the underlying fixed matrix $\mathbf{M}$; the missingness pattern is generated artificially, as detailed below. We focus on daily sales (the unit is 1K dollar) of 1115 drug stores on workdays from Jan 1, 2013 to July

5. `https://www.kaggle.com/c/rossmann-store-sales`



(a) als vs cmc-als: homogeneous missingness.

(b) als vs cmc-als: logistic missingness.

Figure 2.4: Comparison between conformalized and model-based matrix completion with sales dataset.

31, 2015 and hence the underlying matrix has dimension 1115×780. The hypothesized rank r is varied from 5 to 60 with step size 5 and we set the target level to be $1 - \alpha = 0.9$. We apply random masking for 100 independent trials and report the average for $AvgCov$ and $AvgLength$ in Figure 2.4.

- **Homogeneous** p_{ij}. For each entry, we first apply random masking $1 - Z_{ij} \sim \text{Bern}(0.2)$ independently. In Figure 2.4a, conformalized approach has exact coverage at $1 - \alpha$ but als tends to be conservative when r is small and loses coverage when r increases to 50.

- **Heterogeneous** p_{ij}**.** After drawing p_{ij}'s from $\log(p_{ij}/(1-p_{ij})) = -\log\left(h_{ij}\right) = \phi(A_{ij})$ in Section 2.4.1 with $k^* = 1$, $\widehat{\mathbf{P}}$ is obtained via the constrained maximum likelihood estimator in Example 2.3.3. In Figure 2.4b, $cmc - als$ has coverage around $1-\alpha$ and in comparison, when r is greater than 40, als fails to guarantee the coverage and $AvgCov$ decreases to 0.75.

2.5 Discussion

In this paper, we introduce the conformalized matrix completion method cmc, which offers a finite-sample coverage guarantee for all missing entries on average and requires no assumptions on the underlying matrix or any specification of the algorithm adopted, relying only on an estimate of the entrywise sampling probabilities. Given an estimate of the entrywise sampling probability, we provide an upper bound for the coverage gap and give the explicit

form with examples of the logistic as well as the general one-bit low-rank missingness. In the implementation, an efficient one-shot weighted conformal approach is proposed with the provable guarantee and achieves nearly exact coverage.

We can compare our findings to a few related results in the literature. The work of Chernozhukov et al. (2021) applies conformal prediction to counterfactual and synthetic control methods, where they include matrix completion as an example with a regression-type formulation. However, their approach relies on a test statistic that is a function of estimated residuals. Consequently, their method requires the assumption of stationarity and weak dependence of errors. Furthermore, the validity of their approach is contingent upon the estimator being either consistent or stable. Additionally, Wieczorek (2023) studies conformal prediction with samples from a deterministic and finite population, but the validity under the sampling without replacement remains an open question in their work.

Our results suggest a number of questions for further inquiry. In the setting of matrix completion, while the results are assumption-free with regard to the matrix $\mathbf{M}$ itself, estimating the sampling probabilities $\mathbf{P}$ that determine which entries are observed remains a key step in the procedure; while our empirical results suggest the method is robust to estimation error in this step, studying the robustness properties more formally is an important open question. More broadly, our algorithm provides an example of how the conformal prediction framework can be applied in a setting that is very different from the usual i.i.d. sampling regime, and may lend insights for how to develop conformal methodologies in other applications with non-i.i.d. sampling structure.

CHAPTER 3

DISTRIBUTIONALLY ROBUST INFERENCE WITH SHAPE CONSTRAINTS

Statistical learning under distribution shift is challenging when neither prior knowledge nor data from the target distribution is available. Distributionally robust learning (DRL) aims to control the worst-case statistical performance within a set of candidate distributions, but how to properly specify the set remains challenging. To enable distributional robustness without being overly conservative, in this paper we propose a shape-constrained approach to DRL, which incorporates prior information about the way in which the unknown target distribution differs from its estimate—specifically, we assume the unknown density ratio between the target distribution and its estimate is isotonic with respect to some partial order. At the population level, we provide a solution to the shape-constrained optimization problem that can be solved without the challenge of an explicit isotonic constraint. At the sample level, we provide consistency results for an empirical estimator of the target in a range of different settings. Empirical studies on both synthetic and real data demonstrate the improved efficiency of the proposed shape-constrained approach. [1]

3.1 Introduction

Evaluating the performance of an estimator is of significant importance in statistics. To give several motivating examples, we first consider supervised learning settings, where our observations consist of features $X \in \mathcal{X} \subseteq \mathbb{R}^d$ and a response $Y \in \mathcal{Y} \subseteq \mathbb{R}$:

- Given a fitted model $\widehat{\mu} : \mathcal{X} \to \mathbb{R}$, we may want to estimate the expected value of the squared error $(Y - \widehat{\mu}(X))^2$ with respect to a target distribution on (X, Y).

1. This chapter is based on Gui et al. (2024b).

- Or, in predictive inference, suppose we have constructed a prediction band $\widehat{C}_{1-\alpha}$, where $\widehat{C}_{1-\alpha}(X) \subseteq \mathbb{R}$ is a confidence region for the response Y given features X, and $1-\alpha$ denotes the target coverage level. Then to determine whether $\widehat{C}_{1-\alpha}$ does in fact achieve coverage at level $1-\alpha$ for data points drawn from some target distribution, we would like to estimate the expected value of $\mathbb{1}\{Y \notin \widehat{C}_{1-\alpha}(X)\}$ with respect to this target distribution. This is the probability that our interval fails to cover the response.

We can also consider unsupervised learning settings, where observations consist only of features $X \in \mathcal{X} \subseteq \mathbb{R}^d$:

- In principal component analysis (PCA), suppose we have obtained a set of pre-fitted principal components $\widehat{\mathcal{V}}_K = \{\widehat{v}_1, \ldots, \widehat{v}_K\}$ which forms an orthonormal basis for a K-dimensional subspace of $\mathbb{R}^d$. To evaluate how well the variance in X is explained by the top K principal components, it would be of interest to analyze the expected value of the reconstruction error $\|X - \sum_{k=1}^{K}(X^\top \widehat{v}_k)\widehat{v}_k\|^2$ with respect to the distribution of X.

- Another example is density estimation. In this case, given a density estimate P_θ learned from data, we may want to evaluate its performance using the expected log-likelihood $-\log \mathsf{d}P_\theta(X)$ over a target distribution P_{target}. In fact, $\mathbb{E}_{P_{\text{target}}}[-\log \mathsf{d}P_\theta(X)]$ is the cross-entropy of P_θ relative to P_{target}.

A key challenge for any of these problems is that the target distribution (say, the distribution of the general population) may be unknown, and our available data (say, individuals who participate in our study) may be drawn from a different distribution.

3.1.1 Problem formulation

To make the problem more concrete, and unify the examples mentioned above, here we introduce some notation to formulate the question at hand.

The unsupervised setting. Let $R : \mathcal{X} \to \mathbb{R}_+$ denote a risk function, where our goal is to evaluate the expected value $\mathbb{E}_{P_{\text{target}}}[R(X)]$ with respect to some target distribution P_{target} over $\mathcal{X}$. However, the available data only provides information about P, a potentially different distribution. Using a calibration data set comprised of samples $X_1, \ldots, X_n$ drawn from P, we can estimate $\mathbb{E}_P[R(X)]$ with the empirical mean, $n^{-1}\sum_{i=1}^n R(X_i))$. Our aim, though, is to provide a bound on the risk $\mathbb{E}_{P_{\text{target}}}[R(X)]$—or, at least, to bound the difference in risks (often called the excess risk), $\mathbb{E}_{P_{\text{target}}}[R(X)] - \mathbb{E}_P[R(X)]$.

If we assume that the unknown distribution P_{target} lies in some class $\mathcal{Q}$ (to be specified later on), then defining the worst-case excess risk

$$\Delta(R; \mathcal{Q}) = \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q[R(X)] - \mathbb{E}_P[R(X)], \tag{3.1}$$

we can bound the risk under distribution P_{target} by $\mathbb{E}_{P_{\text{target}}}[R(X)] \leq \mathbb{E}_P[R(X)] + \Delta(R; \mathcal{Q})$.

The supervised setting: covariate shift assumption. In the supervised learning setting, the data contains both features X and a response Y. Here we will consider a loss function $r : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, for instance, $r(x, y) = (y - \widehat{\mu}(x))^2$ for the squared error in a regression, or $r(x, y) = \mathbb{1}\{y \notin \widehat{C}_{1-\alpha}(x)\}$ for characterizing the (mis)coverage of a prediction interval in predictive inference.

Throughout this paper, for the supervised learning setting, we will assume the covariate shift setting, where the distribution of the available data and the target distribution may differ in the marginal distribution of the covariates X, but share the same conditional distribution $Y \mid X$. To make this concrete, if our calibration data consists of n data points $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn from $\widetilde{P}$, while our goal is to control the expected loss with respect to the target distribution $\widetilde{P}_{\text{target}}$ on (X, Y), we will assume that we can write

$$\text{data distribution: } \widetilde{P} = P \times P_{Y|X}, \quad \text{target distribution: } \widetilde{P}_{\text{target}} = P_{\text{target}} \times P_{Y|X},$$

so that $\widetilde{P}$ and $\widetilde{P}_{\text{target}}$ share the same conditional distribution $P_{Y|X}$ for $Y \mid X$.

In fact, under covariate shift, this supervised setting can be unified with the unsupervised one by defining the risk $R(X) = \mathbb{E}[r(X,Y) \mid X]$, which is the conditional expectation of $r(X,Y)$ under either $\widetilde{P}$ or $\widetilde{P}_{\text{target}}$. The quantity of interest is then given by $\mathbb{E}_{P_{\text{target}}}[R(X)] = \mathbb{E}_{\widetilde{P}_{\text{target}}}[r(X,Y)]$, but our calibration data, which is sampled from P, instead enables us to estimate $\mathbb{E}_P[R(X)] = \mathbb{E}_{\widetilde{P}}[r(X,Y)]$. If we again assume that $P_{\text{target}} \in \mathcal{Q}$, then $\Delta(R;\mathcal{Q})$ again allows us to bound the risk of our estimator under the target distribution:

$$\mathbb{E}_{\widetilde{P}_{\text{target}}}[r(X,Y)] \leq \mathbb{E}_{\widetilde{P}}[r(X,Y)] + \Delta(R;\mathcal{Q}).$$

Estimating the risk or tuning the model? In this paper, we consider the setting where our estimator—say, a prediction band $\widehat{C}_{1-\alpha}$—is pretrained, meaning that we have available calibration data sampled from P (in the unsupervised setting) or $\widetilde{P}$ (in the supervised setting) that is independent of the fitted estimator. Consequently, our available calibration data provides us with an unbiased estimate of $\mathbb{E}_P[R(X)]$ (or, equivalently in the supervised setting, $\mathbb{E}_{\widetilde{P}}[r(X,Y)]$); given a constraint set $\mathcal{Q}$, we can then use this estimate to bound $\mathbb{E}_{P_{\text{target}}}[R(X)]$ (or, in the supervised setting, $\mathbb{E}_{\widetilde{P}_{\text{target}}}[r(X,Y)]$).

In some settings, the goal may be to estimate the risk of each estimator within a family of (pretrained) options, in order to select a good estimator. Returning again to the example of a prediction band, suppose, we actually are given a nested family of prediction bands, $\{\widehat{C}_{1-a} : a \in [0,1]\}$, where $1-a$ denotes the confidence level. Choosing $R_a(X) = \mathbb{P}_{P_{Y|X}}(Y \notin \widehat{C}_{1-a}(X))$ or accordingly, $r_a(X,Y) = \mathbb{1}\{Y \notin \widehat{C}_{1-a}(X)\}$, then, if we can compute a bound on the miscoverage rate $\mathbb{E}_{P_{\text{target}}}[R_a(X)]$ for each a, then we can choose a value of a that

achieves some desired level of coverage. More generally, we may do the same in other settings as well—that is, given a family of candidate estimators, bounding the risk of each one under the distribution P_{target} provides an intermediate step towards choosing the tuning parameter.

Throughout this paper, then, we will primarily discuss the question of estimating the expected risk. Later on, in our experiments, we will turn to the aim of using these estimates to tune a procedure for achieving a desired bound on the error.

3.1.2 Prior work: distributionally robust learning

Our work builds upon the distributionally robust learning (DRL) literature (Ben-Tal and Nemirovski, 1998; El Ghaoui et al., 1998; Lam, 2016; Duchi and Namkoong, 2018), which is a well-established framework for risk evaluation under distribution shift. In this framework, the target distribution P_{target} is assumed to lie in some neighborhood around the distribution P of the available data—for instance, we might assume that $D_{\text{KL}}(P_{\text{target}} \| P) \leq \rho$, where D_{KL} denotes the Kullbeck–Leibler (KL) divergence. DRL takes a conservative approach and evaluate the performance on P_{target} via its upper bound, i.e., the worst-case performance over all distributions within the specified neighborhood of P,

$$\mathbb{E}_{P_{\text{target}}}[R(X)] \leq \sup \left\{ \mathbb{E}_Q[R(X)] : D_{\text{KL}}(Q\|P) \leq \rho \right\}, \tag{3.2}$$

or, equivalently, $\mathbb{E}_{P_{\text{target}}}[R(X)] \leq \mathbb{E}_P[R(X)] + \Delta(R; \mathcal{Q}_{\text{KL}}(\rho))$, where $\Delta(R; \mathcal{Q}_{\text{KL}}(\rho))$ is defined as in (3.1) with $\mathcal{Q} = \mathcal{Q}_{\text{KL}}(\rho) = \{Q : D_{\text{KL}}(Q\|P) \leq \rho\}$. More generally, we can consider divergence measures beyond the KL distance, as we will describe in more detail below.

3.1.3 Our proposal: iso-DRL

If the assumption $D_{\text{KL}}(P_{\text{target}}\|P) \leq \rho$ is correct, then the upper bound (3.2) is valid. However, since this bound uses only the KL divergence to define the constraint $P_{\text{target}} \in \mathcal{Q}$ on

the target distribution, it could be quite conservative. In many practical settings, additional side information or prior knowledge on the structure of the distribution shift may allow for a tighter bound, which would be less conservative than the worst-case excess risk of DRL (3.2). This raises the following key question:

Can we use side information on the distribution shift between the distribution P and the target distribution P_{target}, to improve the worst-case excess risk of DRL in risk evaluation?

In this paper, we study one specific example of this type of setting: we assume that the density ratio $(\mathsf{d}P_{\text{target}}/\mathsf{d}P)(\cdot)$ between the target distribution and the data distribution is (approximately) isotonic (i.e., monotone) with respect to some order or partial order on $\mathcal{X}$.

Motivation: recalibration of an estimated density ratio. To motivate the use of such side information, consider a practical supervised setting where we have an initial estimate w_0 for the density ratio:

$$w_0(x) \approx \frac{\mathsf{d}P_{\text{target}}}{\mathsf{d}P}(x).$$

This ratio is possible to estimate in settings where, in addition to labeled data (i.e., (X, Y) pairs) sampled from the data distribution $P \times P_{Y|X}$, we also have access to unlabeled (i.e., X only) data from the target population P_{target}. We may use these two data sets to train w_0. Although there is no guarantee that the estimate w_0 is accurate, the shape or relative magnitude of w_0 may provide us with useful side information: large values of w_0 can identify portions of the target population that are underrepresented under the data distribution P. This motivates us to recalibrate w_0 within the set of density ratios that are isotonic in w_0.

To express this scenario in the notation of the problem formulation above, we assume that the target distribution P_{target} satisfies an isotonicity constraint, $P_{\text{target}} \in \mathcal{Q}_{\text{iso}}(w_0)$, where

$$\mathcal{Q}_{\text{iso}}(w_0) = \left\{ Q : \frac{\mathsf{d}Q}{\mathsf{d}P}(x) \text{ is a monotonically nondecreasing function of } w_0(x) \right\}.$$

If we assume as before that the target distribution P_{target} satisfies $D_{\text{KL}}(P_{\text{target}}\|P) \leq \rho$, then we can bound

$$\mathbb{E}_{P_{\text{target}}}[R(X)] \leq \mathbb{E}_P[R(X)] + \Delta(R; \mathcal{Q}_{\text{KL}}(\rho) \cap \mathcal{Q}_{\text{iso}}(w_0)). \tag{3.3}$$

The benefits of iso-DRL. What are the benefits of iso-DRL, as compared to the existing DRL framework? Of course, thus far the idea is quite straightforward—if we have stronger constraints on P_{target}, then we can place a tighter bound on the excess risk $\mathbb{E}_{P_{\text{target}}}[R(X)] - \mathbb{E}_P[R(X)]$. But as we will see below, adding the isotonic constraint plays a crucial role in enabling DRL to provide bounds that are useful in practical scenarios. Specifically, consider a practical setting where the bound ρ on the distribution shift is a positive constant. As we will see below, the existing worst-case excess risk $\Delta(R; \mathcal{Q}_{\text{KL}}(\rho))$ of DRL is often quite large, leading to extremely conservative statistical conclusions; in contrast, the worst-case excess risk $\Delta(R; \mathcal{Q}_{\text{KL}}(\rho) \cap \mathcal{Q}_{\text{iso}}(w_0))$ given by iso-DRL is often vanishingly small, leading to much more informative conclusions. Moreover, surprisingly, this improvement in the bound does not incur any additional computational challenges—even though the constraint set $\mathcal{Q}_{\text{KL}}(\rho) \cap \mathcal{Q}_{\text{iso}}(w_0)$ appears more complex than the original set $\mathcal{Q}_{\text{KL}}(\rho)$, we will see that $\Delta(R; \mathcal{Q}_{\text{KL}}(\rho) \cap \mathcal{Q}_{\text{iso}}(w_0))$ can be computed nearly as easily as the original quantity $\Delta(R; \mathcal{Q}_{\text{KL}}(\rho))$. In addition, we further show in Appendix Section B.3.6 that the worst-case excess risk of iso-DRL can be consistently estimated with noisy observations of $R(X)$, while the estimation of the worst-case excess risk of DRL can be challenging even with bounded risks.

Empirical example: predictive inference for the wine quality dataset. To illustrate the advantage of the proposed approach, Figure 3.1 presents a numerical example for a predictive inference problem on the wine quality dataset (Cortez et al., 2009).[2] (See

2. Available at `https://archive.ics.uci.edu/dataset/186/wine+quality`

Section 3.5.2 for full details of this experiment.)

We are given a pretrained family of prediction bands $\widehat{C}_{1-a}$, indexed by the target coverage level $1-a$. At each value $a \in [0,1]$, we define $R_a(X) = \mathbb{P}(Y \notin \widehat{C}_{1-a}(X) \mid X)$, the probability of the prediction band failing to cover the true response value Y given features X. Our goal is to return a prediction band with 90% coverage—that is, we would like to choose a value of a such that the expected risk $\mathbb{E}_{P_{\text{target}}}[R_a(X)] = \mathbb{P}_{\widetilde{P}_{\text{target}}}(Y \notin \widehat{C}_{1-a}(X))$ is bounded by $0.1 = 1 - 90\%$. In our experiment, the available data is given by all samples that are white wines (with distribution $\widetilde{P}$), while the target population is comprised of the samples that are red wines (with a different distribution $\widetilde{P}_{\text{target}}$). In Figure 3.1, we compare four methods

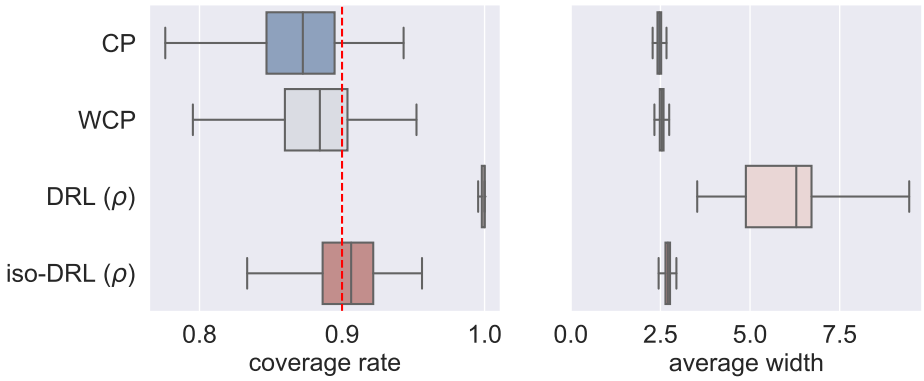


Figure 3.1: Coverage rate and average width of intervals for the wine quality dataset. The red dashed line (in the left-hand plot) marks the nominal coverage level, $1 - \alpha = 90\%$.

(see Section 3.5.2 for details):

- An uncorrected interval—using conformal prediction (CP) (Vovk et al., 2005): the value a is chosen by tuning on the calibration data set (i.e., we choose a to satisfy $\mathbb{E}_P[R_a(X)] \leq 0.1$), without correcting for the distribution shift.

- A corrected interval—using weighted conformal prediction (WCP) (Tibshirani et al., 2019): the value a is chosen by tuning on the calibration data set using an estimated density ratio w_0 to correct for the covariate shift between distributions $\widetilde{P}$ and $\widetilde{P}_{\text{target}}$. Since w_0 is estimated from data, this correction is imperfect.

- The DRL interval: we choose a to satisfy $\mathbb{E}_P[R_a(X)] + \Delta(R_a; \mathcal{Q}_{\mathrm{KL}}(\rho)) \leq 0.1$, where $\mathbb{E}_P[R_a(X)]$ and $\Delta(R_a; \mathcal{Q}_{\mathrm{KL}}(\rho))$ are estimated using the calibration data.

- The iso-DRL interval: we choose a to satisfy $\mathbb{E}_P[R_a(X)] + \Delta(R_a; \mathcal{Q}_{\mathrm{KL}}(\rho) \cap \mathcal{Q}_{\mathrm{iso}}(w_0)) \leq 0.1$, where $\mathbb{E}_P[R_a(X)]$ and $\Delta(R_a; \mathcal{Q}_{\mathrm{KL}}(\rho) \cap \mathcal{Q}_{\mathrm{iso}}(w_0))$ are estimated using the calibration data.[3]

As we can see in Figure 3.1, the CP and WCP intervals both undercover—for CP, this is because the method does not correct for distribution shift, while for WCP, this is because the ratio w_0 that corrects for distribution shift is imperfectly estimated. At the other extreme, DRL shows substantial overcoverage with extremely wide prediction intervals due to the worst-case nature of the bound $\Delta(R_a; \mathcal{Q}_{\mathrm{KL}}(\rho))$. In contrast, our proposed method, iso-DRL, achieves the target coverage rate 90% without excessive increase in the size of the prediction interval, showing the benefit of adding the isotonic constraint to the DRL framework.

The motivating example demonstrates that, when we have access to meaningful—but imperfect—side information (e.g., in the form of the density ratio w_0), adding the isotonic constraint to iso-DRL can provide an estimate of the risk that is more reliable than a non-distributionally-robust approach, but less conservative than the original DRL approach.

3.1.4 Organization of paper

Section 3.2 introduces a general class of uncertainty sets for candidate distributions and further studies the property of the worst-case excess risk defined in (3.1) for generic DRL. For the worst-case excess risk with the isotonic constraint, we prove that it is equivalent to the worst-case excess risk for a projected risk function without the isotonic constraint in Section 3.3. In Section 3.4, we propose an estimator of the worst-case excess risk with the isotonic constraint and establish the estimation error bounds. Numerical results for both

3. For both the DRL and iso-DRL methods, the parameter ρ is an estimate of the actual KL distance $D_{\mathrm{KL}}(P_{\mathrm{target}} \| P)$—see Section 3.5.2 and Appendix B.4.2 for details

synthetic and real data are shown in Section 3.5 and additional related work is summarized in Section 3.6. We defer technical proofs and additional simulations to the Appendix.

Notation. Before proceeding, we introduce useful notation for theoretical developments later on. To begin with, we write $(a)_+$ as the positive part of $a \in \mathbb{R}$. We denote by $L_p(P)$ $(1 \le p \le \infty)$ the L_p function space under the probability measure P, i.e., when $p \neq \infty$, $L_p(P) = \{f : \|f\|_p = (\int_{\mathcal{X}} |f(x)|^p \mathsf{d}P(x))^{1/p} < \infty\}$. When $p = \infty$, the set $L_\infty(P)$ consists of measurable functions that are bounded almost surely under P. In addition, for a measurable function w defined on $\mathcal{X}$ and a measure P on $\mathcal{X}$, the pushforward measure $w_\# P$ denotes the measure satisfiying that $(w_\# P)(B) = P(w^{-1}(B))$ for any measurable set B, where $w^{-1}(B) = \{x \in \mathcal{X} : w(x) \in B\}$ denotes the preimage of B under w. In other words, if $X \sim P$, then $w(X)$ follows the distribution $w_\# P$. We say a function h is A_h-bounded if $\sup_x |h(x)| \le A_h$. Fix a partial (pre)order $\preceq$ on $\mathcal{X} \subseteq \mathbb{R}^d$. A function g is isotonic if $g(x_1) \le g(x_2)$ for any $x_1 \preceq x_2$. Correspondingly, we define the cone of isotonic functions by $\mathcal{C}^{\text{iso}}_{\preceq} = \{w : w \text{ is isotonic w.r.t. partial order } \preceq\}$. Lastly, to compare two probability distributions Q and P, the convex ordering $\overset{cvx}{\preceq}$ is defined as $Q' \overset{cvx}{\preceq} Q$ if and only if $\mathbb{E}_{Q'}[\psi(X)] \le \mathbb{E}_Q[\psi(X)]$ for all convex functions ψ.

3.2 The distributional robustness framework

As we have explained in Section 3.1.1, both the unsupervised and supervised setting under covariate shift can be unified. Therefore, throughout this section, to develop our theoretical results we will use the notation of the unsupervised setting with the risk function $R(X)$, with the understanding that this also covers the supervised setting under covariate shift.

Recall that $\mathcal{X}$ is the feature domain. We consider a bounded risk function $R : \mathcal{X} \to [0, B_R]$ with $0 < B_R < \infty$. The goal is to evaluate (or bound) the target risk $\mathbb{E}_{P_{\text{target}}}[R(X)]$ using samples from P, by assuming that the target distribution P_{target} is in some sense similar

to the available distribution P—more concretely, by assuming that the target distribution P_{target} lies in some neighborhood $\mathcal{Q}$ around the distribution P of the available data.

Reformulating the neighborhood. To unify the different examples of constraints described in Section 3.1, we will start by considering settings where we can express the constraint $Q \in \mathcal{Q}$ using conditions on the density ratio $w = \mathsf{d}Q/\mathsf{d}P$. This type of framework includes the sensitivity analysis setting via bounds on w (Cornfield et al., 1959; Rosenbaum, 1987; Tan, 2006; Ding and VanderWeele, 2016; Zhao et al., 2019; Yadlowsky et al., 2018; Jin et al., 2022), and f-divergence constraints such as a bound on the KL divergence (Duchi et al., 2021; Namkoong and Duchi, 2017; Duchi and Namkoong, 2018; Cauchois et al., 2020).

Concretely, we can reparameterize the distribution Q using the density ratio $w(x) = (\mathsf{d}Q/\mathsf{d}P)(x)$. Then we can reformulate the constraint $Q \in \mathcal{Q}$ into a constraint on this density ratio, i.e., $Q \in \mathcal{Q} \iff w_{\#}P \in \mathcal{B}$, where $\mathcal{B}$ is a set of distributions, and where $w_{\#}P$ denotes the pushforward measure (as defined in Section 3.1.4). Let us now consider the two examples mentioned above.

Example 1: bound-constrained distribution shift. In sensitivity analysis, it is common to assume that the likelihood ratio $\mathsf{d}P_{\text{target}}/\mathsf{d}P$ is bounded from above and below. This corresponds to a constraint set of the form $\mathcal{Q} = \{Q : a \leq (\mathsf{d}Q/\mathsf{d}P)(X) \leq b \ P\text{-almost surely}\}$, for some constants $0 \leq a < 1 < b < +\infty$. In particular, when $a = \Gamma^{-1}$ and $b = \Gamma$ for some $\Gamma > 1$, this constraint set represents the marginal Γ-selection model for the density ratio in sensitivity analysis (Rosenbaum, 1987; Tan, 2006). By defining

$$\mathcal{B} = \mathcal{B}_{a,b} = \left\{\widetilde{Q} : \ \mathbb{E}_{Z\sim\widetilde{Q}}[Z] = 1, \ \mathbb{P}_{Z\sim\widetilde{Q}}(a \leq Z \leq b) = 1\right\},$$

we can verify that $Q \in \mathcal{Q} \iff w_{\#}P \in \mathcal{B}_{a,b}$ with $w(x) = (\mathsf{d}Q/\mathsf{d}P)(x)$.

Example 2: f-divergence constrained distribution shift. The f-divergence is a generalized way of measuring the distance between distributions, which includes common metrics such as KL divergence or chi-squared divergence as special cases. For a convex function $f : [0, \infty) \to \mathbb{R}$ satisfying $f(1) = 0$, the f-divergence (Ali and Silvey, 1966; Rényi, 1961) of Q from P is defined as $D_f(Q||P) = \mathbb{E}_P[f((\mathsf{d}Q/\mathsf{d}P)(X))]$. In this example, we consider a constraint set $\mathcal{Q}$ defined via a bound on the f-divergence: $\mathcal{Q} = \{Q : D_f(Q||P) \leq \rho\}$. For instance, if we take $\mathcal{Q} = \mathcal{Q}_{\mathrm{KL}}(\rho) = \{Q : D_{\mathrm{KL}}(Q\|P) \leq \rho\}$, this corresponds to choosing $f(x) = x\log(x)$. Choosing

$$\mathcal{B} = \mathcal{B}_{f,\rho} = \{\widetilde{Q} : \mathbb{E}_{Z\sim\widetilde{Q}}[Z] = 1,\ \mathbb{E}_{Z\sim\widetilde{Q}}[f(Z)] \leq \rho,\ \mathbb{P}_{Z\sim\widetilde{Q}}(Z \geq 0) = 1\},$$

we can verify that $Q \in \mathcal{Q} \iff w_{\#}P \in \mathcal{B}_{f,\rho}$ with $w(x) = (\mathsf{d}Q/\mathsf{d}P)(x)$.

3.2.1 Worst-case excess risk with DRL

In this section, we explore some properties of the generic DRL, without the isotonic constraint. Building this framework will help us to introduce the isotonic constraint as follows.

Based on the equivalence of $\mathcal{Q}$ and $\mathcal{B}$ in representing the uncertainty set, we focus on the following equivalent representation of $\Delta(R; \mathcal{Q})$:

$$\begin{aligned} \Delta(R;\mathcal{B}) \quad = \quad & \sup_{w\geq 0} \quad \mathbb{E}_P\left[w(X)R(X)\right] - \mathbb{E}_P[R(X)] \\ & \text{subject to} \quad w_{\#}P \in \mathcal{B}, \end{aligned} \tag{3.4}$$

where abusing notation we now write $\Delta(\cdot;\mathcal{B})$ to express that $\mathcal{B}$ is a constraint on the distribution of the density ratio $w(X) = (\mathsf{d}Q/\mathsf{d}P)(X)$, where previously we instead wrote $\Delta(\cdot;\mathcal{Q})$. We will say that $\Delta(R;\mathcal{B})$ is attainable if this supremum is attained by some w^* in the constraint set. Throughout the paper, we assume that the set of distributions $\mathcal{B}$ satisfies the

following condition.

Condition 3.2.1. The set $\mathcal{B}$ contains the point mass on the value 1. Moreover, $\mathcal{B}$ is closed under convex ordering, that is, if $Q \in \mathcal{B}$, then for any $Q' \overset{cvx}{\preceq} Q$, it holds that $Q' \in \mathcal{B}$.

This condition enables the following reformulation of the quantity of interest, $\Delta(R; \mathcal{B})$:

Proposition 3.2.2. *Assume Condition 3.2.1 holds. Then* $\Delta(R; \mathcal{B})$ *can be written as*

$$\begin{aligned} \Delta(R; \mathcal{B}) \quad = \quad & \sup_{\phi:\mathbb{R}\to\mathbb{R}_+} \quad \mathbb{E}_P\left[(\phi \circ R)(X) R(X)\right] - \mathbb{E}_P[R(X)] \\ & \text{subject to} \quad (\phi \circ R)_{\#} P \in \mathcal{B}, \qquad \phi \textit{ is nondecreasing.} \end{aligned}$$

Moreover, if $\Delta(R; \mathcal{B})$ is attainable (i.e., the supremum is attained by some w^* satisfying the constraints), then this equivalent formulation is attainable as well (i.e., the supremum is attained by some ϕ^* satisfying the constraints), and it then holds that $w^*(X) = \phi^*(R(X))$ P-almost surely.

In words, this proposition shows that the excess risk is maximized by considering functions $w(x)$ that are monotonically nondecreasing with respect to $R(x)$. This is intuitive, since maximizing the expected value of $w(X)R(X)$ implies that we should choose a function w that is large when R is large. Most importantly, Proposition 3.2.2 implies that for a class of constraint sets $\mathcal{B}$, the optimal value in the constrained optimization problem (3.4) only depends on the distribution of X through the distribution of $R(X)$. We note that, in the special case when $\mathcal{B}$ is specified in terms of an f-divergence (as in Example 2 above), the conclusion of Proposition 3.2.2 is established by Donsker and Varadhan, 1976; Lam, 2016; Namkoong et al., 2022. We verify that this result holds in aforementioned settings as follows.

Returning to Example 1: bound-constrained distribution shift. Recall that in this example, we take the constraint set $\mathcal{B} = \mathcal{B}_{a,b} = \{\widetilde{Q} : \mathbb{E}_{Z\sim\widetilde{Q}}[Z] = 1,\ \mathbb{P}_{Z\sim\widetilde{Q}}(a \le Z \le b) = 1\}$,

for some $0 \le a < 1 < b < +\infty$. It is straightforward to verify that $\mathcal{B}_{a,b}$ satisfies Condition 3.2.1, implying that Proposition 3.2.2 can be applied.

Moreover, we can actually calculate the maximizing density ratio $w^*(x)$ explicitly. The worst-case density ratio that attains the worst-case excess risk takes the form

$$w^*(x) = a \cdot \mathbb{1}\left\{R(x) < q_R\left(\tfrac{b-1}{b-a}\right)\right\} + b \cdot \mathbb{1}\left\{R(x) > q_R\left(\tfrac{b-1}{b-a}\right)\right\} + c \cdot \mathbb{1}\left\{R(x) = q_R\left(\tfrac{b-1}{b-a}\right)\right\},$$

where $q_R(t)$ is the t-quantile of the distribution of $R(X)$ under $X \sim P$ and $c \in [a, b]$ is defined as the unique value ensuring that $\mathbb{E}[w^*(X)] = 1$, namely,

$$c = a + \frac{(b-a)t^* - (b-1)}{\mathbb{P}\left\{R(X) = q_R\left(\tfrac{b-1}{b-a}\right)\right\}} \qquad \text{with} \qquad t^* = \mathbb{P}\left\{R(X) \le q_R\left(\tfrac{b-1}{b-a}\right)\right\} \ge \frac{b-1}{b-a}.$$

We can see that $w^*(x)$ is nondecreasing in $R(x)$, validating the conclusion of Proposition 3.2.2.

Returning to Example 2: f-divergence constrained distribution shift. Recall that for an f-divergence constraint, we define $\mathcal{B} = \mathcal{B}_{f,\rho} = \{\widetilde{Q} : \mathbb{E}_{Z\sim\widetilde{Q}}[Z] = 1,\ \mathbb{E}_{Z\sim\widetilde{Q}}[f(Z)] \le \rho,\ Z \ge 0\}$. Since f is convex, this immediately implies that $\mathcal{B}_{f,\rho}$ satisfies Condition 3.2.1. If we further assume that f is differentiable, by the results of Shapiro (2017); Donsker and Varadhan (1976); Lam (2016), the worst-case excess risk $\Delta_\rho(R; \mathcal{B}_{f,\rho})$ is attained at

$$w^*(x) = w(x; \lambda^*, \nu^*) = \left\{(f')^{-1}\left(\frac{R(x) - \nu^*}{\lambda^*}\right)\right\}_+,$$

where λ^*, ν^* are the solutions to the dual problem

$$\inf_{\lambda \ge 0,\, \nu} \left\{\lambda\rho + \nu + \mathbb{E}_P\Big[w(X; \lambda, \nu)(R(X) - \nu) - \lambda f(w(X; \lambda, \nu))\Big]\right\}.$$

Since f is convex, its inverse derivative $(f')^{-1}$ is nondecreasing, meaning that $w^*(x)$ is nondecreasing in $R(x)$, which again validates the result in Proposition 3.2.2.

3.3 Worst-case excess risk with an isotonic constraint

In this section, we will now formally introduce our iso-DRL method, adding an isotonic constraint to the DRL framework developed in Section 3.2 above. As in Section 3.2, throughout this section we use the notation of the unsupervised learning setting with risk $R(X)$, since the supervised case can also be reduced to this setting.

Recall the cone of isotonic functions $\mathcal{C}^{\text{iso}}_{\preceq} = \{w : \mathcal{X} \to \mathbb{R} : \ w \text{ is isotonic w.r.t. } \preceq\}$. In this paper, we actually allow $\preceq$ to be a partial preorder rather than a partial order, meaning that it may be the case that both $x \preceq x'$ and $x' \preceq x$, even when $x \neq x'$. As an example, we denote $\mathcal{C}^{\text{iso}}_{w_0} = \{w : w(x) \text{ is a monotonically nondecreasing function of } w_0(x)\}$—this is obtained by the (pre)order given by $x \preceq x'$ whenever $w_0(x) \leq w_0(x')$.

Our focus is the worst-case excess risk with the isotonic constraint:

$$\begin{aligned} \Delta^{\text{iso}}(R;\mathcal{B}) \quad = \quad & \sup_{w \geq 0} \quad \mathbb{E}_P\left[w(X)R(X)\right] - \mathbb{E}_P[R(X)] \\ & \text{subject to} \quad w_{\#}P \in \mathcal{B}, \quad w \in \mathcal{C}^{\text{iso}}_{\preceq}. \end{aligned} \tag{3.5}$$

To make this more concrete with a specific example, in the bound (3.3), this example corresponds to choosing $\mathcal{B} = \mathcal{B}_{f,\rho}$ for the f-divergence $f(x) = x \log x$, as for the KL distance constraint. In particular, the bound (3.3) assumed two constraints on the distribution P_{target}—first, $D_{\text{KL}}(P_{\text{target}} \| P) \leq \rho$ (which corresponds to assuming $(\mathsf{d}P_{\text{target}}/\mathsf{d}P)_{\#}P \in \mathcal{B}_{f,\rho}$, in our new notation), and second, $P_{\text{target}} \in \mathcal{Q}_{\text{iso}}(w_0)$ (which is expressed by assuming $w \in \mathcal{C}^{\text{iso}}_{\preceq}$ when we take the partial (pre)order defined as $x \preceq x'$ whenever $w_0(x) \leq w_0(x')$—or equivalently, we can write this as $w \in \mathcal{C}^{\text{iso}}_{w_0}$).

3.3.1 Equivalent formulation

Optimization problems with isotonic constraints may be difficult to tackle both theoretically and computationally, since the isotonic cone, despite being convex, may be challenging to optimize over when working with an infinite-dimensional object such as the density ratio. In this section, we will show that the maximization problem (3.5) can equivalently be reformulated as an optimization problem without an isotonic constraint, by drawing a connection to the original (not isotonic) DRL maximization problem (3.4).

Given the probability measure P, we will define π as the projection to the isotonic cone $\mathcal{C}_{\preceq}^{\text{iso}}$ with respect to $L_2(P)$, i.e., $\pi(a) = \operatorname{argmin}_{b \in \mathcal{C}_{\preceq}^{\text{iso}}} \int (a(x) - b(x))^2 \mathsf{d}P(x)$. As $L_2(P)$ is reflexive and strictly convex, the projection $\pi(a)$ exists and is unique (up to sets of measure zero) for all $a \in L_2(P)$ (Megginson, 2012). Then, with the projection π in place, we are ready to state our main equivalence result.

Theorem 3.3.1. *For any $\mathcal{B}$ and any partial (pre)order $\preceq$ on $\mathcal{X}$, it holds that $\Delta^{\text{iso}}(R; \mathcal{B}) \leq \Delta(\pi(R); \mathcal{B})$. If in addition Condition 3.2.1 holds, then we have*

$$\Delta^{\text{iso}}(R; \mathcal{B}) = \Delta(\pi(R); \mathcal{B}),$$

and moreover, $\Delta^{\text{iso}}(R; \mathcal{B})$ is attainable if and only if $\Delta(\pi(R); \mathcal{B})$ is attainable.

To interpret this theorem, recall from the definition (3.4) that we have

$$\begin{aligned} \Delta(\pi(R); \mathcal{B}) \quad = \quad & \sup_{w \geq 0} \quad \mathbb{E}_P\big[w(X)[\pi(R)](X)\big] - \mathbb{E}_P\big[[\pi(R)](X)\big] \\ & \text{subject to} \quad w_{\#}P \in \mathcal{B}. \end{aligned} \tag{3.6}$$

Compared with the formulation (3.5) that defines the isotonic worst-case risk $\Delta^{\text{iso}}(R; \mathcal{B})$, we see that this equivalent formulation removes the constraint $w \in \mathcal{C}_{\preceq}^{\text{iso}}$ by replacing R with

its isotonic projection $\pi(R)$. This brings computational benefits. The equivalent formulation (3.6) separates two constraints $w_{\#}P \in \mathcal{B}$ and $w \in \mathcal{C}^{\text{iso}}_{\preceq}$, allowing us to first project the risk function R onto $\mathcal{C}^{\text{iso}}_{\preceq}$, and then solve a problem that is as simple as the problem stated earlier in (3.4). More concretely, as seen in Examples 1 and 2, for many common choices of $\mathcal{B}$, we have closed-form solutions to (3.6) in terms of the projected risk $\pi(R)$.

3.3.2 Setting: iso-DRL with estimated density ratio

We now return to the scenario described in (3.3) in Section 3.1.3, where we would like to recalibrate a pretrained density ratio w_0 that estimates $\mathsf{d}P_{\text{target}}/\mathsf{d}P$. As the shape or relative magnitude of w_0 could contain useful information about the true density ratio, we consider candidate distributions with the density ratio as an isotonic function of w_0, which is equivalent to considering the partial (pre)order $x \preceq x' \iff w_0(x) \leq w_0(x')$. We will denote the specific isotonic cone under this partial (pre)order as $\mathcal{C}^{\text{iso}}_{w_0}$ and its isotonic projection as π_{w_0}, and abusing notation, we write $\Delta^{\text{iso}}(R; \mathcal{B}, w_0)$ to denote the excess risk for this particular setting, to emphasize the role of w_0. Under the condition of Theorem 3.3.1, we have the equivalence $\Delta^{\text{iso}}(R; \mathcal{B}, w_0) = \Delta(\pi_{w_0}(R); \mathcal{B})$. To understand the projection onto the cone $\mathcal{C}^{\text{iso}}_{w_0}$ in a more straightforward way, we can derive a further simplification. Write π_1 to denote the isotonic projection of functions $\mathbb{R} \to \mathbb{R}$ under the measure $(w_0)_{\#}P$, and define a function $\widetilde{R} : \mathbb{R} \to \mathbb{R}$ to satisfy $\widetilde{R}(w_0(X)) = \mathbb{E}_P[R(X) \mid w_0(X)]$ P-almost surely. We then have the following simplified equivalence:

Proposition 3.3.2. *Assume Condition 3.2.1 holds. We have the equivalence* $\Delta^{\text{iso}}(R; \mathcal{B}, w_0) = \Delta(\pi_1(\widetilde{R}) \circ w_0; \mathcal{B}, w_0)$*, where we define*

$$\begin{aligned} \Delta(R; \mathcal{B}, w_0) \quad = \quad & \sup_{h:\ h \circ w_0 \geq 0} \quad \mathbb{E}_P\left[(h \circ w_0)(X) R(X)\right] - \mathbb{E}_P[R(X)] \\ & \text{subject to} \quad (h \circ w_0)_{\#} P \in \mathcal{B}. \end{aligned} \tag{3.7}$$

In comparison to the equivalence $\Delta^{\text{iso}}(R; \mathcal{B}, w_0) = \Delta(\pi_{w_0}(R); \mathcal{B})$, the equivalence in the proposition relies on an isotonic projection with respect to the canonical order on the real line (i.e., the projection π_1).

Moreover, when the true distribution shift does not obey the isotonic constraint exactly, in Appendix Section B.2.4, we can nonetheless provide a bound on the worst-case excess risk, which is tighter than the (non-iso) DRL bound whenever the isotonic constraint provides a reasonable approximation.

3.4 Estimation of worst-case excess risk with isotonic constraint

So far, our focus has been on the population level problem, namely, we have assumed full access to the data distribution P and the risk function R. In practice, however, we may only be able to access the data distribution P via samples, and we may only be able to learn about the risk function R via noisy evaluations of $R(X)$ on each sampled point X in the unsupervised setting. Or, in the supervised setting, we can only access $\widetilde{P}$ via samples of labeled data points drawn from this distribution, and can learn about r only through evaluating $r(X, Y)$ on these sampled data points.

In this section, we propose a fully data dependent estimator for the worst-case excess risk $\Delta^{\text{iso}}(R; \mathcal{B})$. Moreover, we characterize the estimation error for different choices of $\mathcal{B}$, including the bounds constraint and the f-divergence constraint for the distribution shift.

3.4.1 Plug-in estimators

We start with presenting the plug-in estimators of the worst-case excess risk under the isotonic constraint in both the unsupervised and supervised settings.

The unsupervised setting. We have n i.i.d. observations $\{X_i\}_{i \le n}$ from a distribution P. Given a risk function $R : \mathcal{X} \to \mathbb{R}_+$, and the uncertainty set $\mathcal{B}$, we estimate the worst-case

excess risk $\Delta^{\text{iso}}(R;\mathcal{B})$ (cf. Equation (3.5)) via

$$\begin{aligned}\widehat{\Delta}^{\text{iso}}(R;\mathcal{B}) \quad &:= \quad \max_{w\geq 0} \qquad \frac{1}{n}\sum_{i\leq n} w(X_i)R(X_i) - \frac{1}{n}\sum_{i\leq n} R(X_i) \\ &\qquad\quad \text{subject to} \qquad w_{\#}\widehat{P}_n \in \mathcal{B}, \quad w \in \mathcal{C}^{\text{iso}}_{\preceq}. \end{aligned} \tag{3.8}$$

Here, $\widehat{P}_n$ denotes the empirical distribution of the sample $\{X_i\}_{i\leq n}$ drawn from P.

The supervised setting. In this case, we have $\{(X_i, Y_i)\}_{i\leq n}$ drawn i.i.d. from $\widetilde{P} = P \times P_{Y|X}$. Given a risk function r, and the uncertainty set $\mathcal{B}$, we propose to estimate the worst-case excess risk $\widehat{\Delta}^{\text{iso}}(r;\mathcal{B})$ by replacing $R(X_i)$ with $r(X_i, Y_i)$ in (3.8).

Adding a boundedness constraint

When calculating the excess risk at the population level, the constraint set $\mathcal{B}$ may not require w to be bounded—specifically, while $\mathcal{B}_{a,b}$ imposes an upper bound on w, the f-divergence constraint set $\mathcal{B}_{f,\rho}$ does not. In the empirical setting, however, a boundedness constraint is more crucial: we want to avoid degenerate scenarios, such as $w(X_i)$ taking an arbitrarily large value for a single i, and being zero for the remaining $n-1$ data points. To this end, we will assume from this point on that $\mathcal{B}$ includes a boundedness constraint:

Condition 3.4.1. There exists Ω such that any distribution $Q \in \mathcal{B}$ is supported on $[0, \Omega]$.

This is trivially true for $\mathcal{B} = \mathcal{B}_{a,b}$ with $\Omega = b$. But this constraint actually allows us to work with the f-divergence example, as well, as established by the following result.

Proposition 3.4.2. *Assume the convex function f is differentiable on $\mathbb{R}_+$. The worst-case excess risk $\Delta^{\text{iso}}(R;\mathcal{B}_{f,\rho})$ is attained at some $w^{*\text{iso}}_{f,\rho} \in \mathcal{C}^{\text{iso}}_{\preceq}$ with $\|w^{*\text{iso}}_{f,\rho}\|_\infty < \infty$.*

In particular, defining $\mathcal{B}_{f,\rho,\Omega} = \{\widetilde{Q} : \ \mathbb{E}_{Z\sim\widetilde{Q}}[Z] = 1, \ \mathbb{E}_{Z\sim\widetilde{Q}}[f(Z)] \leq \rho, \ \mathbb{P}_{Z\sim\widetilde{Q}}(0 \leq Z \leq \Omega) = 1\}$, which adds a boundedness requirement in addition to the f-divergence constraint,

we can see that for sufficiently large Ω (namely, $\Omega \geq \|w^{*\mathrm{iso}}_{f,\rho}\|_\infty$), even though $\mathcal{B}_{f,\rho,\Omega} \subsetneq \mathcal{B}_{f,\rho}$, it nonetheless holds that $\Delta^{\mathrm{iso}}(R; \mathcal{B}_{f,\rho,\Omega}) = \Delta^{\mathrm{iso}}(R; \mathcal{B}_{f,\rho})$. Therefore, by working with the constraint set $\mathcal{B}_{f,\rho,\Omega}$, we are estimating the same excess risk, but Condition 3.4.1 nonetheless holds. (Of course, in practice, the value of $\|w^{*\mathrm{iso}}_{f,\rho}\|_\infty$ is unknown and so we can simply set Ω to be a large constant.)

3.4.2 Computation: estimation after projection

Before moving onto the statistical performance of the two estimators $\widehat{\Delta}^{\mathrm{iso}}(R; \mathcal{B})$ and $\widehat{\Delta}^{\mathrm{iso}}(r; \mathcal{B})$, we pause to discuss fast computational methods for these. The key is Theorem 3.3.1—we may accelerate the computation of both estimators via an equivalent optimization problem without the isotonic constraint.

To be more specific, we begin by considering the supervised setting. Denote $r^{\mathrm{iso}} = (r^{\mathrm{iso}}_i)_{i \leq n} \in \mathbb{R}^n$ as the isotonic projection of $(r(X_i, Y_i))_{i \leq n}$ with respect to the empirical distribution $\widehat{P}_n$ under the partial order $\preceq$. Then, consider the optimization problem

$$\begin{aligned} \widehat{\Delta}(r^{\mathrm{iso}}; \mathcal{B}) \quad := \quad & \max_{w \geq 0} \quad \frac{1}{n} \sum_{i \leq n} w(X_i) r^{\mathrm{iso}}_i - \frac{1}{n} \sum_{i \leq n} r^{\mathrm{iso}}_i \\ & \text{subject to} \quad w_\# \widehat{P}_n \in \mathcal{B}. \end{aligned}$$

By Theorem 3.3.1 (applied with $\widehat{P}_n$ in place of P), we have $\widehat{\Delta}(r^{\mathrm{iso}}; \mathcal{B}) = \widehat{\Delta}^{\mathrm{iso}}(r; \mathcal{B})$. Analogously, in the unsupervised setting, we instead have $\widehat{\Delta}^{\mathrm{iso}}(R; \mathcal{B}) = \widehat{\Delta}(R^{\mathrm{iso}}; \mathcal{B})$, where $R^{\mathrm{iso}} = (R^{\mathrm{iso}}_i)_{i \leq n} \in \mathbb{R}^n$ as the isotonic projection of $(R(X_i))_{i \leq n}$ with respect to the empirical distribution $\widehat{P}_n$ under the partial order $\preceq$.

Note that in iso-DRL with estimated density ratio in Section 3.3.2, we can simply apply the isotonic regression for $(r(X_i, Y_i))_{i \leq n}$ on $(w_0(X_i))_{i \leq n}$ to obtain the projected risk. We can now see concretely that this equivalence allows for a much more efficient calculation. For example, in the case $\mathcal{X} = \mathbb{R}$, this isotonic projection can be computed in $\mathcal{O}(n)$ time

(e.g., via the PAVA algorithm, which provides an exact calculation of isotonic projection in $\mathbb{R}^n$ (Grotzinger and Witzgall, 1984)), this leads to a very simple implementation for computing $\widehat{\Delta}(r^{\text{iso}}; \mathcal{B})$—in particular, once the vector r^{iso} has been computed, the remaining optimization problem is simple since there is no remaining isotonic constraint.

3.4.3 Performance guarantees for plug-in estimators

In this section, we present the performance guarantees for plug-in estimators for a general constraint set $\mathcal{B}$. To jump ahead to the conclusion, we will see that our results imply the following consistency properties for the settings $\mathcal{B} = \mathcal{B}_{a,b}$ and $\mathcal{B} = \mathcal{B}_{f,\rho,\Omega}$. Here we define the Rademacher complexity of a function class $\mathcal{G}$ by $\mathcal{R}_n(\mathcal{G}) = \mathbb{E}[\sup_{g \in \mathcal{G}} |n^{-1} \sum_{i \leq n} \sigma_i g(Z_i)|]$, where $\{Z_i\}_{i \leq n}$ is a sample of size n from P and $\{\sigma_i\}_{i \leq n}$ are independent random variables drawn uniformly from $\{+1, -1\}$.

Proposition 3.4.3 (Informal result for examples)**.** *For both $\mathcal{B} = \mathcal{B}_{a,b}$ and $\mathcal{B} = \mathcal{B}_{f,\rho,\Omega}$, and for both supervised ($\widehat{\Delta}^{\text{iso}}(\mathcal{B}) = \widehat{\Delta}^{\text{iso}}(r; \mathcal{B})$) and unsupervised ($\widehat{\Delta}^{\text{iso}}(\mathcal{B}) = \widehat{\Delta}^{\text{iso}}(R; \mathcal{B})$) learning, and under some mild additional conditions specified below, it holds with probability $\geq 1 - 3n^{-1}$ that*

$$\left| \widehat{\Delta}^{\text{iso}}(\mathcal{B}) - \Delta^{\text{iso}}(R; \mathcal{B}) \right| \leq C \left(\mathcal{R}_n(\mathcal{C}^{\text{iso}}_{\preceq,\Omega}) + \sqrt{\frac{\log n}{n}} \right),$$

where $\mathcal{C}^{\text{iso}}_{\preceq,\Omega} = \{w \in \mathcal{C}^{\text{iso}}_{\preceq} : \ 0 \leq w \leq \Omega\}$ is the bounded isotonic cone, where the constant C will be defined in the theorems below, and where we set $\Omega = b$ for the case $\mathcal{B} = \mathcal{B}_{a,b}$.

Our bounds rely on the Rademacher complexity term $\mathcal{R}_n(\mathcal{C}^{\text{iso}}_{\preceq,\Omega})$, which will naturally depend on the properties of the (pre)ordering $\preceq$ that defines this isotonic cone. To provide further intuition, we now give two concrete examples to make apparent the dependence of the Rademacher complexity on the sample size.

(1) When $d = 1$, e.g., in the setting of density ratio recalibration in Section 3.3.2, similar to the results of Chatterjee and Lafferty (2019), one can show by Dudley's theorem (Dudley, 1967) that $\mathcal{R}_n\left(\mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}\right) \lesssim n^{-1/2}$ up to logarithmic factors.

(2) For $\mathbb{R}^d$ with a fixed dimension $d \geq 2$ and a bounded domain $\mathcal{X}$ equipped with the componentwise order (Han et al., 2019; Deng and Zhang, 2020; Gao and Wellner, 2007), i.e., $x \preceq z$ if and only if $x_j \leq z_j$ for all $j \in [d]$, by Han et al. (2019), if the density $\mathsf{d}P(x)$ is bounded below (away from zero) and above, then we have $\mathcal{R}_n\left(\mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}\right) \lesssim n^{-1/d}$ up to logarithmic factors.

Formal results

Now we turn to developing these results formally, in a general framework. We will begin with a deterministic result, which shows that, if certain concentration inequalities hold, then $\widehat{\Delta}^{\mathrm{iso}}(\mathcal{B}_{a,b})$ is an accurate estimate of $\Delta^{\mathrm{iso}}(R; \mathcal{B}_{a,b})$. Then we will show that the concentration results hold with high probability, in both of our two settings, $\mathcal{B} = \mathcal{B}_{a,b}$ and $\mathcal{B} = \mathcal{B}_{f,\rho,\Omega}$.

We first need a few definitions. For any distributions P_0, P_1, if $w_{\#}P_0 \in \mathcal{B}$, we define

$$\varepsilon_{\mathcal{B}}\left(w; P_0, P_1\right) = \inf\left\{s \geq 0 \ : \ \exists\, t \geq 0,\ \left((1-s)\cdot w + t \cdot \mathbf{1}\right)_{\#} P_1 \in \mathcal{B}\right\}.$$

In other words, if weight function w satisfies the constraints relative to distribution P_0, we need to find constants s, t such that the modified weight function $(1-s)\cdot w + t \cdot \mathbf{1}$ satisfies the constraints relative to distribution P_1. (Note that we must have $\varepsilon_{\mathcal{B}}\left(w; P_0, P_1\right) \leq 1$, since choosing $s = t = 1$ will always be feasible, because $\mathbf{1}_{\#}P_1$ is the point mass on the value 1, and therefore satisfies the constraints of $\mathcal{B}$, by assumption.) Of importance is the quantity

$$\varepsilon_{\mathcal{B}} = \sup_{w \in \mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}} \max\left\{\varepsilon_{\mathcal{B}}\left(w; P, \widehat{P}_n\right), \varepsilon_{\mathcal{B}}\left(w; \widehat{P}_n, P\right)\right\},$$

which characterizes the feasibility gap between the population and sample problems. In addition, we define

$$\varepsilon_R = \sup_{w \in \mathcal{C}^{\text{iso}}_{\preceq,\Omega}} \left| \mathbb{E}_{\widehat{P}_n}[(w(X)-1)r(X,Y)] - \mathbb{E}_P[(w(X)-1)R(X)] \right|$$

in the case of unsupervised learning, or

$$\varepsilon_R = \sup_{w \in \mathcal{C}^{\text{iso}}_{\preceq,\Omega}} \left| \mathbb{E}_{\widehat{P}_n}[(w(X)-1)R(X)] - \mathbb{E}_P[(w(X)-1)R(X)] \right|$$

in the case of supervised learning. The value of ε_R measures the concentration between the empirical risk and the population one. With these definitions in place, we are ready to state the generic performance guarantee of the plug-in estimators.

Theorem 3.4.4. *Suppose that the risk is B_R-bounded (i.e., R or r, in the unsupervised or supervised case, respectively), and that the constraint set $\mathcal{B}$ satisfies Condition 3.4.1. Then, it holds for both supervised ($\widehat{\Delta}^{\text{iso}}(\mathcal{B}) = \widehat{\Delta}^{\text{iso}}(r;\mathcal{B})$) and unsupervised ($\widehat{\Delta}^{\text{iso}}(\mathcal{B}) = \widehat{\Delta}^{\text{iso}}(R;\mathcal{B})$) learning that*

$$\left| \widehat{\Delta}^{\text{iso}}(\mathcal{B}) - \Delta^{\text{iso}}(R;\mathcal{B}) \right| \leq \varepsilon_R + 2B_R\Omega \cdot \varepsilon_{\mathcal{B}}.$$

Of course, in order for this result to be meaningful, we need to ensure that ε_R and $\varepsilon_{\mathcal{B}}$ are likely to be small, with high probability, We now turn to the question of establishing such concentration results. First we bound ε_R.

Lemma 3.4.5. *Suppose that the risk is B_R-bounded (i.e., R or r, in the unsupervised or supervised case, respectively). Then, with probability at least $1 - n^{-1}$, it holds that $\varepsilon_R \leq 4B_R\mathcal{R}_n(\mathcal{C}^{\text{iso}}_{\preceq,\Omega}) + B_R\Omega\sqrt{\log n/(2n)}$.*

Next we turn to bounding $\varepsilon_{\mathcal{B}}$, which we will do separately for our two examples.

Lemma 3.4.6. *Let $\mathcal{B} = \mathcal{B}_{a,b}$, where $a < 1 < b$. Then, with probability at least $1 - n^{-1}$, it holds that $\varepsilon_{\mathcal{B}} \leq C(\mathcal{R}_n(\mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}) + \Omega\sqrt{\log n/(2n)})$, where we take $\Omega = b$ and C depends only on a, b.*

Finally, to complete this section, we turn to the f-divergence constraint, $\mathcal{B}_{f,\rho}$.

Lemma 3.4.7. *Let $\mathcal{B} = \mathcal{B}_{f,\rho,\Omega}$, where we take any $\Omega \geq \|w^{*\mathrm{iso}}_{f,\rho}\|_\infty$ for $w^{*\mathrm{iso}}_{f,\rho}$ defined as in Proposition 3.4.2. Assume also that f is L_Ω-Lipschitz on $[0, \Omega]$. Then, with probability at least $1 - 2n^{-1}$, it holds that $\varepsilon_{\mathcal{B}} \leq C(\mathcal{R}_n(\mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}) + \sqrt{\log n/(2n)})$, where C depends only on Ω, L_Ω, and ρ.*

3.4.4 The role of the isotonic constraint

The consistency bounds developed above show that, under appropriate conditions, the error in estimating $\Delta^{\mathrm{iso}}(R; \mathcal{B})$ can be controlled whenever the appropriate Rademacher complexity terms are small. In the Appendix Section B.3.6, we will construct an example with $\mathcal{B} = \mathcal{B}_{a,b}$ such that, without an isotonic constraint, $\Delta(R; \mathcal{B}_{a,b}) = 0$ but $\widehat{\Delta}(r; \mathcal{B}_{a,b}) > c$ with high probability, where $c > 0$ doesn't vanish with n, i.e., this empirical estimate is not a consistent estimator of the true excess risk. This suggests that the isotonic constraint plays an important role: essentially, the isotonic constraint induces a form of regularization, ensuring that we work with a low-complexity class of functions.

3.5 Numerical experiments

In this section, we demonstrate the benefits of iso-DRL in calibrating prediction sets under covariate shift with empirical examples, as previewed in Section 3.1.3. Throughout all experiments, we have a training data set $\mathcal{D}_{\mathrm{train}}$ containing data points (X_i, Y_i) drawn from the data distribution $\widetilde{P}$, and a test set $\mathcal{D}_{\mathrm{test}}$ containing data points $(\widetilde{X}_i, \widetilde{Y}_i)$ drawn from $\widetilde{P}_{\mathrm{target}}$.

We consider both synthetic and real datasets. Code to reproduce all experiments is available at `https://github.com/yugjerry/iso-DRL`.

Background. When covariate shift is present, Tibshirani et al. (2019) proposes the weighted conformal prediction (WCP) method, which produces a prediction set $C_{1-\alpha}^{w_0}(X)$ with an estimated density ratio w_0, which is only valid for the shifted covariate distribution $\widehat{P}$ defined by $\mathtt{d}\widehat{P} \propto w_0 \cdot \mathtt{d}P$. The validity for the target distribution $\widetilde{P}_{\text{target}}$ is only guaranteed up to a coverage gap due to the estimation error or potential misspecification in w_0 (Lei and Candès, 2020; Candès et al., 2023; Gui et al., 2024a, 2023b).

Dataset partition. The datasets $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ are partitioned as follows: (1) first, we use a subset $\mathcal{D}_1 \subseteq \mathcal{D}_{\text{train}}$ of the training data of size $|\mathcal{D}_1| = n_{\text{pre}}$, and a subset $\mathcal{D}_{\text{test},1} \subseteq \mathcal{D}_{\text{test}}$ of the test data of size $|\mathcal{D}_{\text{test},1}| = n_{\text{pre}}$, to train the function w_0; (2) next, we use a subset $\mathcal{D}_2 \subseteq \mathcal{D}_{\text{train}} \backslash \mathcal{D}_1$ of the training data of size $|\mathcal{D}_2| = n_{\text{train}}$ to train CP or WCP prediction intervals; (3) then, $\mathcal{D}_3 = \mathcal{D}_{\text{train}} \backslash (\mathcal{D}_1 \cup \mathcal{D}_2)$ is used to for estimating upper bounds on the excess risk for the DRL and iso-DRL methods. We further define $n = |\mathcal{D}_3|$ to ease notations; (4) finally, $\mathcal{D}_{\text{test},0} = \mathcal{D}_{\text{test}} \backslash \mathcal{D}_{\text{test},1}$ is used for estimating the actual performance of each method relative to the target distribution: defining $n_{\text{test}} = |\mathcal{D}_{\text{test},0}|$, we compute $\mathit{Coverage\ rate}(C, \alpha) = n_{\text{test}}^{-1} \sum_{i \in \mathcal{D}_{\text{test},0}} \mathbb{1}\{\widetilde{Y}_i \in C(\widetilde{X}_i)\}$. We next turn to the details of how these steps are carried out.

Initial density ratio estimation. Using data from $\mathcal{D}_1$ and $\mathcal{D}_{\text{test},1}$, we construct a data set comprised of the covariate X and a binary label $L \in \{0, 1\}$ (0 for the training points, 1 for the test points). We then fit a logistic regression model and obtain the estimated probability $\widehat{p}(x)$ for $\mathbb{P}(L = 1 \mid X = x)$, with which we define $w_0(x) = \widehat{p}(x)/(1 - \widehat{p}(x))$.

Split conformal prediction and weighted split conformal prediction. With data from $\mathcal{D}_2$, we use Ordinary Least Squares (OLS) as the base algorithm, where we denote $\widehat{\mu}$

as the fitted regression model, and apply split conformal prediction with the nonconformity score $V(x,y) = |y - \widehat{\mu}(x)|$ to obtain the following prediction intervals for comparison: (1) CP: conformal prediction interval $C_{1-\alpha}$ without adjusting for covariate shift; (2) WCP-oracle: weighted conformal prediction interval $C_{1-\alpha}^{w^*}$ with true density ratio $w^* = \mathsf{d}P_{\text{target}}/\mathsf{d}P$; (3) WCP: weighted conformal prediction interval $C_{1-\alpha}^{w_0}$ with estimated density ratio w_0.

DRL methods: estimation of worst-case excess risks. We then consider two distributionally robust methods. Using the subset $\mathcal{D}_3$ of the training data, the observed risks can be calculated by $r_i = \mathbb{1}\{Y_i \notin C_{1-\alpha}(X_i)\}$, $i \in \mathcal{D}_3$. We adopt the KL divergence constraint $D_{\text{KL}}(Q\|P) \leq \rho$ to measure the magnitude of distribution shift. Then, we obtain

$$\widehat{\Delta}(\alpha) = \max_{\|w\|_\infty \leq \Omega} \quad \frac{1}{n}\sum_{i\in\mathcal{D}_3}(w_i - 1)r_i \quad \text{s.t.} \quad \frac{1}{n}\sum_{i\in\mathcal{D}_3} w_i = 1,\ \frac{1}{n}\sum_{i\in\mathcal{D}_3} w_i \log w_i \leq \rho,$$

with the upper bound set as $\Omega = 100$ throughout the experiments. Next, given the estimated density ratio w_0, we run isotonic regression for $(r_i)_{i\leq n}$ on $(w_0(X_i^{(3)}))_{i\leq n}$ to obtain the projected risk $(r_i^{\text{iso}})_{i\in\mathcal{D}_3}$, with which we can calculate the worst-case excess risk

$$\widehat{\Delta}^{\text{iso}}(\alpha) = \max_{\|w\|_\infty \leq \Omega} \quad \frac{1}{n}\sum_{i\in\mathcal{D}_3}(w_i - 1)r_i^{\text{iso}} \quad \text{s.t.} \quad \frac{1}{n}\sum_{i\in\mathcal{D}_3} w_i = 1,\ \frac{1}{n}\sum_{i\in\mathcal{D}_3} w_i \log w_i \leq \rho.$$

Given these estimates of the worst-case excess risks, we compare the following methods: (1) DRL: CP interval $C_{1-\widetilde{\alpha}}$, where $\widetilde{\alpha} = \max\{0, \alpha - \widehat{\Delta}(\alpha)\}$.[4]; (2) iso-DRL-$w_0$: CP interval $C_{1-\alpha_{\text{iso}}}$, where $\alpha_{\text{iso}} = \max\{0, \alpha - \widehat{\Delta}^{\text{iso}}(\alpha)\}$.

4. To explain this construction, recall from Section 3.1 that we can use the excess risk estimate to choose a tuning parameter that achieves a desired bound on risk. Specifically, for any value of $\widetilde{\alpha}$, we can bound the risk (i.e., the miscoverage) for the CP interval $C_{1-\widetilde{\alpha}}$ as $\mathbb{E}_{\widetilde{P}_{\text{target}}}[Y \notin C_{1-\widetilde{\alpha}}(X)] \leq \mathbb{E}_{\widetilde{P}}[Y \notin C_{1-\widetilde{\alpha}}(X)] + \Delta(R_{\widetilde{\alpha}}; \mathcal{B}_{f,\rho}) \leq \widetilde{\alpha} + \Delta(R_{\widetilde{\alpha}}; \mathcal{B}_{f,\rho})$ (where $R_{\widetilde{\alpha}}$ is the risk defined by the CP interval $C_{1-\widetilde{\alpha}}$, for any value of $\widetilde{\alpha}$). Since $a \mapsto R_a$ is nondecreasing, this also implies that $a \mapsto \Delta(R_a; \mathcal{B}_{f,\rho})$ is nondecreasing (recall from Section 3.2.1 that $\Delta(R; \mathcal{B})$ is monotone in R, as a corollary of Proposition 3.2.2). Thus, for $\widetilde{\alpha} \leq \alpha$ we have $\mathbb{E}_{\widetilde{P}_{\text{target}}}[Y \notin C_{1-\widetilde{\alpha}}(X)] \leq \widetilde{\alpha} + \Delta(R_\alpha; \mathcal{B}_{f,\rho}) \approx \widetilde{\alpha} + \widehat{\Delta}(\alpha)$. Consequently, the above choice of $\widetilde{\alpha}$ ensures that miscoverage will be (approximately) bounded by α. A similar argument also holds for iso-DRL-w_0.

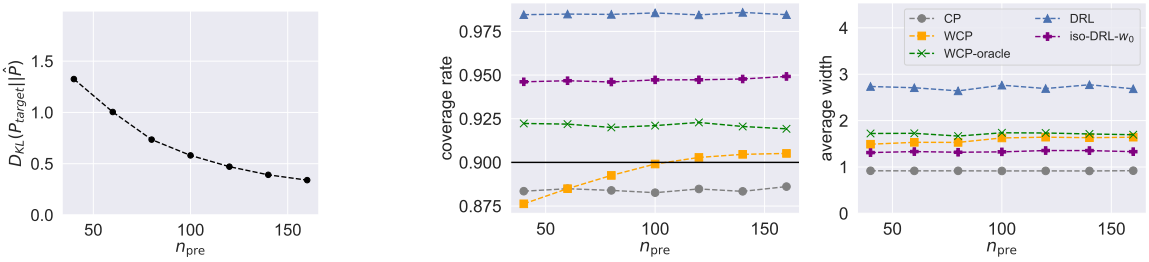
3.5.1 Synthetic dataset

We start with a synthetic example, in which we fix $n_{\text{train}} = n = n_{\text{test}} = 500$ and will vary n_{pre} to see how will the initial density ratio estimation w_0 affect the result. We will consider two settings—the "well-specified" and "misspecified" settings. Specifically, for the marginal distributions of X, we set the well-specified setting with $P : X \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ and $P_{\text{target}} : X \sim \mathcal{N}(\mu, \mathbf{I}_d)$, and the misspecified setting with $P : X \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ and $P_{\text{target}} : X \sim \mathcal{N}(\mu, \mathbf{I}_d + \frac{\zeta}{d}\mathbf{1}_d\mathbf{1}_d^\top)$, where $d = 20$, $\mu = (2/\sqrt{d}) \cdot (1, \cdots, 1)^\top$, and $\zeta = 6$. Since the estimate w_0 for the density ratio will be fitted via logistic regression as described above, the first setting is indeed well-specified since, due to the fact that P and P_{target} have the same covariance, the logistic model is correct for the distribution shift from P to P_{target}. In contrast, the second setting is misspecified since, due to the change in covariance matrix, the underlying log-density ratio is no longer a linear function of $\mu^\top X$, and therefore cannot be characterized exactly by a logistic regression model. Finally, for the conditional distribution of $Y \mid X$, we set $Y \mid X \sim 0.2 \cdot \mathcal{N}(X^\top\beta + \sin(X_1) + 0.4X_3^3 + 0.2X_4^2, 1)$ for both training and target distributions, where $\beta \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$.

Results with varying sample size n_{pre} for estimating w_0

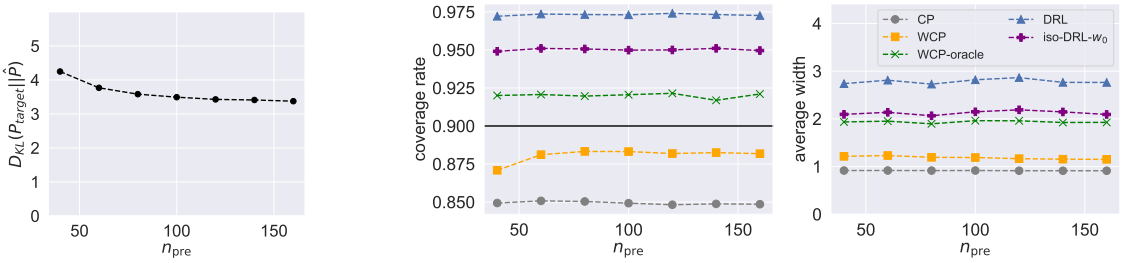
We first consider the scenario with an estimated density ratio w_0. Recall that we use the subsets $\mathcal{D}_1 \subset \mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test},1} \subset \mathcal{D}_{\text{test}}$ with $|\mathcal{D}_1| = |\mathcal{D}_{\text{test},1}| = n_{\text{pre}}$ for estimating w_0; consequently, for larger values of n_{pre}, we will expect a more accurate w_0. By varying n_{pre} in $\{40, 60, 80, 100, 120, 140, 160\}$, we aim to investigate the robustness of WCP and iso-DRL with respect to the accuracy in w_0, where we fix $\rho = \rho^* \coloneqq D_{\text{KL}}(P_{\text{target}} \| P)$.

Well-specified setting. In Figure 3.2b, where the solid horizontal line (in the middle plot) marks the nominal coverage level, $1 - \alpha = 90\%$, we can see that the uncorrected CP exhibits undercoverage due to the mismatch between P_{target} and P, while the coverage of



(a) $D_{\mathrm{KL}}(P_{\mathrm{target}}\|\widehat{P})$ vs. n_{pre}. (b) Comparison of all methods with varying n_{pre}.

Figure 3.2: Results in the well-specified setting.

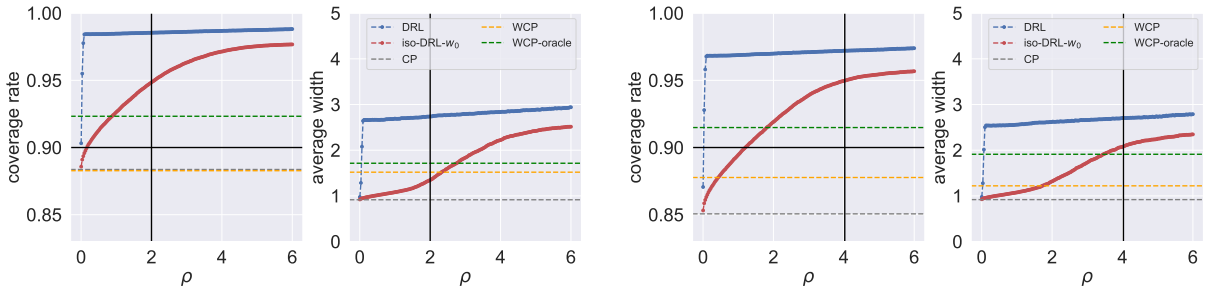


(a) $D_{\mathrm{KL}}(P_{\mathrm{target}}\|\widehat{P})$ vs. n_{pre}. (b) Comparison of all methods with varying n_{pre}.

Figure 3.3: Results in the misspecified setting.

WCP using w_0 increases to 90% as n_{pre} increases, since w_0 becomes more accurate with larger n_{pre} (cf. Figure 3.2a). The generic DRL, even with $\rho = \rho^*$, tends to be conservative and has the widest interval. In comparison, iso-DRL-w_0 has coverage very close to the target level.

Misspecified w_0. In Figure 3.3b, we show results for the misspecified setting. Since w_0 is estimated from a model class that does not contain the true density ratio, consequently $D_{\mathrm{KL}}(P_{\mathrm{target}}\|\widehat{P})$ does not converge to zero as n_{pre} increases (cf. Figure 3.3a). As a result, both uncorrected CP and WCP (which is weighted with the misspecified w_0) exhibit undercoverage. The proposed iso-DRL-w_0 method has coverage slightly above 90% but has interval width close to that of WCP-oracle, while DRL is overly conservative.



(a) Results in the well-specified setting.

(b) Results in the misspecified setting.

Figure 3.4: Results with varying ρ.

Results with varying ρ

In this section, we investigate the sensitivity of each approach (DRL and iso-DRL-w_0) to the choice of ρ. Fixing $n_{\text{pre}} = 50$, we vary ρ in $[0.002, 6]$. The solid vertical line in each plot denotes the true KL divergence, $\rho^* = D_{\text{KL}}(P_{\text{target}} \| P)$. Other methods that do not depend on ρ behave in the same way as shown in the previous section.

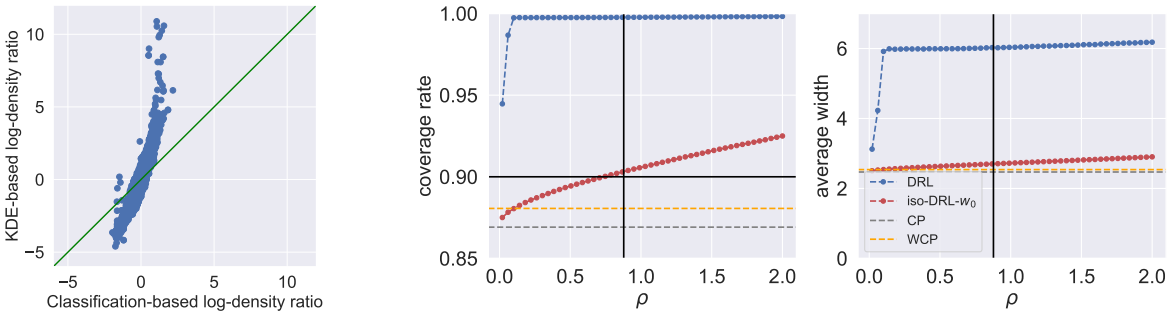
We can see from both plots that the prediction intervals produced by DRL are quite conservative and much wider than the oracle interval across nearly the entire range of ρ, even values ρ much smaller than the true distribution shift magnitude $\rho^* = D_{\text{KL}}(P_{\text{target}} \| P)$. In comparison, for iso-DRL-w_0, when $\rho = \rho^*$, the width of intervals is comparable to the oracle interval in both cases, and the coverage and width vary slowly as we change the value of ρ. From this we can see that the isotonic constraint offers a significant gain in accuracy if we have a reasonable estimate of ρ^*.

3.5.2 Real data: wine quality dataset

We next consider a real dataset: the wine quality dataset (Cortez et al., 2009)[5]. The dataset includes 12 variables that measure the physicochemical properties of wine and we treat the variable quality as the response of interest. The entire dataset consists of two groups: the

5. Available at `https://archive.ics.uci.edu/dataset/186/wine+quality`

white and red variants of the Portuguese "Vinho Verde" wine (1599 data points for the red wine and 4898 data points for the white wine). The subset of red wine is treated as the test dataset and that of white wine is viewed as the training set. All variables are nonnegative and we scale each variable by its largest value such that the entries are bounded by 1. Similar to the dataset partition in synthetic simulation, we fix $n_{\text{pre}} = 50$, $n_{\text{train}} = n = 1900$, and $n_{\text{test}} = 1000$. We first fit a kernel density estimator (Gaussian kernel with a bandwidth suggested by cross-validation) using the entire dataset as a proxy of the oracle density ratio. Figure 3.5a plots this against the log-density ratio obtained from logistic regression fitted on n_{pre} samples from each group. It can be seen that the two density ratios exhibit an approximately isotonic trend, which motivates us to consider the isotonic constraint with respect to w_0.



(a) Log-density ratio estimation: KDE versus logistic regression.

(b) Comparison with varying ρ (sample size $n_{\text{pre}} = 50$).

Figure 3.5: Results for wine quality dataset.

To assess the performance of the proposed approach, we estimate w_0 using the same procedure as for the simulated data, with sample size $n_{\text{pre}} = 50$ for each group. We consider the uncertainty set defined by KL-divergence and choose ρ from 50 uniformly located grid points in $[0.02, 2]$ in Figure 3.5b. The solid vertical line (in the middle and right plots) denotes an estimate $\widehat{\rho}$ of the KL divergence $D_{\text{KL}}(P_{\text{target}} \| P)$ to ensure that we are considering a reasonable range of values of ρ (see Appendix B.4.2 for details on this estimate). In Figure 3.5b, similar to the performance in Section 3.5.1 for simulated data, DRL tends to

be conservative: the coverage rate quickly approaches 1 while ρ is still below 0.1 and the intervals tend to be wide. In the meantime, iso-DRL-w_0 captures the approximate isotonic trend in Figure 3.5a and achieves valid coverage by recalibrating the weighted approach. The key message is that in the real data case, even when there is no oracle information for selecting ρ and the isotonic trend is not exact, the proposed iso-DRL-w_0 with the isotonic constraint with respect to the pre-fitted density ratio is less sensitive to the selection of ρ.

3.6 Additional related work

In this section, we discuss some additional literature in several related areas, including transfer learning, DRL, sensitivity analysis, shape-constrained learning, and conformal prediction.

Transfer learning. Transfer learning, in which data from one distribution is used to improve performance on a related but different distribution, is usually categorized into domain adaptation and inductive transfer learning (Redko et al., 2020).

Domain adaptation focuses on the scenario with covariate shift. From the theoretical side, the performance of machine learning models is analyzed in Ben-David et al. (2010); Ben-David and Urner (2012); Pathak et al. (2022); Pathak and Ma (2024); Hanneke and Kpotufe (2019). To implement efficient predictions, weighted methods are adopted as the first trial to draw P closer to Q (Cortes et al., 2008; Gretton et al., 2009; Ma et al., 2023; Ge et al., 2023). Another scenario requires a small number of labeled target samples, which can be feasible in reality and related works include Chen et al. (2011); Chattopadhyay et al. (2013); Yang et al. (2012), etc. Inductive transfer learning, on the other hand, assumes that the marginal distributions of X are invariant for training and target distributions and is studied in difference statistical settings (Bastani, 2021; Cai and Wei, 2019; Li et al., 2021; Tian and Feng, 2021).

Distributionally robust learning (DRL). Our work is directly related to DRL (Ben-Tal and Nemirovski, 1998; El Ghaoui and Lebret, 1997), which aims to control certain statistical risks uniformly over a set of candidate distributions for the target distribution. Different classes of the uncertainty set are studied in the literature, such as the optimal transport discrepancy (Shafieezadeh Abadeh et al., 2015; Blanchet and Murthy, 2019; Blanchet et al., 2019; Esfahani and Kuhn, 2015) and f-divergence (Duchi et al., 2021; Duchi and Namkoong, 2018; Weiss et al., 2023). Further constraints on the uncertainty set as the improvement of DRL are explored by Duchi et al. (2019); Setlur et al. (2023); Esteban-Pérez and Morales (2022); Liu et al. (2023); Popescu (2007); Shapiro and Pichler (2023). The recent work of Wang et al. (2023) considers the constraint that the unseen target distribution is a mixture of data distributions from multiple sources.

Sensitivity analysis. Sensitivity analysis is closely related to DRL and is widely studied in the field of causal inference (Cornfield et al., 1959; Rosenbaum, 1987; Tan, 2006; Ding and VanderWeele, 2016; Zhao et al., 2019; De Bartolomeis et al., 2023) with the goal of evaluating the effect of unmeasured confounders and relaxing untestable assumptions. Sensitivity models can be viewed as a specific example of constraints on distribution shift. For example, the marginal Γ-selection model (Tan, 2006) with a binary treatment T imposes a bound constraint on the distribution shift from the data distribution $P_{Y(1)|X,T=1}$ to the counterfactual $P_{Y(1)|X,T=0}$. Recent works also investigate the sensitivity model from the perspective of DRL, such as Yadlowsky et al. (2018); Jin et al. (2022, 2023); Sahoo et al. (2022). Sensitivity analysis that incorporates more informative constraints is explored in Huang and Pimentel (2024); Nie et al. (2021).

Statistical learning with shape constraints. Our work also borrows ideas from shape-constrained learning, which have been studied across various applications (Grenander, 1956; Matzkin, 1991). The isotonic constraint is the most common one among these. Since Rao

(1969), the properties of isotonic regression are well studied in the literature (Brunk et al., 1957, 1972; Zhang, 2002; Han et al., 2019; Yang and Barber, 2019; Durot and Lopuhaä, 2018). Moreover, the isotonic constraint is also widely applied to calibration for distributions in regression and classification settings (Zadrozny and Elkan, 2002; Niculescu-Mizil and Caruana, 2012; van der Laan et al., 2023; Henzi et al., 2021; Berta et al., 2024).

Conformal prediction. One important application of iso-DRL is to recalibrate conformal prediction intervals. Conformal prediction (Vovk et al., 2005; Shafer and Vovk, 2008; Angelopoulos et al., 2024, 2022; Zhou et al., 2024) provides a framework for distribution-free uncertainty quantification, which constructs prediction intervals that are valid with exchangeable data from any underlying distribution and with any "black-box" algorithm. As the validity of WCP (Tibshirani et al., 2019) with the estimated density ratio only holds up to a coverage gap due to the error the estimate w_0 (Lei and Candès, 2020; Candès et al., 2023; Gui et al., 2024a), the work Jin et al. (2023) further establish a robust guarantee via sensitivity analysis. Besides the weighted approaches, there are other solutions in the literature: Cauchois et al. (2020); Ai and Ren (2024) address the issue of joint distribution shift via the DRL; Qiu et al. (2023); Yang et al. (2024); Chen and Lei (2024) formulate the covariate shift problem within the semiparametric/nonparametric framework and utilize the doubly-robust theory to correct the distributional bias.

3.7 Discussion

In this paper, we focus on distributionally robust risk evaluation with the isotonic constraint on the density ratio as the regularization, which aims to avoid over-pessimistic candidate distributions. This is similar in flavor to many tools in high-dimensional statistical learning, where regularization/inductive bias is introduced to improve generalization. We provide an efficient approach to solve the shape-constrained optimization problem via an equivalent

reformulation, for which estimation error bounds for the worst-case excess risk are also provided. To conclude, we provide further discussions on the proposed iso-DRL framework and highlight several open questions.

Stability against distribution shift. Excess risk can also be interpreted from the perspective of stability against distribution shift (Lam, 2016; Namkoong et al., 2022; Rothenhäusler and Bühlmann, 2023). With a fixed budget $\varepsilon \ll 1$ for the excess risk, the largest tolerance of distribution shift such that the excess risk is under control is of interest. Taking the f-divergence constraint as an example, to ensure $\Delta_\rho(R; \mathcal{B}_{f,\rho}) \leq \varepsilon \ll 1$, then ρ needs to obey $\rho \leq (2\mathrm{Var}(R(X)))^{-1} f''(1) \cdot \varepsilon^2 + o(\varepsilon^2)$ (Lam, 2016; Duchi and Namkoong, 2018; Blanchet and Shapiro, 2023). However, with the additional isotonic constraint on the density ratio, we can tolerate larger distribution shift: $\rho \leq (2\mathrm{Var}([\pi(R)](X)))^{-1} f''(1) \cdot \varepsilon^2 + o(\varepsilon^2)$. This improvement implies that when side information of the underlying distribution shift is provided, risk evaluation will be less sensitive to the hyperparameters describing the uncertainty set (e.g., ρ), thus is more robust with the presence of distribution shift.

From risk evaluation to distributionally robust optimization. Different from risk evaluation, distributionally robust optimization (DRO) focuses on the optimization problem with a loss function $\ell_\theta(x)$, i.e., $\widehat{\theta} \in \operatorname{argmin}_{\theta \in \Theta} \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q\, \ell_\theta(X)$. Under smoothness conditions on ℓ_θ, asymptotic normality for $\widehat{\theta}$ is established in the literature (Duchi and Namkoong, 2018). The DRO framework is shown to regularize $\widehat{\theta}$ in terms of variance penalization (Lam, 2016; Duchi and Namkoong, 2018) or explicit norm regularization (Blanchet and Murthy, 2019). It is interesting to incorporate the isotonic constraint into DRO and to understand the effect of the isotonic constraint in the asymptotics of $\widehat{\theta}^{\mathrm{iso}}$.

CHAPTER 4

UNRAVELING PROJECTION HEADS IN CONTRASTIVE LEARNING: INSIGHTS FROM EXPANSION AND SHRINKAGE

We investigate the role of projection heads, also known as projectors, within the encoder-projector framework (e.g., SimCLR) used in contrastive learning. We aim to demystify the observed phenomenon where representations learned before projectors outperform those learned after—measured using the downstream linear classification accuracy, even when the projectors themselves are linear.

In this paper, we make two significant contributions towards this aim. Firstly, through empirical and theoretical analysis, we identify two crucial effects—expansion and shrinkage—induced by the contrastive loss on the projectors. In essence, contrastive loss either expands or shrinks the signal direction in the representations learned by an encoder, depending on factors such as the augmentation strength, the temperature used in contrastive loss, etc. Secondly, drawing inspiration from the expansion and shrinkage phenomenon, we propose a family of linear transformations to accurately model the projector's behavior. This enables us to precisely characterize the downstream linear classification accuracy in the high-dimensional asymptotic limit. Our findings reveal that linear projectors operating in the shrinkage (or expansion) regime hinder (or improve) the downstream classification accuracy. This provides the first theoretical explanation as to why (linear) projectors impact the downstream performance of learned representations. Our theoretical findings are further corroborated by extensive experiments on both synthetic data and real image data.[1]

1. This chapter is based on Gui et al. (2023c).

4.1 Introduction

Representation learning (Bengio et al., 2013) is a fundamental task in machine learning and statistics with the aim of extracting representations from the data that are useful for building future classifiers or predictors. While supervised learning is effective for this purpose (for instance, deep neural networks such as ResNet (He et al., 2016) have achieved remarkable performance in image classification), it is limited by the availability of massive labeled data.

Self-supervised learning (SSL) (Balestriero et al., 2023) has recently emerged as a novel paradigm to learn meaningful representations from huge unlabeled datasets (Misra and van der Maaten, 2019; Chen et al., 2020a; He et al., 2020; Dwibedi et al., 2021; HaoChen et al., 2021; Jing et al., 2021; Wang and Isola, 2020; Ji et al., 2021). Among SSL methods (Chen et al., 2020a; Zbontar et al., 2021; Bardes et al., 2021), contrastive learning (Chen et al., 2020a) is arguably the most popular one, which is also the focus of this paper. In essence, contrastive learning learns representations by encouraging proximity between the representations of similar inputs (also known as positive pairs), while forcing the representations of dissimilar inputs (i.e., negative pairs) to be far from each other.

Below, we compare contrastive learning with the classical representation learning methods to help readers better understand the former one. Readers familiar with contrastive learning can jump directly to Section 4.1.3.

4.1.1 Classical unsupervised learning based on encoders and decoders

Principal component analysis (PCA), dating back to Pearson (1901); Hotelling (1933), is perhaps the oldest unsupervised representation learning method. In a nutshell, PCA aims to find a linear function $\boldsymbol{f}_{\mathbf{W}}(\boldsymbol{x}) = \mathbf{W}\boldsymbol{x}$ of the input $\boldsymbol{x}$ that preserves as much information about the original data as possible. Mathematically, PCA can be formulated as minimizing the reconstruction loss: denoting $\boldsymbol{h} = \boldsymbol{f}_{\mathbf{W}}(\boldsymbol{x})$, we search for another linear function $\boldsymbol{g}_{\mathbf{T}}(\boldsymbol{h}) = \mathbf{T}\boldsymbol{h}$

such that the empirical reconstruction loss

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i - \boldsymbol{g}_{\mathbf{T}}(\boldsymbol{f}_{\mathbf{W}}(\boldsymbol{x}_i))\right\|^2$$

is minimized over all possible linear maps $\boldsymbol{f}_{\mathbf{W}}$ and $\boldsymbol{g}_{\mathbf{T}}$ of fixed dimensions. Here, $\{\boldsymbol{x}_i\}_{i\leq n}$ denotes the input data.

Using the machine learning terminology, the linear function $\boldsymbol{f}_{\mathbf{W}}(\cdot)$ is called an *encoder* that maps input data to latent representations while the linear function $\boldsymbol{g}_{\mathbf{T}}(\cdot)$ is called a *decoder* that reproduces the original input as accurately as possible. More generally, this encoder-decoder approach forms the core principle of many other representation learning methods (Ghojogh et al., 2023) including autoencoders (Bourlard and Kamp, 1988) and variational autoencoders (Kingma and Welling, 2013).

This encoder-decoder approach serves as the precursor of modern deep learning. Before 2010, learning features from large unlabeled data (also called pretraining), often followed by fine-tuning on a smaller labeled dataset, is known to be beneficial for downstream tasks (Hinton and Salakhutdinov, 2006). However, it is realized that the decoder component is not essential if we are not asking for a generative model, which can be time-consuming to train (Chen et al., 2020a). This gives rise to new attempts and ideas of learning with no or limited labeled data. Contrastive learning is one emerging approach that achieves outstanding empirical performance, producing state-of-the-art models such as CLIP (Radford et al., 2021).

4.1.2 An encoder-projector framework for contrastive learning

Contrastive learning marks its departure from the classical representation learning methods (e.g., autoencoders (Bourlard and Kamp, 1988; Kingma and Welling, 2013) in two aspects: (1) first, instead of the encoder-decoder framework, contrastive learning generally follows

the encoder-projector framework (cf. Figure 4.1); (2) second, instead of minimizing the reconstruction loss, contrastive learning minimizes a contrastive loss with the aim to pull positive pairs closer and push negative ones farther. Below we detail these two modifications and other essential components of contrastive learning using images as a running example for the input data.

Encoder. Similar to autoencoders, contrastive learning starts with an encoder $\boldsymbol{h} = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$ that maps an input image $\boldsymbol{x}$ to its feature representation $\boldsymbol{h}$. Oftentimes, the encoder $\boldsymbol{f}_{\boldsymbol{\theta}}$ is a deep neural network, e.g., ResNet-50 (He et al., 2016).

Projector. Contrastive learning has a unique component $\boldsymbol{z} = \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{h})$ that projects the representation $\boldsymbol{h}$ to its embedding $\boldsymbol{z}$, used later for calculating loss functions. This map $\boldsymbol{g}_{\boldsymbol{\varphi}}$ is usually called the *projection head* or *projector*. One often uses a simple network, e.g., a multilayer perceptron (MLP) with one or two hidden layers for the projector.

Augmentation and contrastive loss. As we have mentioned, contrastive learning aims to learn representations that are close for positive pairs and far for negative pairs. Here we provide one example for building positive and negative pairs from unlabeled data. Each sample $\boldsymbol{x}_i$ is augmented by transformations (e.g., color distortion) to produce semantically similar positive samples $\{\boldsymbol{x}_{i,k}\}_{1\leq k\leq K}$, where K denotes the number of augmentations (a.k.a. views). Two augmentations $(\boldsymbol{x}_{i,k}, \boldsymbol{x}_{j,k'})$ are called a positive pair if $i = j$, and a negative pair if $i \neq j$. Figure 4.2 gives an example of building positive and negative pairs from unlabeled image data.

With the positive and negative pairs in place, we can introduce the contrastive loss used in contrastive learning, as opposed to the reconstruction loss in PCA and autoencoders. Let $\boldsymbol{z}_{i,k} \coloneqq \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_{i,k}))$. Denoting the (cosine) similarity score by $\mathrm{sim}(\boldsymbol{z}, \boldsymbol{z}') \coloneqq \langle \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|}, \frac{\boldsymbol{z}'}{\|\boldsymbol{z}'\|} \rangle$, we train the encoder and projector by minimizing a contrastive loss, e.g., the canonical SimCLR

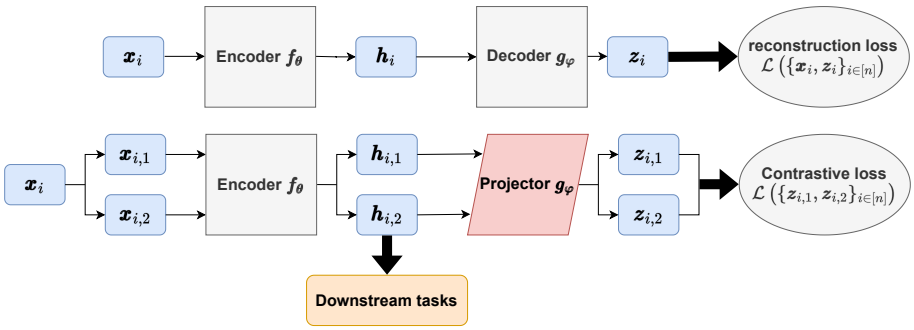


Figure 4.1: Encoder-decoder framework vs. encoder-projector framework.

loss (Chen et al., 2020a):

$$\min_{\boldsymbol{\theta},\boldsymbol{\varphi}} \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\varphi}) := \underbrace{-\frac{1}{\tau}\sum_{i}\sum_{k\neq k'}\operatorname{sim}(\boldsymbol{z}_{i,k},\boldsymbol{z}_{i,k'})}_{\text{alignment loss}} + \underbrace{(K-1)\sum_{i,k}\log\Big(\sum_{(j,k')\neq(i,k)} e^{\operatorname{sim}(\boldsymbol{z}_{i,k},\boldsymbol{z}_{j,k'})/\tau}\Big)}_{\text{uniformity loss}} \tag{4.1}$$

where $\tau > 0$ is known as the *temperature* parameter. Following Wang and Isola (2020), we call the first term the alignment loss that promotes feature proximity of positive pairs, and the second term the uniformity loss that repels negative pairs.

In fact, the contrastive loss can be viewed as pairwise cross-entropy loss (a.k.a. logistic loss). Let $s = (i,k)$ denote the index tuple for simplicity. In this case, s, s' are a positive pair if and only if $i = i'$. Then equivalently, one has $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\varphi}) = \sum_s \mathcal{L}_s(\boldsymbol{\theta},\boldsymbol{\varphi})$ where

$$\mathcal{L}_s(\boldsymbol{\theta},\boldsymbol{\varphi}) = -\sum_{s'\neq s}\mathbf{1}\{i=i'\}\log\left(\frac{\exp\big(\operatorname{sim}(\boldsymbol{z}_{s'},\boldsymbol{z}_s)/\tau\big)}{\sum_{\bar{s}\neq s}\exp\big(\operatorname{sim}(\boldsymbol{z}_{\bar{s}},\boldsymbol{z}_s)/\tau\big)}\right).$$

Downstream accuracy. After training, we freeze $\boldsymbol{\theta}$ and only keep the encoder $\boldsymbol{f}_{\boldsymbol{\theta}}$ (i.e., throwing away the projector). Later, given a downstream task, say a classification problem with labeled data $(\widetilde{y}_i, \widetilde{\boldsymbol{x}}_i)_{i\leq m}$, one can simply apply logistic regression to learned features $\{\boldsymbol{f}_{\boldsymbol{\theta}}(\widetilde{\boldsymbol{x}}_i)\}$ and labels $\{\widetilde{y}_i\}$.

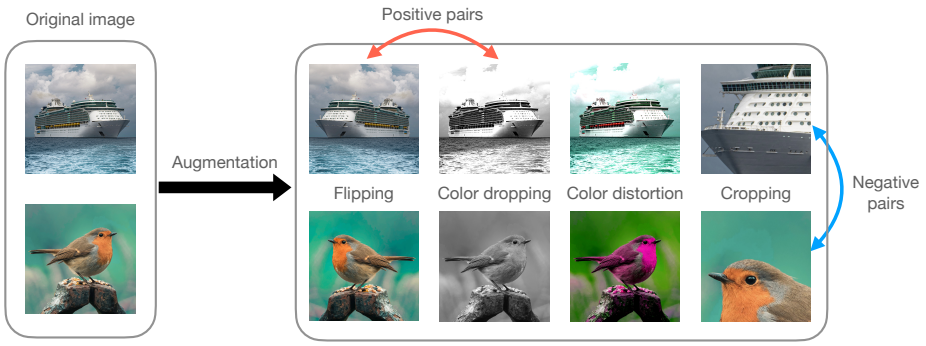


Figure 4.2: Positive and negative pairs after image augmentation.

4.1.3 Puzzling effect of projectors on the representations

The success of contrastive learning, or more specifically SimCLR (Chen et al., 2020a), can be attributed (at least) to two ingredients: strong data augmentation and the use of projectors. However, the role of projectors is quite puzzling. It is observed in Chen et al. (2020a) that representations learned before projectors outperform those learned after—measured using the downstream linear classification accuracy. This is even true when the projectors themselves are linear; see Figure 8 therein. As a result, the standard practice in contrastive learning is to jointly train the encoder and the projector, and then remove the projector completely after training (Balestriero et al., 2023).

In this paper, we aim to demystify this puzzling phenomenon about projectors in contrastive learning. In particular, we focus on answering the following two questions:

Q1: *What geometric structure does contrastive loss minimization induce on the projectors?*

Q2: *Why do (linear) projectors affect the generalization properties of learned features?*

For Q1, the contrastive loss is relatively new and thus less understood. It is unclear what properties the optimal solution to contrastive loss minimization possesses. A quantitative characterization will be helpful for understanding contrastive learning.

For Q2, heuristics in the literature are insufficient (see e.g., Section 3.2 in Balestriero et al. (2023)). One may view a projector as a buffer component: it protects features from

distortion due to loss minimization, so its removal after training improves the generalization properties of features. Yet, this argument cannot reconcile with the observation that even linear projectors affect the downstream linear classification accuracy. Therefore, without a thorough investigation, the role of projectors on generalization remains mysterious.

4.1.4 Our contributions

To delineate the effect of projectors, we do not attempt to analyze the encoder or its training dynamics, but rather assume access to a well-trained encoder, which is often achieved at the later stage of training. Through fixing such a good encoder, we make the following empirical and theoretical discoveries that are fundamental to addressing the aforementioned two questions.

1. First, we identify two crucial effects—expansion and shrinkage—induced by the contrastive loss on the projectors. In essence, contrastive loss either expands or shrinks the signal direction in the representations learned by encoders.

2. Secondly, under a simpler projector model, we precisely characterize the downstream linear classification accuracy in the high-dimensional asymptotic limit. Our findings reveal that linear projectors operating in the shrinkage (resp. expansion) regime hinder (resp. improve) the downstream classification accuracy. This provides the first theoretical explanation as to why (linear) projectors impact the downstream performance of learned representations.

We also discuss connections to other empirical phenomena such as dimensional collapse, feature transferability, neural collapse, etc.

4.1.5 Paper organization

In Section 4.2, we present the empirical findings regarding the effects of the contrastive loss on the projectors, including expansion and shrinkage. In Section 4.3, we introduce the feature-level Gaussian mixture model and provide empirical evidence as motivation for our modeling approach. Moving on to Section 4.4, we present a precise approximation of the population contrastive loss under the Gaussian mixture model. Additionally, we theoretically characterize the sharp phase transition between the expansion and shrinkage regimes, and provide the approximation bound along with the finite-sample loss. Section 4.5 analyzes the impact of the expansion/shrinkage phenomenon of the linear projection head on downstream tasks. We calculate the precise generalization error as a function of the expansion effect. Furthermore, in Section 4.6, we extend our expansion/shrinkage results from Section 4.4 to inhomogeneous augmentations. We identify the simultaneous expansion and shrinkage, which aligns with our empirical discoveries using the STL-10 dataset. Simulation details and theoretical proofs are deferred to the appendix.

4.1.6 Notations

$\mathbb{R}_+$ denotes $\{a \in \mathbb{R} : a \geq 0\}$ and $\bar{\mathbb{R}}$ denotes $\mathbb{R} \cup \{\infty\}$. For a vector $\boldsymbol{u}$, we use $\|\boldsymbol{u}\|$ or $\|\boldsymbol{u}\|_2$ to denote its Euclidean norm. For vectors $\boldsymbol{u}, \boldsymbol{v}$ of the same length, we use $\langle \boldsymbol{u}, \boldsymbol{v} \rangle := \boldsymbol{u}^\top \boldsymbol{v}$ to denote the inner product. For a matrix $\mathbf{A}$, we use $\|\mathbf{A}\|_{\mathrm{F}}$ to denote its Frobenius norm. The identity matrix of size $p \times p$ is denoted by $\mathbf{I}_p$ or simply $\mathbf{I}$. We use $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The notation Φ means the cumulative distribution function of the standard Gaussian variable.

For two real-valued sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$, we use the standard small-o notation: $a_n = o(b_n)$ means $\lim_{n\to\infty} a_n/b_n = 0$, and sometimes we also write $a_n \ll b_n$. For random variable X_n, $X_n = o_{\mathbb{P}}(b_n)$ means $|X_n|/b_n$ converges in probability to 0 as $n \to \infty$. Moreover, if $\mathbf{X}_n$ is a random vector, then $\mathbf{X}_n = o_{\mathbb{P}}(b_n)$ means $\|\mathbf{X}_n\|/b_n$ converges in probability to 0

as $n \to \infty$.

4.2 Empirical discovery: expansion and shrinkage

We begin with presenting our empirical findings on the two crucial effects—expansion and shrinkage, of contrastive loss on projectors.

Experimental setup. We provide a brief description about our experimental setup, and leave the details to the appendix. We freeze the encoder (based on ResNet-18)[2] pretrained on the STL-10 image dataset.[3] Then we apply standard data augmentation techniques such as random cropping and color distortion to generate positive/negative pairs. In the end, we train a linear projector $\mathbf{W} \in \mathbb{R}^{512\times 512}$ under different configurations of hyperparameters, including 5 different temperatures.

In what follows, we summarize our empirical findings.

4.2.1 Insights from spectral decomposition

First, for two different class labels c_1, c_2 (e.g., c_1 denotes airplane while c_2 denotes dog), we use $\boldsymbol{\mu}_{c_k} \coloneqq \mathrm{Ave}_{y_i=c_k}\boldsymbol{h}_i$ $(k = 1, 2)$ to denote the average representation after the encoder, where Ave is the average operator. Correspondingly, we denote $\boldsymbol{\mu}_{c_1,c_2} \coloneqq \boldsymbol{\mu}_{c_1} - \boldsymbol{\mu}_{c_2} \in \mathbb{R}^p$ to be the difference between class means. It turns out that $\boldsymbol{\mu}_{c_1,c_2}$is closely connected to the top/bottom right singular subspaces of the linear projector $\mathbf{W}$.

To be more precise, let $\mathcal{V}_{\mathrm{top}}, \mathcal{V}_{\mathrm{bottom}}$ be the top/bottom singular subspaces containing a few right singular vectors (SVs) of $\mathbf{W}$ and $\mathcal{V}_{\mathrm{bulk}}$ be the singular subspace containing the remaining SVs. Clearly one has $\mathbb{R}^p = \mathcal{V}_{\mathrm{top}} \oplus \mathcal{V}_{\mathrm{bulk}} \oplus \mathcal{V}_{\mathrm{bottom}}$.

2. Downloaded from `https://github.com/sthalles/SimCLR`.

3. Source from `https://cs.stanford.edu/~acoates/stl10/`.

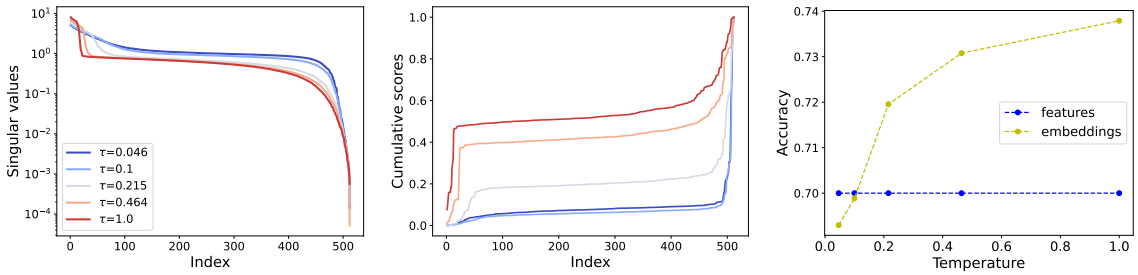


Figure 4.3: Results with the pretrained encoder and a one-layer linear projector $\mathbf{W} \in \mathbb{R}^{512\times512}$ under the standard SimCLR loss (ResNet-18 and 10-class STL-10 dataset). **Left**: singular values of $\mathbf{W}$ with varying temperature τ. **Middle**: cumulative score plot $\text{score}_i = \sum_{j\le i}\langle \boldsymbol{v}_j, \boldsymbol{\mu}_{c_1,c_2}\rangle^2/\|\boldsymbol{\mu}_{c_1,c_2}\|^2$ where $(c_1, c_2) = (7, 2)$. **Right**: downstream task accuracy using features $\boldsymbol{h}$ versus using embeddings $\boldsymbol{z}$ on the test set containing all 10 classes.

Our first main empirical finding is that, empirically, the following holds approximately:

$$\boldsymbol{\mu}_{c_1,c_2} \in \mathcal{V}_{\text{top}} \oplus \mathcal{V}_{\text{bottom}}\,.$$

In other words, the energy is concentrated on the span of extreme (right) singular vectors. Moreover, on $\mathcal{V}_{\text{top}}$ and $\mathcal{V}_{\text{bottom}}$, the corresponding singular values experience sharp drops. Indeed, in Figure 4.3, we train $\mathbf{W}$ on the 10-class STL-10 dataset and calculate the alignment score $\kappa_{j,c_1,c_2} := \langle \boldsymbol{v}_j, \boldsymbol{\mu}_{c_1,c_2}\rangle^2/\|\boldsymbol{\mu}_{c_1,c_2}\|^2$, where $\boldsymbol{v}_j$ is the jth right SV of $\mathbf{W}$. For each index i, we report the cumulative score $\sum_{j\le i}\kappa_{j,c_1,c_2}$, which satisfies the normalization $\sum_{j=1}^{p}\kappa_{j,c_1,c_2} = 1$. A wide flat cumulative score suggests orthogonality $\boldsymbol{\mu}_{c_1,c_2} \perp \mathcal{V}_{\text{bulk}}$. A geometric interpretation is that $\mathbf{W}$ is *expanding* vectors in $\mathcal{V}_{\text{top}}$ and *shrinking* vectors in $\mathcal{V}_{\text{bottom}}$.

4.2.2 Expansion/shrinkage affects generalization

Our second empirical finding is that the expansion/shrinkage effects are highly correlated with downstream accuracy.

As we observed earlier, when the temperature τ increases, expansion gets stronger because the alignment of the top SVs with $\mu_{7,2}$ increases. Figure 4.3 (right) shows that a larger

temperature also leads to an increase in the classification accuracy using the embeddings $(z_i)_{i\leq n}$. Figure 4.3 (right) confirms the surprise that a linear projector can change the generalization performance significantly.

Can we theoretically justify the expansion and shrinkage effects of contrastive loss and their impact on the generalization power of the learned representations?

4.2.3 A simple simulation and heuristic explanations

Here, we present a simple simulation using synthetic input data that recreates the salient characteristics of our empirical findings. We hope to provide some heuristic explanations for the observed phenomena, while leaving the formal proof to later sections.

A visual illustration. The main characteristics of the empirical structure of projectors are displayed in a simple clean model. Consider a 2-component Gaussian mixture model $\frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_p)+\frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_p)$. We generate data according to this model, and then generate augmented data by adding independent perturbation (Gaussian noise). The data pass through a linear layer $\mathbf{W} \in \mathbb{R}^{p\times p}$ treated as the projector. We obtain $\mathbf{W}^*$ by minimizing the SimCLR loss over $\mathbf{W}$.

We project and visualize our simulated data in Figure 4.4 using $\mathbf{W}^*$. Under two different hyperparameter settings, the original data (shown in the left plot) are either (i) extended along the separation direction $\boldsymbol{\mu}$ in the middle plot, or (ii) compressed along $\boldsymbol{\mu}$ in the right plot.

The simulation result matches earlier empirical findings. (i) In the expansion regime, normalized features (projected to a circle) are more aligned, signaling a large top singular value; and in the shrinkage regime, normalized features are more uniform, signaling a smaller singular value. (ii) The test error decreases in the expansion regime that has a higher temperature, and it increases in the shrinkage regime that has a lower temperature.

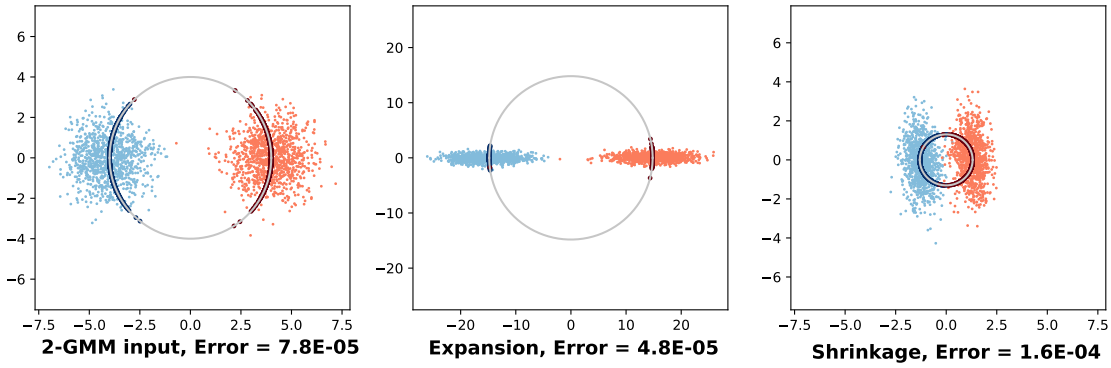


Figure 4.4: Contrastive loss drives expansion or shrinkage. **Left:** we generate clean 2-component GMM data. **Middle and right:** we use SimCLR loss for training 100-dim data, and then visualize embeddings on a 2D plane under large/small augmentation and large/small temperature (middle/right plot). Also, we show the projection of embeddings onto a circle with varying radii, which is $\boldsymbol{z}_i/\|\boldsymbol{z}_i\|$ used for calculating cosine similarities. **Bottom text:** expansion decreases test error, while shrinkage reduces signals and increases test error.

Expansion and shrinkage promote alignment and uniformity. In Wang and Isola (2020), the contrastive loss is decomposed into two components, which are called the *alignment* loss and the *uniformity* loss—which correspond to the two terms in the RHS of (4.1). Feature embeddings in the hypersphere are driven by the two opposite forces induced by the two loss components. Our expansion and shrinkage perspective offers a consistent explanation. If the projector stretches the features along the signal direction thus increasing variance in that direction, then after normalization to the hypersphere the features become more aligned. Conversely, if the projector compressed the features along the signal direction, after normalization features will become more uniform. See Section 4.4 for a formal analysis.

Projector as reparametrization changes inductive bias. Why does the projector, even in the linear case, change the generalization performance on downstream tasks? Reparametrization techniques such as skip connections (He et al., 2016) and batch normalization (Ioffe and Szegedy, 2015) are commonly used in deep learning, yet how they impact generalization is rarely elucidated. Here, we give an explanation from the lens of inductive bias. When the

downstream task involves linearly separable data, gradient descent on the commonly used logistic loss produces a sequence of iterates $\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \ldots$ that converge in direction to the max-margin solution (Soudry et al., 2018).

$$\begin{aligned} \max_{\boldsymbol{\beta}} \quad & \min_{i \le n} y_i \langle \boldsymbol{z}_i, \boldsymbol{\beta} \rangle \\ \text{subject to} \quad & \|\boldsymbol{\beta}\|_2 \le 1. \end{aligned}$$

The ℓ_2 norm in the constraint plays the role of implicit regularization that is induced by the gradient descent. Applying a linear transformation to features, however, will effectively lead to a different norm. Thus, the generalization properties on downstream tasks are affected by even a linear projector. See Section 4.5 for the formal analysis.

4.3 Feature-level modeling via GMM

To decouple the encoder and the projector, we assume access to a well-trained encoder and model the output using a well-separated Gaussian mixture model (GMM), where each Gaussian component represents a class. This assumption is also supported by at least two empirical observations: (i) the cluster structure in the feature space, and (ii) convergence speed during training. We defer detailed justifications about this assumption to Section 4.3.2, before which we detail the model setup using GMMs.

4.3.1 Model setup

Mathematically, we assume that the outputs of the encoder (i.e., features) are generated from a 2-component GMM (or 2-GMM), that is

$$\boldsymbol{h}_{0,i} \overset{\text{i.i.d.}}{\sim} \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_p) + \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_p),$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ denotes the mean difference between two classes. Conditional on $\boldsymbol{h}_{0,i}$, we construct two augmentations (views) via

$$\boldsymbol{h}_i, \boldsymbol{h}_i^+ \mid \boldsymbol{h}_{0,i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{h}_{0,i}, \sigma_{\text{aug}}^2 \mathbf{I}_p),$$

where $\sigma_{\text{aug}}^2 > 0$ represents the augmentation strength. As a result, $(\boldsymbol{h}_i, \boldsymbol{h}_i^+)$ is viewed as a positive pair, and $(\boldsymbol{h}_i, \boldsymbol{h}_j)$, $(\boldsymbol{h}_i, \boldsymbol{h}_j^+)$ $(i \neq j)$ are considered negative pairs. For the projector, we focus on linear projectors, i.e., $\boldsymbol{z} = \boldsymbol{g}_{\mathbf{W}}(\boldsymbol{h}) = \mathbf{W}\boldsymbol{h}$, where $\mathbf{W} \in \mathbb{R}^{p \times p}$. We present a modified SimCLR loss that is more amenable to theoretical analysis. Given the temperature $\tau > 0$ and projector outputs $(\boldsymbol{z}_i)_{i \leq n}$ where $\boldsymbol{z}_i = \boldsymbol{g}_{\mathbf{W}}(\boldsymbol{h}_i)$, we define

$$\mathcal{L}_n(\mathbf{W}) \coloneqq -\frac{1}{n\tau} \sum_{i=1}^n \mathbb{E}_{\boldsymbol{z}_i^+, \boldsymbol{z}_i | \boldsymbol{h}_{0,i}} \left[\text{sim}^*(\boldsymbol{z}_i, \boldsymbol{z}_i^+) \right] + \log \left(\frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\boldsymbol{z}_i^-, \boldsymbol{z}_i | \boldsymbol{h}_{0,i}} \left[e^{\text{sim}^*(\boldsymbol{z}_i^-, \boldsymbol{z}_i)/\tau} \right] \right) \quad (4.2)$$

$$\text{where } \text{sim}^*(\boldsymbol{z}_1, \boldsymbol{z}_2) \coloneqq \frac{-\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|^2/2}{\{\mathbb{E}[\|\boldsymbol{z}_1\|^2]\}^{1/2} \cdot \{\mathbb{E}[\|\boldsymbol{z}_2\|^2]\}^{1/2}} .$$

Here, $\mathbb{E}_{\boldsymbol{z}_i^+, \boldsymbol{z}_i | \boldsymbol{h}_{0,i}}$ denotes expectation over random augmentations *conditioning* on $\boldsymbol{h}_{0,i}$ (i.e., using every possible positive pair), $\boldsymbol{z}_i^-$ denotes an independent copy of $\boldsymbol{z}_i$, and $\mathbb{E}_{\boldsymbol{z}_i^-, \boldsymbol{z}_i | \boldsymbol{h}_{0,i}}$ denotes expectation over the negative pair $(\boldsymbol{z}_i, \boldsymbol{z}_i^-)$ conditioning on $\boldsymbol{h}_{0,i}$ (i.e., using every possible negative pair).

Our modified SimCLR loss is different from the original SimCLR in the following ways: (i) We consider full batch and training for infinite time (so averages over augmentations are replaced by expectations); (ii) similar to Wang and Isola (2020), we interchange log with the summation; and (iii) we use a variant of similarity score sim* where we replace instance-based normalization with the population level normalization.

We emphasize that this modified loss is only used for theoretical analysis. In all experiments, we use the original SimCLR contrastive loss.

4.3.2 Why fixing a well-trained encoder?

Arguably, our theoretical study departs from common practice as we assume a well-trained encoder and fix it to be a simple Gaussian mixture model. However, we would like to argue that this assumption sheds light on the practice.

Convergence of projectors. Training an encoder and projector jointly creates complex dynamics that are beyond the scope of this paper. Yet, we observe in experiments that the dynamics are significantly simplified in the later stage of training. Let $(\boldsymbol{\theta}^{(t)}, \mathbf{W}^{(t)})$ be the parameters at epoch t, and $\widetilde{\mathbf{W}}^{(t)}$ be the optimal projector parameters $\widetilde{\mathbf{W}}^{(t)} := \operatorname{argmin}_{\mathbf{W}} \mathcal{L}(\boldsymbol{\theta}^{(t)}, \mathbf{W})$ while freezing encoder parameters $\boldsymbol{\theta}^{(t)}$ at epoch t. It can be seen from Figure C.5 in the appendix that when $t \geq 50$,

$$\|\widetilde{\mathbf{W}}^{(t)} - \mathbf{W}^{(t)}\| \ll \|\mathbf{W}^{(t)}\|,$$

which suggests that the projector is close to (conditionally) optimal. This allows us to decouple the dynamics and encoders and projectors into two separate optimization problems:

1. For fixed encoder $\boldsymbol{\theta}$, solve $\min_{\mathbf{W}} \mathcal{L}(\boldsymbol{\theta}, \mathbf{W})$.
2. Solve $\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \widetilde{\mathbf{W}}(\boldsymbol{\theta}))$ where $\widetilde{\mathbf{W}}(\boldsymbol{\theta})$ is the minimizer of the first part.

Our focus is the first optimization problem, assuming that the encoder $\boldsymbol{\theta}$ is sufficiently good, e.g., $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$ for large t. We remark that the feature-learning process and training dynamics of encoder networks are complicated and elusive for analysis, so it is not uncommon to decouple multiple layers or components by freezing some of them Han et al. (2021); Bietti et al. (2022).

Cluster structure in feature space. When the encoders are trained well, we observe well-separated clusters formed by pretrained features in contrastive learning, and these clus-

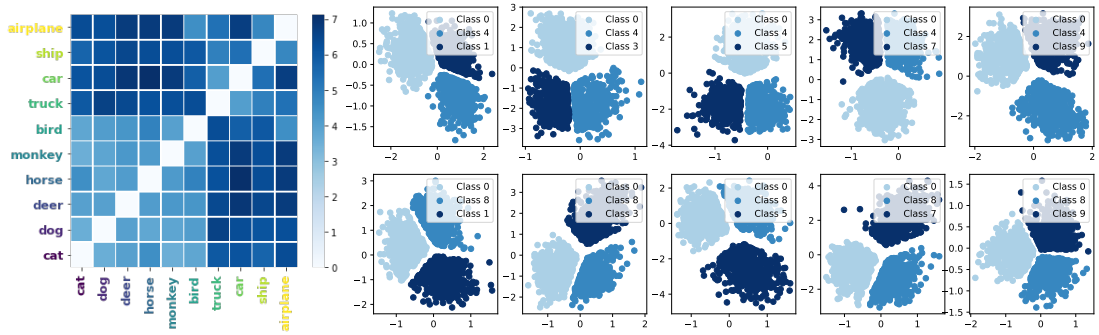


Figure 4.5: Pretrained STL-10 features from the standard SimCLR model exhibit a strong cluster structure. **Left**: positive pairwise margins of linear classifier imply pairwise linear separability. **Right**: 2D visualization of 3-class subsets of features after 30 epochs. The cluster structure emerges quickly.

ters contain class information (Böhm et al., 2022). We examine the cluster structure of pretrained features both visually and quantitatively. In Figure 4.5 (left plot), for every pair of classes, we perform max-margin linear classification on a subset of pretrained features. We find that every pair achieves zero training error with a large margin, which indicates that all classes are linearly separable. In Figure 4.5 (right plot), we visualize pretrained features by following the same projection technique Müller et al. (2019). For well-clustered pretrained features, the GMM is a natural model, which is the starting point of our analysis.

4.4 Contrastive loss drives expansion/shrinkage

Under the model setup in Section 4.3.1, we hope to characterize the minimizer of $\mathcal{L}_n(\mathbf{W})$. However, the loss $\mathcal{L}_n$ is very complex, and quantitative results are hard to obtain. It turns out that, under reasonable simplification, the minimization admits an explicit solution. We then quantify the approximation error in the aforementioned simplification.

4.4.1 Quantitative results under simplification

We start our analysis with the infinite-sample case, i.e., $n = \infty$. Let $\mathcal{L}(\mathbf{W}) \coloneqq \mathbb{E}_{\boldsymbol{z}}\mathcal{L}_n(\mathbf{W})$ be the population loss. Denote by $\mathbf{W} = \sum_{j=1}^{p} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^\top$ the SVD of $\mathbf{W} \in \mathbb{R}^{p\times p}$, where

$\sigma_1 \geq \ldots \sigma_p \geq 0$, $[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p]$ and $[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p]$ are orthogonal matrices. Define

$$\alpha := \alpha(\mathbf{W}) = (1 + \sigma_{\text{aug}}^2)\|\mathbf{W}\|_{\text{F}}^2 + \|\mathbf{W}\boldsymbol{\mu}\|^2.$$

A straightforward calculation yields the following decomposition of the population loss.

Proposition 4.4.1. *We have* $\mathcal{L}(\mathbf{W}) = \mathcal{L}_{\text{align}}(\mathbf{W}) + \mathcal{L}_{\text{unif}}(\mathbf{W})$ *with*

$$\begin{aligned}
\mathcal{L}_{\text{align}}(\mathbf{W}) &= \frac{\sigma_{\text{aug}}^2}{\tau\alpha}\|\mathbf{W}\|_{\text{F}}^2, \\
\mathcal{L}_{\text{unif}}(\mathbf{W}) &= -\log 2 - \frac{1}{2}\sum_{j=1}^{p} \log\left(1 + \frac{2(1+\sigma_{\text{aug}}^2)\sigma_j^2}{\tau\alpha}\right) \\
&\qquad + \log\Big(1 + \exp\Big(-\sum_{j=1}^{p} \frac{2\sigma_j^2\langle\boldsymbol{\mu}, \boldsymbol{v}_j\rangle^2}{2(1+\sigma_{\text{aug}}^2)\sigma_j^2 + \tau\alpha}\Big)\Big) \\
&=: -\log 2 + \mathcal{L}_{\text{unif}}^{(1)}(\mathbf{W}) + \mathcal{L}_{\text{unif}}^{(2)}(\mathbf{W}).
\end{aligned}$$

See Appendix C.2.1 for the proof.

To understand the two loss components, let us make some observations.

1. The main effect of $\mathcal{L}_{\text{align}}(\mathbf{W})$ is to align positive pairs. If we only minimize $\mathcal{L}_{\text{align}}(\mathbf{W})$, then we need to maximize[4] $\|\mathbf{W}\boldsymbol{\mu}\|$ for any fixed value of $\|\mathbf{W}\|$. In other words, $\mathbf{W}$ needs to stretch $\boldsymbol{\mu}$ as much as possible. This is intuitive since positive pairs after such stretching and normalization will be closer to each other; see the middle plot of Figure 4.4.

2. The main effect of $\mathcal{L}_{\text{unif}}(\mathbf{W})$ is to repel negative pairs. If we only minimize one critical part $\mathcal{L}_{\text{unif}}^{(1)}(\mathbf{W})$, then given fixed $(\sigma_j)_{j\leq p}$ we need to minimize $\|\mathbf{W}\boldsymbol{\mu}\|$ (i.e., $\mathbf{W}$ compresses $\boldsymbol{\mu}$ as much as possible), leading to $|\langle\boldsymbol{\mu}, \boldsymbol{v}_p\rangle| = 1$. Embeddings after

4. Note that α also depends on $\mathbf{W}$.

shrinkage and normalization will be more diversely distributed; see the right plot of Figure 4.4.

We can gain further insights by considering a first-order approximation of the loss: intuitively, if the signal strength $\|\boldsymbol{\mu}\|^2$ is large, then we may expect $\alpha \gg \|\mathbf{W}\|_{\mathrm{F}}^2 \geq \max_j \sigma_j^2$, so we can try a first-order expansion by treating α as a diverging quantity. This motivates us to define an approximate loss:

$$\begin{aligned}\widetilde{\mathcal{L}}(\mathbf{W}) &\coloneqq \frac{\sigma_{\mathrm{aug}}^2}{\tau\alpha}\|\mathbf{W}\|_{\mathrm{F}}^2 - \log 2 - \sum_{j=1}^{p}\frac{(1+\sigma_{\mathrm{aug}}^2)\sigma_j^2}{\tau\alpha} + \log\Big(1+\exp\Big(-\sum_{j=1}^{p}\frac{2\sigma_j^2\langle\boldsymbol{\mu},\boldsymbol{v}_j\rangle^2}{\tau\alpha}\Big)\Big) \\ &= -\frac{t}{\tau} + \log\Big(\frac{1}{2}+\frac{1}{2}\exp\big(-\frac{2}{\tau}+\frac{2(1+\sigma_{\mathrm{aug}}^2)}{\tau}t\big)\Big), \end{aligned} \tag{4.3}$$

where $t \coloneqq t(\mathbf{W}) = \frac{\|\mathbf{W}\|_{\mathrm{F}}^2}{\alpha(\mathbf{W})}$. As an approximation, it always holds that $\mathcal{L}(\mathbf{W}) \geq \widetilde{L}(\mathbf{W})$. Now that $\widetilde{L}(\mathbf{W})$ is actually a univariate function, we can state precise quantitative results.

Theoretical prediction. Now we are ready to theoretically demonstrate the two key effects—expansion and shrinkage. Let $F(t;\sigma_{\mathrm{aug}}^2,\tau) \coloneqq \frac{\mathsf{d}}{\mathsf{d}t}\widetilde{\mathcal{L}}(\mathbf{W})$.

Definition 4.4.2. A three-parameter configuration $(\sigma_{\mathrm{aug}}^2, \tau, \|\boldsymbol{\mu}\|^2)$ is said to be in the

- expansion regime if $F\big(\frac{1}{1+\sigma_{\mathrm{aug}}^2+\|\boldsymbol{\mu}\|^2};\sigma_{\mathrm{aug}}^2,\tau\big) > 0$,
- shrinkage regime if $F\big(\frac{1}{1+\sigma_{\mathrm{aug}}^2+\|\boldsymbol{\mu}\|^2};\sigma_{\mathrm{aug}}^2,\tau\big) < 0$.

The expansion and shrinkage regimes characterize the distinctive behavior of the projector $\mathbf{W}$ at minimization, as stated in the next theorem.

Theorem 4.4.3 (expansion/shrinkage phase transition)**.** *The properties of the minimizer* $\mathbf{W}^*$ *of* $\widetilde{\mathcal{L}}(\mathbf{W})$ *depends on the configuration of* $(\sigma_{\text{aug}}^2, \tau, \|\boldsymbol{\mu}\|^2)$*. Specifically, with the notation* $t^* \coloneqq t(\mathbf{W}^*)$*,*

- *in the expansion regime,* $\widetilde{\mathcal{L}}(\mathbf{W})$ *is minimized at* $t^* = 1/(1+\sigma_{\text{aug}}^2+\|\boldsymbol{\mu}\|^2)$*, which happens if and only if* $\sigma_2 = \cdots = \sigma_p = 0$ *and* $\langle \boldsymbol{v}_1, \boldsymbol{\mu}\rangle^2 = \|\boldsymbol{\mu}\|^2$*;*
- *in the shrinkage regime,* $\widetilde{\mathcal{L}}(\mathbf{W})$ *is minimized at*

$$t^* = \frac{1}{1+\sigma_{\text{aug}}^2}\left(1 - \frac{\tau}{2}\log(1+2\sigma_{\text{aug}}^2)\right) \in \left[\frac{1}{1+\sigma_{\text{aug}}^2+\|\boldsymbol{\mu}\|^2}, \frac{1}{1+\sigma_{\text{aug}}^2}\right]$$

that solves $F(t) = 0$*. Moreover, if* $\sigma_{\text{aug}}^2 \to 0$*, then* $t^* \to 1$ *and* $\|\mathbf{W}\boldsymbol{\mu}\|/\|\mathbf{W}\|_{\text{F}} \to 0$*.*

See Appendix C.2.3 for the proof.

Several remarks are in order. First, when the configuration $(\sigma_{\text{aug}}^2, \tau, \|\boldsymbol{\mu}\|^2)$ is in the expansion regime, optimizing the contrastive loss leads to a linear projector $\mathbf{W}^*$ that maximally expands along the signal direction $\boldsymbol{\mu}$. Second, in contrast, in the shrinkage regime, the signal direction is compressed in $\mathbf{W}^*$, since for fixed value $\|\mathbf{W}\|_{\text{F}}$, a vanishing $\|\mathbf{W}\boldsymbol{\mu}\|$ implies $\max_j |\sigma_j \langle \boldsymbol{v}_j, \boldsymbol{\mu}\rangle| \to 0$—any singular vector positively correlated with $\boldsymbol{\mu}$ must have a vanishing singular value.

Numerical experiments. Under the 2-GMM model stated in Section 4.3, in Figure 4.6 (left), we plot the heatmap of the expansion measure $T^* \coloneqq T(\mathbf{W}^*)$ with $T(\mathbf{W}) \coloneqq \|\mathbf{W}\boldsymbol{\mu}\|^2/(\|\mathbf{W}\|_{\text{F}}^2\|\boldsymbol{\mu}\|^2) \in [0,1]$ with different pairs of temperature and augmentation strength. The green curve is the transition curve

$$\tau^* = 2\|\boldsymbol{\mu}\|^2\left\{(1+\sigma_{\text{aug}}^2+\|\boldsymbol{\mu}\|^2)\log(1+2\sigma_{\text{aug}}^2)\right\}^{-1}$$

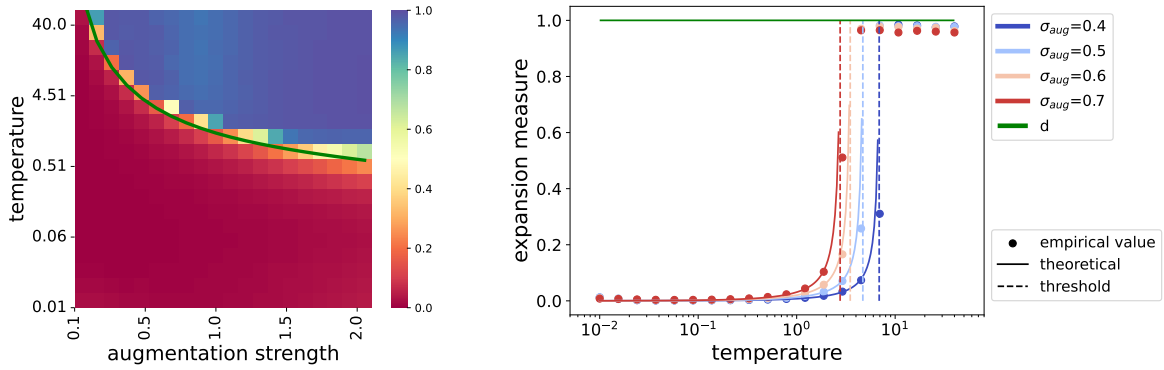


Figure 4.6: Expansion/shrinkage phase transition for the configuration $n = 2000$, $p = 50$, $\|\boldsymbol{\mu}\| = 5$. **Left**: heatmap of expansion measure $T^* := T(\mathbf{W}^*)$; **green curve**: theoretical phase transition. **Right**: T^* vs. τ with different fixed σ_{aug}.

which separates the expansion and shrinkage regimes. In Figure 4.6 (right), with a set of σ_{aug}, we plot the expansion measure against varying temperatures which shows that our theoretical prediction is also precise in the regime when $0 < T^* < 1$. When $\tau < \tau^*$ (shrinkage), T^* is close to zero, indicating a significant compression along $\boldsymbol{\mu}$; when $\tau > \tau^*$ (expansion), $T^* \approx 1$, corresponding to maximal expansion along $\boldsymbol{\mu}$.

Despite simplification and approximation, our theoretical prediction does align well with empirical phase transition with the original SimCLR loss.

4.4.2 Extensions to the general case

Recall that $\widetilde{\mathcal{L}}$ is an approximation to the original loss $\mathcal{L}$. Therefore it is of interest to quantify the approximation error in the optimal solutions. We also briefly discuss the technical difficult of extending these analysis to the finite-sample regime.

Approximation errors. Compared with $\widetilde{\mathcal{L}}$, the original loss $\mathcal{L}$ cannot be expressed as a simple univariate function. Recall that in the approximation, we consider large signal strength $\|\boldsymbol{\mu}\|$. Denote $\mathbf{W}_{\text{simclr}}$ as the solution to the original loss $\mathcal{L}$ and $t_{\text{simclr}} := t(\mathbf{W}_{\text{simclr}}) = \|\mathbf{W}_{\text{simclr}}\|/((1+\sigma_{\text{aug}}^2)\|\mathbf{W}_{\text{simclr}}\|^2 + \|\mathbf{W}_{\text{simclr}}\boldsymbol{\mu}\|^2)$. Approximation bounds are shown in the next proposition for the expansion and shrinkage regimes, respectively.

Proposition 4.4.4. *The difference between t^* and t_{simclr} has the following bounds:*

- *in the expansion regime,*

$$0 \le t_{\text{simclr}} - t^* \le \left(\frac{2(1+\sigma_{\text{aug}}^2)}{1+e^{2/\tau}} - 1 \right)^{-1} \frac{4(1+\sigma_{\text{aug}}^2)}{\tau e^{1/\tau}} \frac{1}{\|\boldsymbol{\mu}\|^2} + C(\tau, \sigma_{\text{aug}}^2) \|\boldsymbol{\mu}\|^{-4};$$

- *in the shrinkage regime,*

$$\|t_{\text{simclr}} - t^*\|^2 \le \frac{e^{2\sigma_{\text{aug}}^2/\tau}}{(1+\sigma_{\text{aug}}^2)^2} \frac{1}{p-1} + C(\tau) \cdot \left(\frac{\sigma_{\text{aug}}^4}{\|\boldsymbol{\mu}\|^2} \right).$$

where $C(\tau, \sigma_{\text{aug}}^2), C(\tau) > 0$ are some constants.

See Appendix C.2.4 for the proof.

We can infer the following properties of the approximation error bounds.

- For the expansion regime, the quality of approximation improves almost linearly in $\|\boldsymbol{\mu}\|^2$, and the dominating bound vanishes as $\tau \to +\infty$;
- For the shrinkage regime, a sufficiently large dimension p implies a smaller approximation error, and when σ_{aug}^2 is adequately small, the error bound becomes tighter as σ_{aug}^2 increases.

Finite-sample regime. Now we comment on the difficulty when dealing with finite samples, as opposed to the infinite-sample regime we have been focusing on thus far. In the finite-sample regime, the loss function becomes

$$\begin{aligned} \mathcal{L}_n(\mathbf{W}) \coloneqq & -\frac{1}{\tau} \mathbb{E}_n \left\{ \mathbb{E}_{\boldsymbol{h}^+, \boldsymbol{h} | \boldsymbol{h}} \left[\text{sim}^*(\mathbf{W}\boldsymbol{h}, \mathbf{W}\boldsymbol{h}^+) \middle| \boldsymbol{h}_0 \right] \right\} \\ & + \log \left(\mathbb{E}_n \left\{ \mathbb{E}_{\boldsymbol{h}^-, \boldsymbol{h} | \boldsymbol{h}_0} \left[e^{\text{sim}^*(\mathbf{W}\boldsymbol{h}, \mathbf{W}\boldsymbol{h}^-)/\tau} \middle| \boldsymbol{h}_0 \right] \right\} \right), \end{aligned}$$

where $\mathbb{E}_n[a]$ denotes the expectation over the empirical distribution of $\{\boldsymbol{h}_{0,i}\}_{i\le n}$. To write $\mathcal{L}_n(\mathbf{W})$ as a more explicit function in $\mathbf{W}$, denote $\mathbf{M} = (\mathbf{I} + (\tau\alpha)^{-1}(1+\sigma_{\text{aug}}^2)\mathbf{W}^\top\mathbf{W})^{-1}$. Additionally, define

$$\begin{aligned} S_{\boldsymbol{h}_0} &\coloneqq \exp\left(-\frac{1}{2(1+2\sigma_{\text{aug}}^2)}(\boldsymbol{h}_0-\boldsymbol{\mu})^\top(\mathbf{I}-\mathbf{M})(\boldsymbol{h}_0-\boldsymbol{\mu})\right) \\ &\qquad + \exp\left(-\frac{1}{2(1+2\sigma_{\text{aug}}^2)}(\boldsymbol{h}_0+\boldsymbol{\mu})^\top(\mathbf{I}-\mathbf{M})(\boldsymbol{h}_0+\boldsymbol{\mu})\right), \\ \widetilde{S}_{\boldsymbol{h}_0} &\coloneqq \exp\left(-\frac{1}{2\tau\alpha}\|\mathbf{W}(\boldsymbol{h}_0-\boldsymbol{\mu})\|^2\right) + \exp\left(-\frac{1}{2\tau\alpha}\|\mathbf{W}(\boldsymbol{h}_0+\boldsymbol{\mu})\|^2\right), \end{aligned}$$

then we have the following proposition.

Proposition 4.4.5. *The finite sample loss $\mathcal{L}_n(\mathbf{W})$ can be written as*

$$\mathcal{L}_n(\mathbf{W}) = -\log 2 + \frac{\sigma_{\text{aug}}^2}{\tau\alpha}\|\mathbf{W}\|_{\text{F}}^2 + \frac{1}{2}\log\det[\mathbf{M}] + \log\left(\mathbb{E}_n S_{\boldsymbol{h}_0}\right).$$

Further, if $\alpha \gg \|\mathbf{W}\|_{\text{F}}^2$, $\mathcal{L}_n(\mathbf{W})$ has the following approximation

$$\widetilde{\mathcal{L}}_n(\mathbf{W}) = -\log 2 - \frac{1}{2\tau\alpha}\|\mathbf{W}\|_{\text{F}}^2 + \log\left(\mathbb{E}_n \widetilde{S}_{\boldsymbol{h}_0}\right).$$

From the proposition, the difficulty in analyzing the finite-sample loss lies in the nonlinearity in $\log\det[\mathbf{M}]$ as well as the terms $S_{\boldsymbol{h}_0}$. If we assume that $\alpha \gg \|\mathbf{W}\|_{\text{F}}^2$, the approximation is based on the fact that $\mathbf{M} \approx \mathbf{I} - (\tau\alpha)^{-1}(1+\sigma_{\text{aug}}^2)\mathbf{W}^\top\mathbf{W}$.

Remark 4.4.6. As $\boldsymbol{h}_0 \sim \mathcal{N}(y\cdot\boldsymbol{\mu}, \mathbf{I})$, if we replace $\mathbb{E}_n$ by $\mathbb{E}_{\boldsymbol{h}_0}$ in $\widetilde{\mathcal{L}}_n(\mathbf{W})$, then the loss can be approximated by

$$-\log 2 - \frac{1}{\tau\alpha}\|\mathbf{W}\|_{\text{F}}^2 + \log\left(1 + \exp\left(-\frac{2}{\tau\alpha}\|\mathbf{W}\boldsymbol{\mu}\|^2\right)\right),$$

which is exactly the loss function $\widetilde{\mathcal{L}}(\mathbf{W})$ defined in (4.3).

4.5 Effect of projectors on generalization

We move on to investigate the effect of projectors on the generalization of learned representations.

4.5.1 One-parameter projector model

To understand the generalization puzzle about projectors, an important empirical observation is that pretrained features typically form *linearly separable* clusters; see Figure 4.5. Adopting our 2-GMM view, we assume that cluster labels $\{y_i\}_{i\le n}$ are revealed to us on downstream tasks, and our data $\{\boldsymbol{h}_{0,i}, y_i\}_{i\le n}$ follow the model

$$\boldsymbol{h}_{0,i} \mid y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(y_i \cdot \boldsymbol{\mu}, \mathbf{I}_p), \quad \text{where} \quad y_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(\{-1, 1\}).$$

Assuming a linear projector as before, we denote $\boldsymbol{z}_i = \mathbf{W}\boldsymbol{h}_{0,i}$ and $\{\boldsymbol{z}_i, y_i\}_{i\le n}$ are the inputs of the classification problem.

In reality, the minimizer from the standard SimCLR loss (4.1) is not perfectly characterized by the ideal expansion/shrinkage; see the singular value plot in Figure 4.3. Nevertheless, to gain insights, we consider a simpler form of linear transforms:

$$\mathcal{W} = \big\{\mathbf{W}_\eta = \mathbf{I}_p + \eta\rho^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^\top : \eta > -1\big\}, \text{ where } \rho \coloneqq \|\boldsymbol{\mu}\|^2.$$

Since only the right singular vectors of $\mathbf{W}$ is of interest in terms of expansion/shrinkage, the symmetric projection head is without loss of generality. If we constrain ourselves to this simple one-parameter space $\mathcal{W}$, then expansion/shrinkage regimes are solely determined by η. Specifically, $\eta = 0$ means effectively no projector, $\eta > 0$ corresponds to the expansion

regime, while $\eta < 0$ corresponds to the shrinkage regime.

$$\begin{cases} \text{If } \eta \in (-1, 0), & \text{then shrinkage regime;} \\ \text{If } \eta \in (0, \infty), & \text{then expansion regime;} \\ \text{If } \eta = 0, & \text{effectively no projector.} \end{cases}$$

4.5.2 Confirming folklore: invariance of test errors in low-dimension regime

In the low-dimensional regime (i.e., when $p/n \to 0$), the data $(\boldsymbol{h}_{0,i}, y_i)_{i \le n}$ are not linearly separable with probability approaching one. We focus on ℓ_2-regularized logistic regression with an intercept in a fixed dimension p. Define

$$\ell_n(\gamma, \boldsymbol{\beta}; \lambda_n) = \mathbb{E}_n \left\{ \log \left[1 + e^{-y\left(\gamma + \boldsymbol{z}^\top \boldsymbol{\beta}\right)} \right] \right\} + \lambda_n \|\boldsymbol{\beta}\|^2, \tag{4.4}$$

where $\mathbb{E}_n$ denotes the sample average over n samples $i = 1, \ldots, n$. With coefficients γ, $\boldsymbol{\beta}$ and a given $\mathbf{W} = \mathbf{W}_\eta$, define the test error as $\texttt{Err}\,(\gamma, \boldsymbol{\beta}; \eta) \coloneqq \mathbb{P}(\gamma + \widetilde{y}\langle \widetilde{\boldsymbol{z}}, \boldsymbol{\beta} \rangle < 0)$, where $(\widetilde{\boldsymbol{z}}, \widetilde{y})$ is a new independent sample. We further define the expected error $\overline{\texttt{Err}}(\eta) = \mathbb{E}[\texttt{Err}(\widehat{\gamma}, \widehat{\boldsymbol{\beta}}; \eta)]$, where $\widehat{\gamma}, \widehat{\boldsymbol{\beta}}$ is the logistic regression estimator, and the expectation is taken w.r.t. the training samples. Intuitively, the linear transformation will not change the generalization error. The following result confirms this intuition by proving that the effects of expansion or shrinkage are negligible, unless we add an unreasonably large regularizer; see the second case below.

Theorem 4.5.1. *Consider minimizing the regularized logistic loss function* (4.4).

1. *If* $0 \le \lambda_n \ll \sqrt{n}$*, then the test error obeys*

$$\overline{\texttt{Err}}\,(\eta) = \Phi(-\|\boldsymbol{\mu}\|) + o(1).$$

   Here, the dominant term (i.e., the first term) in the test error remains the same for varying η*.*

2. *If* $\lambda_n \asymp \sqrt{n}$*, then* $\overline{\texttt{Err}}\,(\eta)$ *is decreasing in* η*.*

See Appendix C.3.1 for the proof.

4.5.3 High-dimensional regime: inductive bias matters

In search of the explanation, we then turn to the high-dimensional regime. In the high-dimensional regime, two distinct phenomena arise: first, $\{\boldsymbol{h}_{0,i}\}_{i\le n}$ are linearly separable with high probability. Second, solving logistic regression on linearly separable data via gradient descent results in a max-margin classifier (Rosset et al., 2003; Soudry et al., 2018)—a form of inductive bias (Neyshabur et al., 2014). Recall that $\boldsymbol{z}_i = \mathbf{W}\boldsymbol{h}_{0,i}$, and the max-margin classifier is given by

$$\max_{\boldsymbol{\beta}} \min_{i\le n} y_i\langle \boldsymbol{z}_i, \boldsymbol{\beta}\rangle \qquad \text{subject to } \|\boldsymbol{\beta}\| \le 1. \tag{4.5}$$

To understand the different inductive biases brought by different linear projectors, let us rewrite the max-margin problem above using the original input data: denoting $\widetilde{\boldsymbol{\beta}} = \mathbf{W}\boldsymbol{\beta}$,

the optimization problem (4.5) can be reformulated into

$$\begin{aligned} \max_{\widetilde{\boldsymbol{\beta}}} \quad & \min_{i\le n} y_i\langle \boldsymbol{h}_{0,i}, \widetilde{\boldsymbol{\beta}}\rangle \\ \text{subject to} \quad & \|\widetilde{\boldsymbol{\beta}}\|_{(\mathbf{W}\mathbf{W}^\top)^{1/2}} \le 1, \end{aligned}$$

where $\|\boldsymbol{u}\|_{\mathbf{A}} := \boldsymbol{u}^\top \mathbf{A}^{-1}\boldsymbol{u}$ is a norm associated with the positive definite matrix $\mathbf{A}$. Clearly the inductive bias by $\mathbf{W}$ manifests itself via the induced norm $\|\cdot\|_{(\mathbf{W}\mathbf{W}^\top)^{1/2}}$. Therefore, it is reasonable to expect that reparametrization due to expansion or shrinkage affects generalization in high dimensions. The question then boils down to provably demonstrate the effect of $\mathbf{W}$ on the downstream classification accuracy $\texttt{Err}(\boldsymbol{\beta};\eta) := \mathbb{P}(\widetilde{y}\langle\widetilde{\boldsymbol{z}},\boldsymbol{\beta}\rangle < 0)$ with new independent sample $(\widetilde{z},\widetilde{y})$.

4.5.4 Precise asymptotic characterization in high dimension

Before presenting the main result, let us define some useful quantities. The solution is denoted by $\widehat{\boldsymbol{\beta}}$ (unique in the separable case). Second, define $\mathbf{P}_{\boldsymbol{\mu}} := \rho^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^\top$, $\mathbf{P}_{\boldsymbol{\mu}}^{\perp} := \mathbf{I}_p - \rho^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^\top$, the ratio $\widehat{u} := \|\mathbf{P}_{\boldsymbol{\mu}}^{\perp}\mathbf{W}_\eta\widehat{\boldsymbol{\beta}}\|/\|\mathbf{P}_{\boldsymbol{\mu}}\mathbf{W}_\eta\widehat{\boldsymbol{\beta}}\|$. We have the following theoretical guarantees.

Theorem 4.5.2. *Suppose that $n/p \to \delta > 0$, $\mathbf{W}_\eta \in \mathcal{W}$, and $\|\boldsymbol{\mu}\|$ is a constant. There exists $\delta^*(\rho) > 0$ such that the following holds.*

1. *(non-separability) If $\delta > \delta^*(\rho)$, then with probability approaching one, $(\boldsymbol{h}_i, y_i)_{i\le n}$ is not linearly separable and $\widehat{\boldsymbol{\beta}} = \mathbf{0}$.*

2. *(separability) If $\delta < \delta^*(\rho)$, then with probability approaching one, there exists a unique solution $\widehat{\boldsymbol{\beta}}$ to (4.5) with the margin*

$$\widehat{\kappa} = \min_{i\le n} y_i \langle \boldsymbol{z}_i, \widehat{\boldsymbol{\beta}} \rangle \xrightarrow{p} \kappa^*(\rho, \eta) > 0.$$

3. *(monotonicity of error) If $\delta < \delta^*(\rho)$, $\widehat{u} \xrightarrow{p} u^*(\rho, \eta)$ and $u^*(\rho, \eta)$ is monotonically decreasing in η. Moreover, the test error obeys*

$$\texttt{Err}\left(\widehat{\boldsymbol{\beta}}; \eta\right) \xrightarrow{p} \Phi\left(-\rho^{1/2}\sqrt{\frac{1}{1+[u^*(\rho,\eta)]^2}}\right),$$

where Φ denotes Gaussian CDF. Thus, the asymptotic test error is decreasing in η.

See Appendix C.3.2 for the proof, and the precise definitions of κ^* and u^*.

While separability thresholds (claims 1 and 2) are known in similar models (Deng et al., 2022), the third claim (the most interesting and important one) shows that in the high-dimensional regime, even a linear projector can change the test accuracy of a "linear" max-margin classifier, and the test accuracy increases with the expansion strength η. This partially explains the puzzling effect of projectors on downstream performance. We remark in passing that this result is based on a recent technique known as the convex Gaussian minimax theorem (Gordon, 1988; Thrampoulidis et al., 2015).

Numerical evidence. In Figure 4.7, we use the 2-GMM with separation parameter $\rho \in \{1, 2, 3, 5\}$ to generate a dataset $(\boldsymbol{h}_{0,i})_{i\le n}$ of size $n = 1000$ and dimension $p \in \{200, 600, 2000\}$.

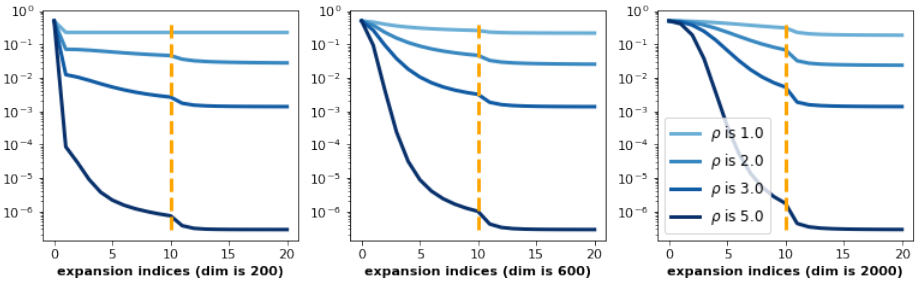


Figure 4.7: Test error is monotone in η under the one-parameter projector model. We apply a linear transform $\mathbf{W} \in \mathcal{W}$ on train data from 2-GMM, and calculate the test errors of the max-margin classifiers. Note $\eta = 0$ (vertical dashed) is equivalent to no transform.

We apply $\mathbf{W} = \mathbf{I}_p + \eta\rho^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^\top \in \mathcal{W}$ to obtain embeddings $(\boldsymbol{z}_i)_{i\le n}$. We treat the mixture membership as labels $(y_i)_{i\le n}$ and compute the max-margin classifier on $(\boldsymbol{z}_i, y_i)_{i\le n}$. For varying ρ, p, η, we report the test error of the max-margin classifier and confirm the monotone decreasing property.

4.6 Extensions to inhomogeneous feature augmentation

So far our analysis applies to scenarios where *either* expansion *or* shrinkage appears. However, in practice, it is not uncommon to encounter situations where both expansion and shrinkage appear. This does not render our previous analysis vacuous. As we will demonstrate in this section, we can extend the previous analysis to the case with inhomogeneous augmentations, under which both expansion and shrinkage can appear. Our treatment throughout this section parallels that in Section 4.4.

4.6.1 Feature augmentation with spiked covariance

Suppose that features are generated from the same 2-GMM but augmentations are inhomogeneous:

$$\boldsymbol{h}_{0,i} \overset{\text{i.i.d.}}{\sim} \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_p) + \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_p), \quad \text{and} \quad \boldsymbol{h}_i, \boldsymbol{h}_i^+ \,|\, \boldsymbol{h}_{0,i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{h}_{0,i}, \sigma_{\text{aug}}^2 \mathbf{A}),$$

where $\mathbf{A} \succeq \mathbf{I}_p$ is covariance matrix. Throughout this section, we assume $p > 2$. This inhomogeneous model is supported by empirical evidence. In the appendix, we show that image-level augmentation (random cropping, color distortion, etc.) does produce inhomogeneous features, which lead to more realistic phenomena.

Now following the same setup as before: we consider linear projectors $\boldsymbol{z} = \mathbf{W}\boldsymbol{h}$, the modified SimCLR loss (4.2), the population counterpart $\mathcal{L}(\mathbf{W}) = \mathbb{E}_{\boldsymbol{z}}\mathcal{L}_n(\mathbf{W})$ and its approximation $\widetilde{\mathcal{L}}(\mathbf{W})$. To ease the notation, let $\widetilde{\mathbf{W}} = (\mathbf{I} + \sigma_{\text{aug}}^2)^{1/2}\mathbf{W}$, $\widetilde{\boldsymbol{\mu}} = (\mathbf{I} + \sigma_{\text{aug}}^2)^{-1/2}\boldsymbol{\mu}$ and $\sum_{j=1}^p \widetilde{\sigma}_j^2 \widetilde{\boldsymbol{u}}_j^\top \widetilde{\boldsymbol{v}}_j$ be the SVD of $\widetilde{\mathbf{W}}$. The following proposition gives the explicit formulas of $\mathcal{L}(\mathbf{W})$ and $\widetilde{\mathcal{L}}(\mathbf{W})$.

Proposition 4.6.1. *Define*

$$\widetilde{\alpha} := \widetilde{\alpha}(\mathbf{W}) = \mathbb{E}\|\mathbf{W}\boldsymbol{h}\|^2 = \|\mathbf{W}\|_{\text{F}}^2 + \|\mathbf{W}\boldsymbol{\mu}\|^2 + \sigma_{\text{aug}}^2 \text{Tr}\left(\mathbf{W}^\top \mathbf{A}\mathbf{W}\right).$$

The loss $\mathcal{L}(\mathbf{W})$ takes the form $\mathcal{L}(\mathbf{W}) = \mathcal{L}_{\text{align}}(\mathbf{W}) + \mathcal{L}_{\text{unif}}(\mathbf{W})$, where

$$\begin{aligned}
\mathcal{L}_{\text{align}}(\mathbf{W}) &= -\frac{1}{\tau\widetilde{\alpha}}\|\mathbf{W}\|_{\text{F}}^2 \\
\mathcal{L}_{\text{unif}}(\mathbf{W}) &= -\frac{1}{2}\sum_{j=1}^p \log\left(1 + \frac{2\widetilde{\sigma}_j^2}{\tau\widetilde{\alpha}}\right) + \log\left(1 + \exp\Big(-\sum_{j=1}^p \frac{2\widetilde{\sigma}_j^2\langle\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{v}}_j\rangle^2}{2\widetilde{\sigma}_j^2 + \tau\widetilde{\alpha}}\Big)\right) - \log 2.
\end{aligned}$$

Similar to before, one can consider a first-order approximation $\widetilde{\mathcal{L}}(\mathbf{W})$:

$$\widetilde{\mathcal{L}}(\mathbf{W}) = -\frac{1}{\tau\widetilde{\alpha}}\|\mathbf{W}\|_{\text{F}}^2 + \log\left(1 + \exp\left(-\frac{2\|\mathbf{W}\boldsymbol{\mu}\|^2}{\tau\widetilde{\alpha}}\right)\right) - \log 2. \tag{4.6}$$

The loss $\widetilde{\mathcal{L}}(\mathbf{W})$ depends on $\mathbf{W}$ through $\|\mathbf{W}\|$, $\|\mathbf{W}\boldsymbol{\mu}\|$ and $\|\mathbf{W}\boldsymbol{v}_{\text{aug}}\|$. We further consider a simple one-spike covariance model

$$\mathbf{A} = \mathbf{I}_p + \rho_{\text{aug}}\boldsymbol{v}_{\text{aug}}\boldsymbol{v}_{\text{aug}}^\top, \qquad \text{where } \|\boldsymbol{v}_{\text{aug}}\| = 1\,. \tag{4.7}$$

Here $\rho_{\text{aug}} \geq 0$ quantifies the strength of the spike in augmentations. In particular, setting $\rho_{\text{aug}} = 0$ recovers the homogeneous case. It is natural to expect that $\text{span}\{\boldsymbol{\mu}, \boldsymbol{v}_{\text{aug}}\}$ plays a critical role in its minimization. For that purpose, let us define an orthogonal basis in $\text{span}\{\boldsymbol{\mu}, \boldsymbol{v}_{\text{aug}}\}$. Let $\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$, $r = \langle \bar{\boldsymbol{\mu}}, \boldsymbol{v}_{\text{aug}} \rangle$, and $\boldsymbol{\mu}_\perp$ be a vector such that $\|\boldsymbol{\mu}_\perp\| = 1$, $\langle \boldsymbol{\mu}_\perp, \boldsymbol{\mu} \rangle = 0$ and $r\langle \boldsymbol{\mu}_\perp, \boldsymbol{v}_{\text{aug}} \rangle \geq 0$.

Theoretical prediction. On the surface, the loss $\widetilde{\mathcal{L}}(\mathbf{W})$ here is not a univariate function. It turns out, as our proof reveals, that minimization of $\widetilde{\mathcal{L}}(\mathbf{W})$ is equivalent to a univariate minimization problem that depends only on $T = \|\mathbf{W}\bar{\boldsymbol{\mu}}\|^2/\|\mathbf{W}\|_{\text{F}}^2$. Define the threshold

$$\tau_1^* = \frac{2(1-r^2)\|\boldsymbol{\mu}\|^2}{\log(1+2\sigma_{\text{aug}}^2)(1+\sigma_{\text{aug}}^2+(1-r^2)\|\boldsymbol{\mu}\|^2)}\,.$$

We have the following result that characterizes the phase transition of the minimizer $\mathbf{W}^*$.

Theorem 4.6.2. *(phase transition under spiked augmentation) Consider the one-spike inhomogeneous model* (4.7) *and the approximate loss* (4.6)*. Recall the definition* $T := T(\mathbf{W}) = \|\mathbf{W}\bar{\boldsymbol{\mu}}\|^2/\|\mathbf{W}\|_{\mathrm{F}}^2$*. Let* $\mathbf{W}^*$ *be a minimizer of* $\widetilde{\mathcal{L}}(\mathbf{W})$ *and* $T^* = T(\mathbf{W}^*)$*. Then,* T^* *is given by the minimization problem*

$$\min_{T\in[0,1]} \quad -\frac{1}{\tau\left[(1+\sigma_{\mathrm{aug}}^2)+\|\boldsymbol{\mu}\|^2 T+\rho_{\mathrm{aug}}\sigma_{\mathrm{aug}}^2[(r\sqrt{T}-\sqrt{1-r^2}\,\sqrt{1-T})_+]^2\right]} + \log\left(1+\exp\left(-\frac{2\|\boldsymbol{\mu}\|^2 T}{\tau\left[(1+\sigma_{\mathrm{aug}}^2)+\|\boldsymbol{\mu}\|^2 T+\rho_{\mathrm{aug}}\sigma_{\mathrm{aug}}^2[(r\sqrt{T}-\sqrt{1-r^2}\,\sqrt{1-T})_+]^2\right]}\right)\right).$$

where x_+ *denotes* $\max\{x,0\}$*. Under* $\rho_{\mathrm{aug}} > 0$ *and* $0 < |r| < 1$*, we have*

- *if* $\tau > \tau_1^*$*, then the minimizer is attained at* $T^* \in (1-r^2, 1)$*, which is associated with a rank-one projector* $\mathbf{W}^*$ *given by*

$$(\mathbf{W}^*)^\top \mathbf{W}^* = (\sqrt{T^*}\bar{\boldsymbol{\mu}} - \sqrt{1-T^*}\boldsymbol{\mu}_\perp)(\sqrt{T^*}\bar{\boldsymbol{\mu}} - \sqrt{1-T^*}\boldsymbol{\mu}_\perp)^\top;$$

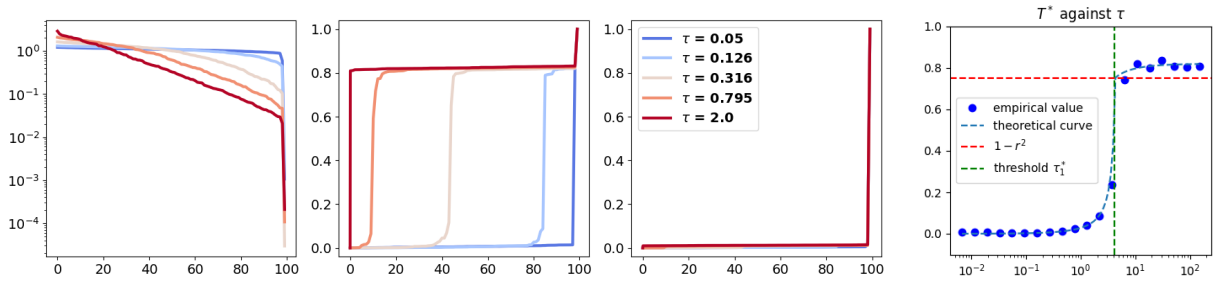
- *if* $\tau \le \tau_1^*$*, the minimizer is attained at* $T^* = \frac{\tau(1+\sigma_{\mathrm{aug}}^2)}{2\|\boldsymbol{\mu}\|^2}\log(1+2\sigma_{\mathrm{aug}}^2)\left[1-\frac{\tau}{2}\log(1+2\sigma_{\mathrm{aug}}^2)\right]^{-1}$ *and* $\mathbf{W}^*\boldsymbol{v}_{\mathrm{aug}} = 0$*.*

See Appendix C.4.2 for the proof.

If $\rho_{\mathrm{aug}} = 0$ in the optimization problem of Theorem 4.6.2, then the analysis reduces to the homogeneous case. However, with *any small* spike strength $\rho_{\mathrm{aug}} > 0$ and assuming nondegeneracy ($0 < |r| < 1$), the phase transition becomes qualitatively very different.

- The phase threshold is smaller as $\tau_1^* \le \tau^*$, which we recall that

$$\tau^* = 2\|\boldsymbol{\mu}\|^2\left\{(1+\sigma_{\mathrm{aug}}^2+\|\boldsymbol{\mu}\|^2)\log(1+2\sigma_{\mathrm{aug}}^2)\right\}^{-1}.$$



(a) (b)

Figure 4.8: GMM with inhomogeneous augmentation. (a) **Left**: $p = 100$ singular values of $\mathbf{W}$ in descending order with varying τ indicated by different colors. (a) **Middle**: Cumulative score for $\boldsymbol{\mu}$, i.e., $\text{score}_i(\boldsymbol{\mu}) = \sum_{j \le i} \langle \boldsymbol{\mu}/\|\boldsymbol{\mu}\|, \boldsymbol{v}_j \rangle^2$. (a) **Right**: Cumulative score for $\boldsymbol{v}_{\text{aug}}$, i.e., $\text{score}_i(\boldsymbol{v}_{\text{aug}}) = \sum_{j \le i} \langle \boldsymbol{v}_{\text{aug}}, \boldsymbol{v}_j \rangle^2$. (b) Phase transition showing T^* against τ: theoretical prediction (dashed curve) versus empirical values (circles).

The difference depends on the cosine angle between the signal $\bar{\boldsymbol{\mu}}$ and spike direction $\boldsymbol{v}_{\text{aug}}$, *irrespective of the spike strength* ρ_{aug}.

- There is no perfect expansion along the signal direction $\bar{\boldsymbol{\mu}}$, as $T^* < 1$ is *always true*.

Degenerate cases. It is beneficial to consider two degenerate examples: $r = 0$ and $r = 1$. When $r = 0$, the signal direction and spike direction are orthogonal; and when $r = 1$, they are perfectly aligned.

Analyzing the optimization problem in Theorem 4.6.2 for the degenerate examples yields the following characterization.

1. When $r = 0$ (namely $\bar{\boldsymbol{\mu}} \perp \boldsymbol{v}_{\text{aug}}$), we have $\tau_1^* = \tau^*$, and $\mathbf{W}^* \boldsymbol{v}_{\text{aug}} = 0$, namely pure shrinkage along $\boldsymbol{v}_{\text{aug}}$. Moreover, similar to the homogeneous case, we have shrinkage along $\bar{\boldsymbol{\mu}}$ if $\tau \le \tau^*$ and expansion along $\bar{\boldsymbol{\mu}}$ if $\tau > \tau^*$.

2. When $|r| = 1$ (namely $\bar{\boldsymbol{\mu}} \parallel \boldsymbol{v}_{\text{aug}}$), we have expansion along $\bar{\boldsymbol{\mu}}$ if

$$\tau \cdot \log\left(\frac{2(1+\sigma_{\text{aug}}^2)\|\boldsymbol{\mu}\|^2}{\|\boldsymbol{\mu}\|^2 + \rho_{\text{aug}}\sigma_{\text{aug}}^2} - 1 \right) > \frac{2\|\boldsymbol{\mu}\|^2}{1 + \sigma^2 + \|\boldsymbol{\mu}\|^2 + \rho_{\text{aug}}\sigma_{\text{aug}}^2} .$$

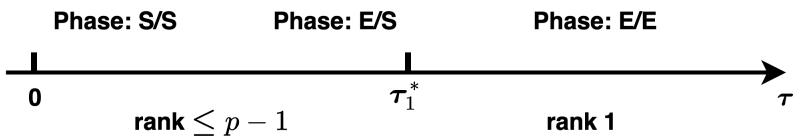


Figure 4.9: An illustration showing the phase changes in the one-spike inhomogeneous model as we increase τ. In the figure 'E' is a shorthand for expansion and 'S' for shrinkage; see Corollary 4.6.3.

In the first example, expansion/shrinkage operates independently along the two directions $\bar{\boldsymbol{\mu}}$ and $\boldsymbol{v}_{\text{aug}}$. In the second example, the presence of a parallel spike produces a threshold that is nonlinear in ρ_{aug} for expansion/shrinkage along $\bar{\boldsymbol{\mu}}$.

Numerical experiments. Figure 4.8a shows the singular vectors of $\mathbf{W}^*$ after training with the SimCLR loss and also the cumulative scores with $\bar{\boldsymbol{\mu}}$ and $\boldsymbol{v}_{\text{aug}}$, respectively. We vary the temperature τ, and fix $\rho_{\text{aug}} = 5$, $\sigma_{\text{aug}} = 0.5$, and $r = \langle \bar{\boldsymbol{\mu}}, \boldsymbol{v}_{\text{aug}} \rangle = 0.5$. When $\tau \leq 2$, there is consistent shrinkage in the direction of $\boldsymbol{v}_{\text{aug}}$. In addition, for example, when $\tau = 2$ or 0.795, expansion and shrinkage in the direction of $\boldsymbol{\mu}$ coexist and $\boldsymbol{\mu}$ is only spanned by either the top singular vector or the bottom singular vector of $\mathbf{W}^*$, which is consistent with Figure 4.3.

4.6.2 Simultaneous expansion/shrinkage

The signal $\boldsymbol{\mu}$ and spike $\boldsymbol{v}_{\text{aug}}$ exhibit different levels of expansion/shrinkage at the same temperature parameter, which sometimes leads to simultaneous expansion/shrinkage.

Consider the one-spike inhomogeneous model (4.7) and the approximate loss (4.6). Assume nondegeneracy $\rho_{\text{aug}} > 0$ and $0 < |r| < 1$. Recall the SVD of $\mathbf{W}^*$ is $\sum_{j \leq p} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^\top$. Without loss of generality we assume $\|\mathbf{W}^*\|_{\text{F}} = 1$.

Corollary 4.6.3 (Phase change under varying τ)**.** *As we increase the temperature parameter τ, treating other parameters $\|\boldsymbol{\mu}\|, \sigma_{\text{aug}}^2, \rho_{\text{aug}}, r$ as constants, we experience the following different phases.*

1. *When $\tau \ll 1$: both shrinkage. We have $\mathbf{W}^* \boldsymbol{v}_{\text{aug}} = 0$, and also $\|\mathbf{W}^* \bar{\boldsymbol{\mu}}\| \ll 1$, the latter of which implies $\max_{1 \le j \le p} |\sigma_j \langle \boldsymbol{v}_j, \bar{\boldsymbol{\mu}} \rangle| \ll 1$.*

2. *When $\omega(1) \le \tau \le \tau_1^*$: simultaneous expansion and shrinkage. We have $\mathbf{W}^* \boldsymbol{v}_{\text{aug}} = 0$, and also $\|\mathbf{W}^* \bar{\boldsymbol{\mu}}\| \ge c$ for certain constant $c > 0$. There are two jumps in the cumulative score: for certain dimension-free constant $\epsilon \in (0, 1)$, we have $\sum_{j \le \epsilon^{-1}} \langle \boldsymbol{v}_j, \bar{\boldsymbol{\mu}} \rangle^2 \ge \epsilon$ and $\langle \boldsymbol{v}_p, \bar{\boldsymbol{\mu}} \rangle^2 \ge \epsilon$.*

3. *When $O(1) \ge \tau > \tau_1^*$: both expansion. $\mathbf{W}^*$ is a rank-one matrix, and its right singular vector $\sqrt{T^*} \bar{\boldsymbol{\mu}} - \sqrt{1 - T^*} \boldsymbol{\mu}_\perp$ has positive cosine angles with both $\bar{\boldsymbol{\mu}}$ and $\boldsymbol{v}_{\text{aug}}$.*

4. *When $\tau \gg 1$: expansion increasingly aligns with $\bar{\boldsymbol{\mu}}$. Note that $\sqrt{T^*}$ is very close to 1 but always strictly smaller than 1.*

See Appendix C.4.3 for the proof.

Figure 4.9 gives an illustration of the phase change in the above corollary. It offers a theoretical explanation for the singular value/vector plots for STL-10 data (Figure 4.3) and simulated GMM data (Figure 4.8a).

4.7 Discussion and related work

4.7.1 Connections to emerging empirical phenomena

The puzzles about projectors in contrastive learning echo several known phenomena in deep learning.

Dimensional collapse. It is often observed that the trained features and embeddings do not span the entire ambient space. To be more precise, the singular values of the feature matrix $\mathbf{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n]^\top$ and the trained embedding matrix $\mathbf{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n]^\top$ (with $\boldsymbol{h}_i = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ and $\boldsymbol{z}_i = \boldsymbol{g}_{\boldsymbol{\varphi}}(\boldsymbol{h}_i)$) contain one or more (approximate) zeros. This phenomenon is known as *dimensional collapse*, which has been repeatedly reported in the literature (Chen et al., 2020a; Jing et al., 2021; Balestriero and LeCun, 2022).[5] When dimensional collapse occurs in the feature space, we obtain less informative representations. This is generally undesirable according to these papers and requires careful handling due to its adverse effects on generalization. Figure 4.3 (left) confirms dimensional collapse by showing that the linear projector does not have full rank.

To address dimensional collapse in contrastive learning (and more so in non-contrastive SSL), a line of work proposes to refine loss functions and design structured projectors (Balestriero et al., 2023), but a systematic treatment is still lacking.

Transferability of intermediate-layer features. In supervised learning, it is well observed that trained deep neural networks contain interpretable features that become progressively complex when moving up layers (Zeiler and Fergus, 2014). Therefore, it is natural to use intermediate-layer features pretrained on large datasets for related tasks (Yosinski et al., 2014). At the very top layers, features are believed to be very specific to a classification task, and thus they need to be finetuned on downstream tasks.

Intuitively, projectors bear similarity to those top layers that require finetuning on downstream tasks. For both supervised learning and contrastive learning, minimizing a specific loss seems to reduce the information in features and thus their generality.

5. In fact, dimensional collapse is a more salient issue for non-contrastive approaches in SSL (Hua et al., 2021; Tian et al., 2021) due to the lack of negative pairs.

Neural collapse. In supervised learning, the features in the penultimate layer tend to form a symmetric structure, if we train neural networks many epochs well past the terminal phase where the train error achieves zero. This phenomenon is known as *neural collapse* (Papyan et al., 2020). In short, as training evolves, the penultimate features gradually collapse to their respective class means, which form an equiangular simplex. The highly symmetric and compact cluster structure is observed on the training dataset and rarely on the test dataset (unless the test error is also zero).

Contrastive learning produces weaker cluster structure but the learned features are more general for downstream tasks (Wang and Isola, 2020). The cluster structure resulting from contrastive learning is desirable for generalization, whereas the strong cluster structure resulting from cross-entropy minimization is partly due to the optimization artifact. In fact, a fine-grained intermediate-layer neural collapse suggests that top layers (including the penultimate layer) do not improve and sometimes even harm the generalization properties of features (Galanti et al., 2022).

4.7.2 Related work

Analysis of contrastive learning. Contrastive learning has received tremendous attention in the past few years, and a large body of work has been done around this topic. We refer interested readers to the recent overview (Balestriero et al., 2023) for the historical account and for the recent updates. Though empirically successful, contrastive learning also brings various intriguing phenomena including dimensional collapse (Jing et al., 2021; Hua et al., 2021) and behavior of projectors (Chen et al., 2020a,b; Cosentino et al., 2022). These motivate quite a few recent theoretical attempts to explain the success of contrastive learning. Most notable and related is the paper by Wang and Isola (2020), where they discover that contrastive loss promotes both alignment and uniformity of the learned representations. This viewpoint is instrumental in our analysis and understanding, as can be seen from e.g.,

Proposition 4.4.1. Our work goes beyond alignment/uniformity by precisely characterizing the effect (namely, expansion and shrinkage) of contrastive loss on the projection head. Jing et al. (2021) also studies the role of projectors by arguing that they prevent dimensional collapse in the representation space. However, no theoretical study of the generalization property is provided. In addition, several recent papers (HaoChen et al., 2021; Wen and Li, 2021; Ji et al., 2021; Lee et al., 2021; Wen and Li, 2022; Von Kügelgen et al., 2021; Saunshi et al., 2022) theoretically study contrastive learning without focusing on the role of projectors.

Implicit bias and interpolating models. It is recently discovered that in over-parametrized models (i.e., interpolating models), gradient descent (GD) type algorithms have implicit regularization effects on the model parameters. Relevant to our paper are the results of Soudry et al. (2018) and Gunasekar et al. (2018), where it is proved that for linearly separable data, GD iterates converge in direction to the *max-margin classifier*. It is also important to characterize the generalization error of the solutions with implicit bias. A line of relevant papers include Belkin et al. (2019); Bartlett et al. (2020); Hastie et al. (2022); Liang and Rakhlin (2020); Bartlett et al. (2020); Hastie et al. (2022); Montanari et al. (2019); Montanari and Zhong (2022); Deng et al. (2022); Mei and Montanari (2022); Liang and Sur (2022); Montanari et al. (2021). Closely related to our Section 4.5 is the recent paper by Deng et al. (2022), but their main purpose is explaining the double-descent phenomenon rather than studying the effects for expansion/shrinkage.

BIBLIOGRAPHY

Ai, J. and Ren, Z. (2024). Not all distributional shifts are equal: Fine-grained robust conformal inference. *arXiv preprint arXiv:2402.13042*.

Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142.

Amjad, M., Shah, D., and Shen, D. (2018). Robust synthetic control. *The Journal of Machine Learning Research*, 19(1):802–852.

Angelopoulos, A., Bates, S., Malik, J., and Jordan, M. I. (2020). Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*.

Angelopoulos, A. N., Barber, R. F., and Bates, S. (2024). Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*.

Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. (2022). Conformal risk control. *arXiv preprint arXiv:2208.02814*.

Arora, S., Cohen, N., Golowich, N., and Hu, W. (2018). A convergence analysis of gradient descent for deep linear neural networks. *arXiv preprint arXiv:1810.02281*.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730.

Bai, J. and Ng, S. (2021). Matrix completion, counterfactuals, and factor analysis of missing data. *Journal of the American Statistical Association*, 116(536):1746–1763.

Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y., et al. (2023). A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*.

Balestriero, R. and LeCun, Y. (2022). Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *arXiv preprint arXiv:2205.11508*.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2021). Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486 – 507.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2022). Conformal prediction beyond exchangeability. *arXiv preprint arXiv:2202.13415*.

Bardes, A., Ponce, J., and LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.

Bastani, H. (2021). Predicting with proxies: Transfer learning in high dimension. *Manag. Sci.*, 67:2964–2984.

Bauschke, H. and Combettes, P. (2019). Convex analysis and monotone operator theory in hilbert spaces, corrected printing.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Ben-David, S., Lu, T., Luu, T., and Pál, D. (2010). Impossibility theorems for domain adaptation. In *AISTATS*.

Ben-David, S. and Urner, R. (2012). On the hardness of domain adaptation and the utility of unlabeled target samples. In *ALT*.

Ben-Tal, A. and Nemirovski, A. (1998). Robust convex optimization. *Mathematics of operations research*, 23(4):769–805.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Berta, E., Bach, F., and Jordan, M. (2024). Classifier calibration with roc-regularized isotonic regression. In *International Conference on Artificial Intelligence and Statistics*, pages 1972–1980. PMLR.

Bietti, A., Bruna, J., Sanford, C., and Song, M. J. (2022). Learning single-index models with shallow neural networks. *arXiv preprint arXiv:2210.15651*.

Blanchet, J., Kang, Y., and Murthy, K. (2019). Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857.

Blanchet, J. and Murthy, K. (2019). Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600.

Blanchet, J. and Shapiro, A. (2023). Statistical limit theorems in distributionally robust optimization. *arXiv preprint arXiv:2303.14867*.

Böhm, J. N., Berens, P., and Kobak, D. (2022). Unsupervised visualization of image datasets using contrastive learning. *arXiv preprint arXiv:2210.09879*.

Bourlard, H. and Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4):291–294.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217.

Brunk, H. (1963). On an extension of the concept conditional expectation. *Proceedings of the American Mathematical Society*, 14(2):298–304.

Brunk, H. (1965). Conditional expectation given a σ-lattice and applications. *The Annals of Mathematical Statistics*, 36(5):1339–1350.

Brunk, H., Barlow, R. E., Bartholomew, D. J., and Bremner, J. M. (1972). Statistical inference under order restrictions.(the theory and application of isotonic regression). *International Statistical Review*, 41:395.

Brunk, H., Ewing, G., and Utz, W. (1957). Minimizing integrals in certain classes of monotone functions. *Pacific Journal of Mathematics*.

Cai, T. T., Liang, T., and Rakhlin, A. (2016). Geometric inference for general high-dimensional linear inverse problems. *The Annals of Statistics*, 44(4):1536 – 1563.

Cai, T. T. and Wei, H. (2019). Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *ArXiv*, abs/1906.02903.

Candès, E., Lei, L., and Ren, Z. (2023). Conformalized survival analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):24–45.

Candes, E. and Recht, B. (2012). Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119.

Candès, E. J., Lei, L., and Ren, Z. (2021). Conformalized survival analysis. *arXiv preprint arXiv:2103.09763*.

Candes, E. J. and Plan, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936.

Candès, E. J. and Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080.

Cao, Y. and Xie, Y. (2015). Categorical matrix completion. In *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 369–372. IEEE.

Carpentier, A., Klopp, O., Löffler, M., and Nickl, R. (2018). Adaptive confidence sets for matrix completion. *Bernoulli*, 24(4A):2429 – 2460.

Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. (2020). Robust validation: Confident predictions even when distributions shift. *arXiv preprint arXiv:2008.04267*.

Chatterjee, S. and Lafferty, J. (2019). Adaptive risk bounds in unimodal regression. *Bernoulli*.

Chattopadhyay, R., Fan, W., Davidson, I., Panchanathan, S., and Ye, J. (2013). Joint transfer and batch-mode active learning. In *ICML*.

Chen, M., Weinberger, K. Q., and Blitzer, J. (2011). Co-training for domain adaptation. In *NIPS*.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020a). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. E. (2020b). Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255.

Chen, W., Chun, K.-J., and Barber, R. F. (2018). Discretized conformal prediction for efficient distribution-free inference. *Stat*, 7(1):e173.

Chen, W., Wang, Z., Ha, W., and Barber, R. F. (2016). Trimmed conformal prediction for high-dimensional models. *arXiv preprint arXiv:1611.09933*.

Chen, Y., Chi, Y., Fan, J., Ma, C., and Yan, Y. (2020c). Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM journal on optimization*, 30(4):3098–3121.

Chen, Y., Fan, J., Ma, C., and Yan, Y. (2019). Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences*, 116(46):22931–22937.

Chen, Y. and Lei, J. (2024). De-biased two-sample u-statistics with application to conditional distribution testing. *arXiv preprint arXiv:2402.00164*.

Chen, Y., Li, C., Ouyang, J., and Xu, G. (2023). Statistical inference for noisy incomplete binary matrix. *Journal of Machine Learning Research*, 24(95):1–66.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2021). An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 116:1849 – 1864.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, 22(1):173–203.

Cortes, C., Mohri, M., Riley, M., and Rostamizadeh, A. (2008). Sample selection bias correction theory. *ArXiv*, abs/0805.2775.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4):547–553.

Cosentino, R., Sengupta, A., Avestimehr, S., Soltanolkotabi, M., Ortega, A., Willke, T., and Tepper, M. (2022). Toward a geometrical understanding of self-supervised contrastive learning. *arXiv preprint arXiv:2205.06926*.

Davenport, M. A., Plan, Y., Van Den Berg, E., and Wootters, M. (2014). 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223.

De Bartolomeis, P., Abad, J., Donhauser, K., and Yang, F. (2023). Hidden yet quantifiable: A lower bound for confounding strength using randomized trials. *arXiv preprint arXiv:2312.03871*.

Deng, H. and Zhang, C.-H. (2020). Isotonic regression in multi-dimensional spaces and graphs.

Deng, Z., Kammoun, A., and Thrampoulidis, C. (2022). A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11(2):435–495.

Ding, P. and VanderWeele, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology (Cambridge, Mass.)*, 27(3):368.

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

Duchi, J. and Namkoong, H. (2018). Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*.

Duchi, J. C., Glynn, P. W., and Namkoong, H. (2021). Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969.

Duchi, J. C., Hashimoto, T., and Namkoong, H. (2019). Distributionally robust losses against mixture covariate shifts. *Under review*, 2(1).

Dudley, R. M. (1967). The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330.

Durot, C. and Lopuhaä, H. P. (2018). Limit Theory in Monotone Function Estimation. *Statistical Science*, 33(4):547 – 567.

Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., and Zisserman, A. (2021). With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9568–9577.

Edwards, R. E. (2012). *Functional analysis: theory and applications*. Courier Corporation.

El Ghaoui, L. and Lebret, H. (1997). Robust solutions to least-squares problems with uncertain data. *SIAM Journal on matrix analysis and applications*, 18(4):1035–1064.

El Ghaoui, L., Oustry, F., and Lebret, H. (1998). Robust solutions to uncertain semidefinite programs. *SIAM Journal on Optimization*, 9(1):33–52.

Esfahani, P. M. and Kuhn, D. (2015). Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *arXiv preprint arXiv:1505.05116*.

Esteban-Pérez, A. and Morales, J. M. (2022). Partition-based distributionally robust optimization via optimal transport with order cone constraints. *4OR*, 20(3):465–497.

Farias, V., Li, A. A., and Peng, T. (2022). Uncertainty quantification for low-rank matrix completion with heterogeneous and sub-exponential noise. In *International Conference on Artificial Intelligence and Statistics*, pages 1179–1189. PMLR.

Fazel, M., Hindi, H., and Boyd, S. (2004). Rank minimization and applications in system theory. In *Proceedings of the 2004 American control conference*, volume 4, pages 3273–3278. IEEE.

Foygel, R. and Srebro, N. (2011). Concentration-based guarantees for low-rank matrix reconstruction. In Kakade, S. M. and von Luxburg, U., editors, *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 315–340, Budapest, Hungary. PMLR.

Galanti, T., Galanti, L., and Ben-Shaul, I. (2022). On the implicit bias towards minimal depth of deep neural networks. *arXiv preprint arXiv:2202.09028*.

Gao, F. and Wellner, J. A. (2007). Entropy estimate for high-dimensional monotonic functions. *Journal of Multivariate Analysis*, 98(9):1751–1764.

Ge, J., Tang, S., Fan, J., Ma, C., and Jin, C. (2023). Maximum likelihood estimation is all you need for well-specified covariate shift. *arXiv preprint arXiv:2311.15961*.

Ghojogh, B., Crowley, M., Karray, F., and Ghodsi, A. (2023). *Elements of Dimensionality Reduction and Manifold Learning*. Springer Nature.

Gordon, Y. (1988). On milman's inequality and random subspaces which escape through a mesh in rn. In *Geometric aspects of functional analysis*, pages 84–106. Springer.

Grenander, U. (1956). On the theory of mortality measurements. *Skandinavisk Aktuarietidskrift*, 39:1–55.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K. M., Schölkopf, B., Candela, Q., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2009). Covariate shift by kernel mean matching. In *NIPS 2009*.

Grotzinger, S. J. and Witzgall, C. (1984). Projections onto order simplexes. *Applied mathematics and Optimization*, 12(1):247–270.

Gui, Y., Barber, R., and Ma, C. (2023a). Conformalized matrix completion. *Advances in Neural Information Processing Systems*, 36:4820–4844.

Gui, Y., Barber, R., and Ma, C. (2024a). Conformalized matrix completion. *Advances in Neural Information Processing Systems*, 36.

Gui, Y., Barber, R. F., and Ma, C. (2024b). Distributionally robust risk evaluation with an isotonic constraint. *arXiv preprint arXiv:2407.06867*.

Gui, Y., Hore, R., Ren, Z., and Barber, R. F. (2023b). Conformalized survival analysis with adaptive cut-offs. *Biometrika*, page asad076.

Gui, Y., Ma, C., and Zhong, Y. (2023c). Unraveling projection heads in contrastive learning: Insights from expansion and shrinkage. *arXiv preprint arXiv:2306.03335*.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. (2018). Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR.

Han, Q., Wang, T., Chatterjee, S., and Samworth, R. J. (2019). Isotonic regression in general dimensions. *The Annals of Statistics*.

Han, X., Papyan, V., and Donoho, D. L. (2021). Neural collapse under mse loss: Proximity to and dynamics on the central path. *arXiv preprint arXiv:2106.02073*.

Hanneke, S. and Kpotufe, S. (2019). On the value of target data in transfer learning. In *NeurIPS*.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. (2021). Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011.

Hardy, G. H., Littlewood, J. E., and Pólya, G. (1952). *Inequalities*. Cambridge university press.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

He, X. and Shao, Q.-M. (2000). On parameters of increasing dimensions. *Journal of multivariate analysis*, 73(1):120–135.

Henzi, A., Ziegel, J. F., and Gneiting, T. (2021). Isotonic distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):963–993.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Hua, T., Wang, W., Xue, Z., Ren, S., Wang, Y., and Zhao, H. (2021). On feature decorrelation in self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9598–9608.

Huang, M. and Pimentel, S. D. (2024). Variance-based sensitivity analysis for weighting estimators results in more informative bounds. *Biometrika*, page asae040.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr.

Jain, P., Netrapalli, P., and Sanghavi, S. (2013). Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, 40(2):633–643.

Ji, W., Deng, Z., Nakada, R., Zou, J., and Zhang, L. (2021). The power of contrast for feature learning: A theoretical analysis. *arXiv preprint arXiv:2110.02473*.

Jin, Y., Ren, Z., and Candès, E. J. (2023). Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6):e2214889120.

Jin, Y., Ren, Z., and Zhou, Z. (2022). Sensitivity analysis under the f-sensitivity models: a distributional robustness perspective. *arXiv preprint arXiv:2203.04373*.

Jing, L., Vincent, P., LeCun, Y., and Tian, Y. (2021). Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*.

Keshavan, R. H., Montanari, A., and Oh, S. (2010). Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302 – 2329.

Lam, H. (2016). Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research*, 41(4):1248–1275.

Ledoux, M. and Talagrand, M. (2013). *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.

Lee, J. D., Lei, Q., Saunshi, N., and Zhuo, J. (2021). Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34:309–323.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Lei, L. and Candès, E. J. (2020). Conformal inference of counterfactuals and individual treatment effects. *arXiv preprint arXiv:2006.06138*.

Li, S., Cai, T. T., and Li, H. (2021). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Liang, T. and Rakhlin, A. (2020). Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347.

Liang, T. and Sur, P. (2022). A precise high-dimensional asymptotic theory for boosting and minimum-ℓ
$_1$-norm interpolated classifiers. *The Annals of Statistics*, 50:1669–1695.

Liu, J., Wu, J., Wang, T., Zou, H., Li, B., and Cui, P. (2023). Geometry-calibrated dro: Combating over-pessimism with free energy implications. *arXiv preprint arXiv:2311.05054*.

Ma, C., Pathak, R., and Wainwright, M. J. (2023). Optimally tackling covariate shift in rkhs-based nonparametric regression. *The Annals of Statistics*, 51(2):738–761.

Ma, C., Wang, K., Chi, Y., and Chen, Y. (2020). Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, 20:451–632.

Matzkin, R. L. (1991). Semiparametric estimation of monotone and concave utility functions for polychotomous choice models. *Econometrica: Journal of the Econometric Society*, pages 1315–1327.

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

Megginson, R. E. (2012). *An introduction to Banach space theory*, volume 183. Springer Science & Business Media.

Mei, S. and Montanari, A. (2022). The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766.

Misra, I. and van der Maaten, L. (2019). Self-supervised learning of pretext-invariant representations. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6706–6716.

Montanari, A., Ruan, F., Sohn, Y., and Yan, J. (2019). The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*.

Montanari, A. and Zhong, Y. (2022). The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *The Annals of Statistics*, 50(5):2816–2847.

Montanari, A., Zhong, Y., and Zhou, K. (2021). Tractability from overparametrization: The example of the negative perceptron. *arXiv preprint arXiv:2110.15824*.

Müller, R., Kornblith, S., and Hinton, G. E. (2019). When does label smoothing help? *Advances in neural information processing systems*, 32.

Nair, J., Wierman, A., and Zwart, B. (2022). *The fundamentals of heavy tails: Properties, emergence, and estimation*, volume 53. Cambridge University Press.

Namkoong, H. and Duchi, J. C. (2017). Variance-based regularization with convex objectives. *Advances in neural information processing systems*, 30.

Namkoong, H., Ma, Y., and Glynn, P. W. (2022). Minimax optimal estimation of stability under distribution shift. *arXiv preprint arXiv:2212.06338*.

Natarajan, N. and Dhillon, I. S. (2014). Inductive matrix completion for predicting gene–disease associations. *Bioinformatics*, 30(12):i60–i68.

Negahban, S. and Wainwright, M. J. (2012). Restricted strong convexity and weighted matrix

completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Neyshabur, B., Tomioka, R., and Srebro, N. (2014). In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*.

Niculescu-Mizil, A. and Caruana, R. A. (2012). Obtaining calibrated probabilities from boosting. *arXiv preprint arXiv:1207.1403*.

Nie, X., Imbens, G., and Wager, S. (2021). Covariate balancing sensitivity analysis for extrapolating randomized trials across locations. *arXiv preprint arXiv:2112.04723*.

Papyan, V., Han, X., and Donoho, D. L. (2020). Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663.

Pathak, R. and Ma, C. (2024). On the design-dependent suboptimality of the lasso. *arXiv preprint arXiv:2402.00382*.

Pathak, R., Ma, C., and Wainwright, M. (2022). A new similarity measure for covariate shift with applications to nonparametric regression. In *International Conference on Machine Learning*, pages 17517–17530. PMLR.

Pearson, K. (1901). Principal components analysis. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 6(2):559.

Popescu, I. (2007). Robust mean-covariance solutions for stochastic optimization. *Operations Research*, 55(1):98–112.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics*, pages 356–366.

Qiu, H., Dobriban, E., and Tchetgen Tchetgen, E. (2023). Prediction sets adaptive to unknown covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkad069.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Rao, B. P. (1969). Estimation of a unimodal density. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 23–36.

Redko, I., Morvant, E., Habrard, A., Sebban, M., and Bennani, Y. (2020). A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv: Learning*.

Rennie, J. D. and Srebro, N. (2005). Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 713–719.

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pages 547–562. University of California Press.

Romano, Y., Sesia, M., and Candes, E. (2020). Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26.

Rosset, S., Zhu, J., and Hastie, T. (2003). Margin maximizing loss functions. In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press.

Rothenhäusler, D. and Bühlmann, P. (2023). Distributionally robust and generalizable inference. *Statistical Science*, 38(4):527–542.

Sahoo, R., Lei, L., and Wager, S. (2022). Learning from a biased sample. *arXiv preprint arXiv:2209.01754*.

Saunshi, N., Ash, J., Goel, S., Misra, D., Zhang, C., Arora, S., Kakade, S., and Krishnamurthy, A. (2022). Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pages 19250–19286. PMLR.

Setlur, A., Dennis, D., Eysenbach, B., Raghunathan, A., Finn, C., Smith, V., and Levine, S. (2023). Bitrate-constrained dro: Beyond worst case robustness to unknown group shifts. *arXiv preprint arXiv:2302.02931*.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3).

Shafieezadeh Abadeh, S., Mohajerin Esfahani, P. M., and Kuhn, D. (2015). Distributionally robust logistic regression. *Advances in Neural Information Processing Systems*, 28.

Shapiro, A. (2017). Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275.

Shapiro, A. and Pichler, A. (2023). Conditional distributionally robust functionals. *Operations Research*.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. (2018). The implicit

bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878.

Srebro, N. and Shraibman, A. (2005). Rank, trace-norm and max-norm. In *Learning Theory: 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30, 2005. Proceedings 18*, pages 545–560. Springer.

Sun, R. and Luo, Z.-Q. (2016). Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637.

Thrampoulidis, C., Oymak, S., and Hassibi, B. (2015). Regularized linear regression: A precise analysis of the estimation error. In Grünwald, P., Hazan, E., and Kale, S., editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1683–1709, Paris, France. PMLR.

Tian, Y. (2022). Understanding deep contrastive learning via coordinate-wise optimization. In *Advances in Neural Information Processing Systems*.

Tian, Y., Chen, X., and Ganguli, S. (2021). Understanding self-supervised learning dynamics without contrastive pairs. *ArXiv*, abs/2102.06810.

Tian, Y. and Feng, Y. (2021). Transfer learning under high-dimensional generalized linear models. *ArXiv*, abs/2105.14328.

Tibshirani, R. J., Barber, R. F., Candès, E. J., and Ramdas, A. (2019). Conformal prediction under covariate shift. In *NeurIPS*.

van der Laan, L., Ulloa-Pérez, E., Carone, M., and Luedtke, A. (2023). Causal isotonic calibration for heterogeneous treatment effects. *arXiv preprint arXiv:2302.14011*.

Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. (2021). Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467.

Vovk, V. (2015). Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74:9–28.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*, volume 29. Springer.

Wang, L. (2011). Gee analysis of clustered binary data with diverging number of covariates. *The Annals of Statistics*.

Wang, T. and Isola, P. (2020). Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

Wang, Z., Bühlmann, P., and Guo, Z. (2023). Distributionally robust machine learning with multi-source data. *arXiv preprint arXiv:2309.02211*.

Weiss, A., Lancho, A., Bu, Y., and Wornell, G. W. (2023). A bilateral bound on the mean-square error for estimation in model mismatch. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 2655–2660. IEEE.

Wellner, J. et al. (2013). *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.

Wen, Z. and Li, Y. (2021). Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pages 11112–11122. PMLR.

Wen, Z. and Li, Y. (2022). The mechanism of prediction head in non-contrastive self-supervised learning. *arXiv preprint arXiv:2205.06226*.

Wieczorek, J. (2023). Design-based conformal prediction. *arXiv preprint arXiv:2303.01422*.

Xia, D. and Yuan, M. (2021). Statistical inferences of linear forms for noisy matrix completion. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(1):58–77.

Xiao, L., Bahri, Y., Sohl-Dickstein, J., Schoenholz, S., and Pennington, J. (2018). Dynamical isometry and a mean field theory of cnns: How to train 10,000-layer vanilla convolutional neural networks. In *International Conference on Machine Learning*, pages 5393–5402. PMLR.

Yadlowsky, S., Namkoong, H., Basu, S., Duchi, J., and Tian, L. (2018). Bounds on the conditional and average treatment effect with unobserved confounding factors. *arXiv preprint arXiv:1808.09521*.

Yang, F. and Barber, R. F. (2019). Contraction and uniform convergence of isotonic regression. *Electronic Journal of Statistics*.

Yang, L., Hanneke, S., and Carbonell, J. G. (2012). A theory of transfer learning with applications to active learning. *Machine Learning*, 90:161–189.

Yang, Y., Kuchibhotla, A. K., and Tchetgen Tchetgen, E. (2024). Doubly robust calibration of prediction sets under covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae009.

Yang, Y. and Ma, C. (2022). Optimal tuning-free convex relaxation for noisy matrix completion. *arXiv preprint arXiv:2207.05802*.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27.

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer.

Zhang, C.-H. (2002). Risk bounds in isotonic regression. *The Annals of Statistics*, 30(2):528–555.

Zhao, Q., Small, D. S., and Bhattacharya, B. B. (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(4):735–761.

Zhao, Y. and Udell, M. (2020). Matrix completion with quantified uncertainty through low rank gaussian copula. *Advances in Neural Information Processing Systems*, 33:20977–20988.

Zhou, X., Chen, B., Gui, Y., and Cheng, L. (2024). Conformal prediction: A data perspective. *arXiv preprint arXiv:2410.06494*.

APPENDIX A

APPENDIX FOR CHAPTER 2

A.1 Proofs

The section collects the proofs for the technical results in the main text. Section A.1.1 and A.1.2 are devoted to the proof of Lemma 2.3.1 and Theorem 2.3.2, respectively. In the end, Section A.1.3 provides proofs for the coverage gap bounds in the two examples in Section 2.3.3.

A.1.1 Proof of Lemma 2.3.1

Fix a set of locations $\mathcal{B} \coloneqq \{(i_1, j_1), \ldots, (i_{n_{\text{cal}}+1}, j_{n_{\text{cal}}+1})\}$, and an index $1 \leq m \leq n_{\text{cal}} + 1$. By the definition of conditional probability, we have

$$\begin{aligned}
&\mathbb{P}\left((i_*, j_*) = (i_m, j_m) \mid \mathcal{S}_{\text{cal}} \cup \{(i_*, j_*)\} = \mathcal{B},\ \mathcal{S}_{\text{train}} = \mathcal{S}_0\right) \\
=&\frac{\mathbb{P}\left((i_*, j_*) = (i_m, j_m), \mathcal{S}_{\text{cal}} = \mathcal{B}\backslash\{(i_m, j_m)\} \mid \mathcal{S}_{\text{train}} = \mathcal{S}_0,\ |\mathcal{S}| = n\right)}{\mathbb{P}\left(\mathcal{S}_{\text{cal}} \cup \{(i_*, j_*)\} = \mathcal{B} \mid \mathcal{S}_{\text{train}} = \mathcal{S}_0,\ |\mathcal{S}| = n\right)} \\
=&\frac{\mathbb{P}\left((i_*, j_*) = (i_m, j_m), \mathcal{S}_{\text{cal}} = \mathcal{B}\backslash\{(i_m, j_m)\} \mid \mathcal{S}_{\text{train}} = \mathcal{S}_0,\ |\mathcal{S}| = n\right)}{\sum_{l=1}^{n_{\text{cal}}+1} \mathbb{P}\left((i_*, j_*) = (i_l, j_l), \mathcal{S}_{\text{cal}} = \mathcal{B}\backslash\{(i_l, j_l)\} \mid \mathcal{S}_{\text{train}} = \mathcal{S}_0,\ |\mathcal{S}| = n\right)}.
\end{aligned}$$

It then boils down to computing $\mathbb{P}\left(\mathcal{S}_{\text{cal}} = \mathcal{S}_1, (i_*, j_*) = (i_l, j_l) \mid \mathcal{S}_{\text{train}} = \mathcal{S}_0, |\mathcal{S}| = n\right)$ for any fixed set $\mathcal{S}_1$ and fixed location $(i_l, j_l) \notin \mathcal{S}_0 \cup \mathcal{S}_1$. To this end, we have the following claim (we defer its proof to the end of this section)

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{S}_{\text{cal}} = \mathcal{S}_1, (i_*, j_*) = (i_l, j_l) \mid \mathcal{S}_{\text{train}} = \mathcal{S}_0, |\mathcal{S}| = n\right) \\
=\ &\frac{1}{d_1 d_2 - n} \cdot \frac{\prod_{(i,j)\in\mathcal{S}_1} \frac{p_{ij}}{1-p_{ij}}}{\sum_{\mathcal{A}\in\Omega_{\mathcal{S}_{\text{train}},n}} \prod_{(i',j')\in\mathcal{A}} \frac{p_{i'j'}}{1-p_{i'j'}}},
\end{aligned} \tag{A.1}$$

where $\Omega_{\mathcal{S}_{\text{train}},n} = \{\mathcal{A} \subseteq [d_1] \times [d_2] : |\mathcal{A}| = n - |\mathcal{S}_{\text{train}}|,\ \mathcal{A} \cap \mathcal{S}_{\text{train}} = \varnothing\}$. As a result, we obtain

$$\begin{aligned}
&\mathbb{P}\left((i_*, j_*) = (i_m, j_m) \mid \mathcal{S}_{\text{cal}} \cup \{(i_*, j_*)\} = \mathcal{B},\ \mathcal{S}_{\text{train}} = \mathcal{S}_0\right) \\
=&\frac{\prod_{(i,j)\in\mathcal{B}\setminus\{(i_m,j_m)\}} \frac{p_{ij}}{1-p_{ij}}}{\sum_{l=1}^{n_{\text{cal}}+1} \prod_{(i,j)\in\mathcal{B}\setminus\{(i_l,j_l)\}} \frac{p_{ij}}{1-p_{ij}}} \\
=&\frac{h_{i_m,j_m}}{\sum_{l=1}^{n_{\text{cal}}+1} h_{i_l,j_l}},
\end{aligned}$$

where $h_{ij} = (1 - p_{ij})/p_{ij}$. This finishes the proof.

Proof of Equation (A.1)**.** Denote $\mathbf{W}$ the matrix consisting of entries W_{ij} defined in Algorithm 1. The matrix $\mathbf{Z}$ consists of Z_{ij} which are the indicators of non-missingness, i.e. $Z_{ij} = \mathbb{1}\{(i, j) \text{ is observed}\}$. Fix any two disjoint subsets $\mathcal{S}_0, \mathcal{S}_1 \subseteq [d_1] \times [d_2]$ with $|\mathcal{S}_0| + |\mathcal{S}_1| = n$. We have

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{S}_{\text{cal}} = \mathcal{S}_1, \mathcal{S}_{\text{train}} = \mathcal{S}_0 \mid |\mathcal{S}| = n\right) \\
=&\mathbb{P}\left(\text{supp}(\mathbf{Z}) = \mathcal{S}_0 \cup \mathcal{S}_1, \mathcal{S}_0 \subseteq \text{supp}(\mathbf{W}), \mathcal{S}_1 \subseteq \text{supp}(\mathbf{W})^c \mid |\mathcal{S}| = n\right).
\end{aligned}$$

Since $\mathbf{W}$ and $\mathbf{Z}$ are independent, one further has

$$\begin{aligned}
&\mathbb{P}\left(\text{supp}(\mathbf{Z}) = \mathcal{S}_0 \cup \mathcal{S}_1, \mathcal{S}_0 \subseteq \text{supp}(\mathbf{W}), \mathcal{S}_1 \subseteq \text{supp}(\mathbf{W})^c \mid |\mathcal{S}| = n\right) \\
=&\mathbb{P}\left(\mathcal{S}_0 \subseteq \text{supp}(\mathbf{W}), \mathcal{S}_1 \subseteq \text{supp}(\mathbf{W})^c\right) \cdot \mathbb{P}\left(\text{supp}(\mathbf{Z}) = \mathcal{S}_0 \cup \mathcal{S}_1 \mid |\mathcal{S}| = n\right) \\
=&q^{|\mathcal{S}_0|}(1-q)^{|\mathcal{S}_1|} \cdot \mathbb{P}\left(\text{supp}(\mathbf{Z}) = \mathcal{S}_0 \cup \mathcal{S}_1 \mid |\mathcal{S}| = n\right) \\
=&\frac{q^{|\mathcal{S}_0|}(1-q)^{|\mathcal{S}_1|}}{\mathbb{P}(|\mathcal{S}| = n)} \prod_{(i',j')\in[d_1]\times[d_2]} (1 - p_{i'j'}) \prod_{(i,j)\in\mathcal{S}_0\cup\mathcal{S}_1} \frac{p_{ij}}{1-p_{ij}},
\end{aligned} \tag{A.2}$$

where the last two identities are based on direct computations.

Based on (A.2), one can further compute

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{S}_{\text{train}} = \mathcal{S}_0 \mid |\mathcal{S}| = n\right) \\
=&\sum_{\mathcal{A}\in\Omega_{\mathcal{S}_{\text{train}},n}} \mathbb{P}\left(\mathcal{S}_{\text{cal}} = \mathcal{A}, \mathcal{S}_{\text{train}} = \mathcal{S}_0 \mid |\mathcal{S}| = n\right) \\
=&\frac{q^{|\mathcal{S}_0|}(1-q)^{n-|\mathcal{S}_0|}}{\mathbb{P}(|\mathcal{S}| = n)} \sum_{\mathcal{A}\in\Omega_{\mathcal{S}_{\text{train}},n}} \prod_{(i',j')\in[d_1]\times[d_2]} (1-p_{i'j'}) \prod_{(i,j)\in\mathcal{S}_0\cup\mathcal{A}} \frac{p_{ij}}{1-p_{ij}}, \qquad \text{(A.3)}
\end{aligned}$$

where the last identity uses Equation (A.2).

Now we are ready to prove (A.1). Recall that the new data point $(i_*, j_*) \mid \mathcal{S}$ is drawn uniformly from $\text{Unif}(\mathcal{S}^c)$. Therefore one has

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{S}_{\text{cal}} = \mathcal{S}_1, (i_*, j_*) = (i_{n+1}, j_{n+1}) \mid \mathcal{S}_{\text{train}} = \mathcal{S}_0, |\mathcal{S}| = n\right) \\
=& \frac{\mathbb{P}\left(\mathcal{S}_{\text{cal}} = \mathcal{S}_1, \mathcal{S}_{\text{train}} = \mathcal{S}_0, (i_*, j_*) = (i_{n+1}, j_{n+1}) \mid |\mathcal{S}| = n\right)}{\mathbb{P}\left(\mathcal{S}_{\text{train}} = \mathcal{S}_0 \mid |\mathcal{S}| = n\right)} \\
=& \frac{\mathbb{P}\left((i_*, j_*) = (i_{n+1}, j_{n+1}) \mid \mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_0\right) \mathbb{P}\left(\mathcal{S}_{\text{cal}} = \mathcal{S}_1, \mathcal{S}_{\text{train}} = \mathcal{S}_0 \mid |\mathcal{S}| = n\right)}{\mathbb{P}\left(\mathcal{S}_{\text{train}} = \mathcal{S}_0 \mid |\mathcal{S}| = n\right)} \\
=& \frac{1}{d_1 d_2 - n} \cdot \frac{\prod_{(i,j)\in\mathcal{S}_1} \frac{p_{ij}}{1-p_{ij}}}{\sum_{\mathcal{A}\in\Omega_{\mathcal{S}_{\text{train}},n}} \prod_{(i',j')\in\mathcal{A}} \frac{p_{i'j'}}{1-p_{i'j'}}},
\end{aligned}$$

where the last line follows from (A.2) and (A.3).

A.1.2 Proof of Theorem 2.3.2

First, fix any $a \in [0, 1]$. Lemma 2.3.1, together with the weighted conformal prediction framework of Tibshirani et al. (2019), implies that

$$\mathbb{P}\left(M_{i_*j_*} \in \widehat{M}_{i_*j_*} \pm q^*_{i_*j_*}(a) \cdot \widehat{s}_{i_*j_*} \,\middle|\, \mathcal{S}_{\text{train}}, \mathcal{S}_{\text{cal}} \cup \{(i_*, j_*)\}\right) \geq 1 - a,$$

where

$$q^*_{i_*j_*}(a) = \text{Quantile}_{1-a}\left(\sum_{(i,j)\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} w_{ij}\cdot\delta_{R_{ij}}\right),$$

and

$$w_{ij} = \frac{h_{ij}}{\sum_{(i',j')\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} h_{i'j'}}.$$

Indeed, here a can be any function of the random variables we are conditioning on—that is, a may depend on n, on $\mathcal{S}_{\text{train}}$, and on $\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}$.

Next define $a = \alpha + \Delta$ where Δ is defined as in (2.6). We observe that, since each $\widehat{h}_{ij}$ is a function of $\mathcal{S}_{\text{train}}$, then Δ (and thus also a) can therefore be expressed as a function of n, $\mathcal{S}_{\text{train}}$, and $\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}$. Therefore, applying the work above, we have

$$\mathbb{P}\left(M_{i_*j_*}\in\widehat{M}_{i_*j_*}\pm q^*_{i_*j_*}(\alpha+\Delta)\cdot\widehat{s}_{i_*j_*}\,\middle|\,\mathcal{S}_{\text{train}},\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}\right)\geq 1-\alpha-\Delta$$

and thus, after marginalizing,

$$\mathbb{P}\left(M_{i_*j_*}\in\widehat{M}_{i_*j_*}\pm q^*_{i_*j_*}(\alpha+\Delta)\cdot\widehat{s}_{i_*j_*}\right)\geq 1-\alpha-\mathbb{E}[\Delta].$$

Next, we verify that

$$\widehat{q}\geq q^*_{i_*j_*}(\alpha+\Delta)$$

holds almost surely—if this is indeed the case, then we have shown that

$$\mathbb{P}\left(M_{i_*j_*}\in\widehat{M}_{i_*j_*}\pm\widehat{q}\cdot\widehat{s}_{i_*j_*}\right)\geq 1-\alpha-\mathbb{E}[\Delta],$$

which establishes the desired result. Thus we only need to show that $\widehat{q}\geq q^*_{i_*j_*}(\alpha+\Delta)$, or

equivalently,

$$\mathrm{Quantile}_{1-\alpha}\left(\sum_{(i,j)\in\mathcal{S}_{\mathrm{cal}}}\widehat{w}_{ij}\cdot\delta_{R_{ij}}+\widehat{w}_{\mathrm{test}}\cdot\delta_{+\infty}\right)\geq\mathrm{Quantile}_{1-\alpha-\Delta}\left(\sum_{(i,j)\in\mathcal{S}_{\mathrm{cal}}\cup\{(i_*,j_*)\}}w_{ij}\cdot\delta_{R_{ij}}\right).$$

Define

$$w'_{ij}=\frac{\widehat{h}_{ij}}{\sum_{(i',j')\in\mathcal{S}_{\mathrm{cal}}\cup\{(i_*,j_*)\}}\widehat{h}_{i'j'}} \tag{A.4}$$

for all $(i,j)\in\mathcal{S}\cup\{(i_*,j_*)\}$. Then by definition of $\widehat{\boldsymbol{w}}$, we see that $w'_{ij}\geq\widehat{w}_{ij}$ for $(i,j)\in\mathcal{S}$, and therefore,

$$\begin{aligned}&\mathrm{Quantile}_{1-\alpha}\left(\sum_{(i,j)\in\mathcal{S}_{\mathrm{cal}}}\widehat{w}_{ij}\cdot\delta_{R_{ij}}+\widehat{w}_{\mathrm{test}}\cdot\delta_{+\infty}\right)\\&\geq\mathrm{Quantile}_{1-\alpha}\left(\sum_{(i,j)\in\mathcal{S}_{\mathrm{cal}}}w'_{ij}\cdot\delta_{R_{ij}}+w'_{(i_*,j_*)}\cdot\delta_{+\infty}\right)\geq\mathrm{Quantile}_{1-\alpha}\left(\sum_{(i,j)\in\mathcal{S}_{\mathrm{cal}}\cup\{(i_*,j_*)\}}w'_{ij}\cdot\delta_{R_{ij}}\right)\end{aligned}$$

holds almost surely. Therefore it suffices to show that

$$\mathrm{Quantile}_{1-\alpha}\left(\sum_{(i,j)\in\mathcal{S}_{\mathrm{cal}}\cup\{(i_*,j_*)\}}w'_{ij}\cdot\delta_{R_{ij}}\right)\geq\mathrm{Quantile}_{1-\alpha-\Delta}\left(\sum_{(i,j)\in\mathcal{S}_{\mathrm{cal}}\cup\{(i_*,j_*)\}}w_{ij}\cdot\delta_{R_{ij}}\right)$$

holds almost surely. Indeed, we have

$$\begin{aligned}&\mathsf{d}_{\mathsf{TV}}\left(\sum_{(i,j)\in\mathcal{S}_{\mathrm{cal}}\cup\{(i_*,j_*)\}}w'_{ij}\cdot\delta_{R_{ij}},\sum_{(i,j)\in\mathcal{S}_{\mathrm{cal}}\cup\{(i_*,j_*)\}}w_{ij}\cdot\delta_{R_{ij}}\right)\\&\leq\frac{1}{2}\sum_{(i,j)\in\mathcal{S}_{\mathrm{cal}}\cup\{(i_*,j_*)\}}|w'_{ij}-w_{ij}|=\Delta,\end{aligned}$$

where d_{TV} denotes the total variation distance. This completes the proof.

A.1.3 Proofs for examples in Section 2.3.3

Recall the definition of Δ:

$$\Delta = \frac{1}{2} \sum_{(i,j)\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} \left| \frac{\widehat{h}_{ij}}{\sum_{(i',j')\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} \widehat{h}_{i'j'}} - \frac{h_{ij}}{\sum_{(i',j')\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} h_{i'j'}} \right|.$$

Note that $\Delta \leq 1$ by definition. We start with stating a useful lemma to bound the coverage gap Δ using the estimation error of $\widehat{h}_{ij}$.

Lemma A.1.1. *By the definition of the coverage gap Δ, we have*

$$\Delta \leq \frac{\sum_{(i,j)\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} |\widehat{h}_{ij} - h_{ij}|}{\sum_{(i',j')\in\mathcal{S}_{\text{cal}}} \widehat{h}_{i'j'}}. \tag{A.5}$$

Proof for Example 2.3.3

Under the logistic model, one has for any (i, j)

$$\mathbb{P}((i,j) \in \mathcal{S}_{\text{train}}) = q \cdot p_{ij} = \frac{q \cdot \exp(u_i + v_j)}{1 + \exp(u_i + v_j)}.$$

In this case, if $\widehat{\boldsymbol{u}}$ and $\widehat{\boldsymbol{v}}$ are the constrained maximum likelihood estimators in Example 2.3.3, Theorem 6 in Chen et al. (2023) implies that

$$\|\widehat{\mathbf{u}} - \mathbf{u}\|_\infty = O_{\mathbb{P}}\left(\sqrt{\frac{\log d_1}{d_2}}\right), \qquad \|\widehat{\mathbf{v}} - \mathbf{v}\|_\infty = O_{\mathbb{P}}\left(\sqrt{\frac{\log d_2}{d_1}}\right),$$

with the proviso that $\|\boldsymbol{u}\|_\infty + \|\boldsymbol{v}\|_\infty \leq \tau < \infty$, $d_2 \gg \sqrt{d_1}\log d_1$ and $d_1 \gg (\log d_2)^2$. Then for $h_{ij} = \exp(-u_i - v_j)$ and $\widehat{h}_{ij} = \exp(-\widehat{u}_i - \widehat{v}_j)$, we have

$$\max_{i,j} |\widehat{h}_{ij} - h_{ij}| = \max_{i,j} e^{-u_i - v_j} \left(e^{-(\widehat{u}_i - u_i) - (\widehat{v}_j - v_j)} - 1\right) = O_{\mathbb{P}}\left(\sqrt{\frac{\log d_1}{d_2}} + \sqrt{\frac{\log d_2}{d_1}}\right).$$

Further, as $\min_{i,j} h_{i,j} = \min_{i,j} \exp(-u_i - v_j) \geq e^{-\tau}$, then with probability approaching one, for every $(i,j) \in [d_1] \times [d_2]$, one has $\widehat{h}_{i,j} \geq h_{i,j} - |h_{ij} - \widehat{h}_{ij}| \geq e^{-\tau}/2 =: h_0$. By the upper bound (A.5), we have

$$\Delta \lesssim \sqrt{\frac{\log d_1}{d_2}} + \sqrt{\frac{\log d_2}{d_1}} \asymp \sqrt{\frac{\log \max\{d_1, d_2\}}{\min\{d_1, d_2\}}}.$$

Further, as $\Delta \leq 1$, we have $\mathbb{E}[\Delta] \lesssim \sqrt{\frac{\log \max\{d_1,d_2\}}{\min\{d_1,d_2\}}}$.

Proof of Example 2.3.3

Define the link function $\psi(t) = q(1 + e^{-\phi(t)})$, where ϕ is monotonic. Applying Theorem 1 in the paper (Davenport et al., 2014), we obtain that with probability at least $1 - C_1/(d_1 + d_2)$

$$\frac{1}{d_1 d_2} \|\widehat{\mathbf{A}} - \mathbf{A}\|_{\mathrm{F}}^2 \leq \sqrt{2} \widetilde{C}_\tau \sqrt{\frac{k(d_1 + d_2)}{d_1 d_2}},$$

with the proviso that $d_1 d_2 \geq (d_1 + d_2) \log(d_1 d_2)$. Here $\widetilde{C}_\tau = 2^{39/4} e^{9/4} (1 + \sqrt{6}) \tau L_\tau \beta_\tau$ with

$$L_\tau = \sup_{-\tau \leq t \leq \tau} \frac{|\psi'(t)|}{\psi(t)(1 - \psi(t))}, \quad \text{and} \quad \beta_\tau = \sup_{-\tau \leq t \leq \tau} \frac{\psi(t)(1 - \psi(t))}{|\psi'(t)|^2}.$$

Denote this high probability event to be $\mathcal{E}_0$. Since $\|\mathbf{A}\|_1 \leq \sqrt{d_1 d_2} \|\mathbf{A}\|_{\mathrm{F}}$, on this event $\mathcal{E}_0$, we further have

$$\frac{1}{d_1 d_2} \|\widehat{\mathbf{A}} - \mathbf{A}\|_1 \leq C_\tau \left(\frac{k(d_1 + d_2)}{d_1 d_2} \right)^{1/4},$$

where $C_\tau = \zeta \sqrt{L_\tau \beta_\tau}$ and ζ is a universal constant.

Recall that $h_{ij} = \exp(-\phi(A_{ij}))$ and $\|\mathbf{A}\|_\infty \leq \tau$. On the same event $\mathcal{E}_0$, we further have

$$\frac{1}{d_1 d_2} \|\widehat{\mathbf{H}} - \mathbf{H}\|_1 \leq C'_\tau \left(\frac{k(d_1 + d_2)}{d_1 d_2} \right)^{1/4},$$

where $C'_\tau = 2e^{\phi(\tau)}\sqrt{\widetilde{C}_\tau}$.

By the feasibility of the minimizer $\widehat{\mathbf{A}}$ and the fact that $\widehat{h}_{ij} = \exp(-\phi(\widehat{A}_{ij}))$, we have $\widehat{h}_{ij} \geq h_0 = e^{-\phi(\tau)}$ for all (i,j). This together with the upper bound (A.5) implies that

$$\Delta \leq \frac{1}{h_0 n_{\text{cal}}} \sum_{(i,j)\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} |h_{ij} - \widehat{h}_{ij}| \leq \frac{1}{h_0 n_{\text{cal}}} \|\widehat{\mathbf{H}} - \mathbf{H}\|_1.$$

Define a second high probability event $\mathcal{E}_1 = \{n_{\text{cal}} \geq (1-c)(1-q)\|\mathbf{P}\|_1\}$. Using the Chernoff bound, we have $\mathbb{P}(\mathcal{E}_1) \geq 1 - C_0(d_1 d_2)^{-5}$. Therefore, on the event $\mathcal{E}_0 \cap \mathcal{E}_1$, we have

$$\Delta \leq \frac{d_1 d_2}{c h_0 (1-q)\|\mathbf{P}\|_1} C'_\tau \left(\frac{k(d_1+d_2)}{d_1 d_2}\right)^{1/4}.$$

Under the assumptions that $p_{ij} = 1/(1+e^{-\phi(A_{ij})})$, and that $\|\mathbf{A}\|_\infty \leq \tau$, we know that $p_{ij} \geq C_2$ for some constant that only depends on τ. As a result, we have $\|\mathbf{P}\|_1 \geq C_2 d_1 d_2$, which further leads to the conclusion that

$$\Delta \leq \frac{1}{c C_2 h_0 (1-q)} C'_\tau \left(\frac{k(d_1+d_2)}{d_1 d_2}\right)^{1/4}.$$

on the event $\mathcal{E}_0 \cap \mathcal{E}_1$. In addition, on the small probability event $(\mathcal{E}_0 \cap \mathcal{E}_1)$, one trivially has $\Delta \leq 1$. Therefore simple combinations of the cases yields the desired bound on $\mathbb{E}[\Delta]$.

Proof of Lemma A.1.1

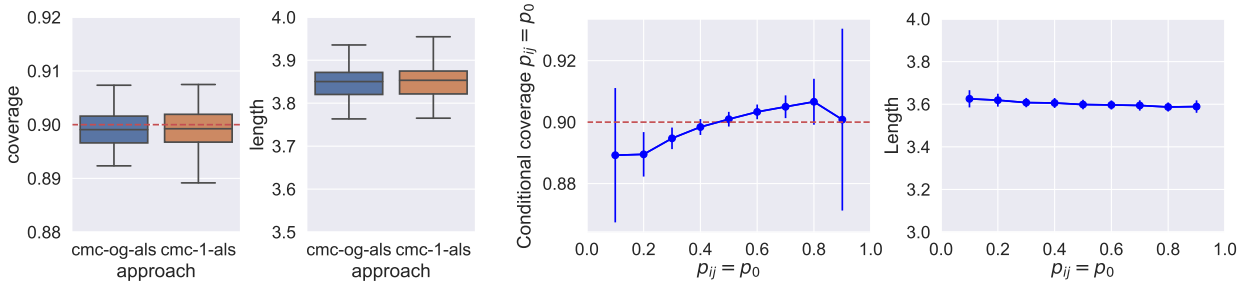
Reusing the definition of w' (A.4), one has

$$
\begin{aligned}
\Delta = & \frac{1}{2} \sum_{(i,j)\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} |w_{ij} - w'_{ij}| \\
= & \frac{1}{2} \sum_{(i,j)\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} \frac{\left|h_{ij} \sum_{(i',j')\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} \widehat{h}_{i'j'} - \widehat{h}_{ij} \sum_{(i',j')\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} h_{i'j'}\right|}{\left(\sum_{(i',j')\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} \widehat{h}_{i'j'}\right)\left(\sum_{(i',j')\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} h_{i'j'}\right)} \\
\leq & \frac{1}{2} \sum_{(i,j)\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} \left\{ \frac{h_{ij}\left| \sum_{(i',j')\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} \widehat{h}_{i'j'} - \sum_{(i',j')\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} h_{i'j'}\right|}{\left(\sum_{(i',j')\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} \widehat{h}_{i'j'}\right)\left(\sum_{(i',j')\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} h_{i'j'}\right)} \right. \\
& \qquad \left. + \frac{|\widehat{h}_{ij} - h_{ij}| \sum_{(i',j')\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} h_{i'j'}}{\left(\sum_{(i',j')\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} \widehat{h}_{i'j'}\right)\left(\sum_{(i',j')\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} h_{i'j'}\right)} \right\} \\
= & \frac{1}{2} \frac{\left| \sum_{(i',j')\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} \widehat{h}_{i'j'} - \sum_{(i',j')\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} h_{i'j'}\right|}{\left(\sum_{(i',j')\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} \widehat{h}_{i'j'}\right)} + \frac{1}{2} \frac{\sum_{(i,j)\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} |\widehat{h}_{ij} - h_{ij}|}{\left(\sum_{(i',j')\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} \widehat{h}_{i'j'}\right)} \\
\leq & \frac{\sum_{(i,j)\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} |\widehat{h}_{ij} - h_{ij}|}{\sum_{(i',j')\in\mathcal{S}_{\text{cal}}\cup\{(i_*,j_*)\}} \widehat{h}_{i'j'}}.
\end{aligned}
$$

This completes the proof.

A.2 Additional numerical experiments

In this section, we provide additional simulation results. In Section A.2.1, we compare the proposed one-shot weighted approach with the exact weighted conformal prediction. In Section A.2.2, the local coverage of cmc-als conditioning on p_{ij} is evaluated. We present the simulation results for the convex relaxation (cvx) and the conformalized convex method (cmc-cvx) in both settings with homogeneous and heterogeneous missingness in Section A.2.3 and A.2.5. We further evaluate the performance of cmc-als when the sampling probability is misspecified in Section A.2.6.



(a) Exact/original (og) cmc vs 1-shot cmc.

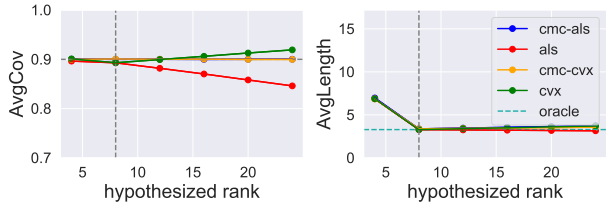
(b) Local performance of coverage given $p_{ij} = p_0$.

A.2.1 Comparison between one-shot and exact weighted approaches

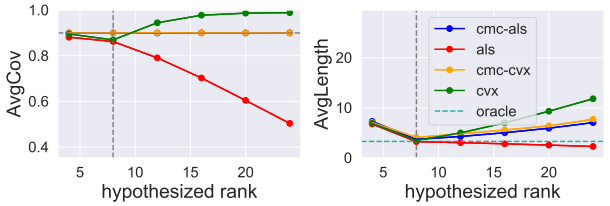
We conduct comparison in the setting with heterogeneous missingness and Gaussian noise as shown in Section 2.4.1 with $k^* = 12$. From the result in Figure A.1a, the performance of two aproaches are essentially the same while the one-shot cmc-als is more computationally efficient.

A.2.2 Local coverage of cmc

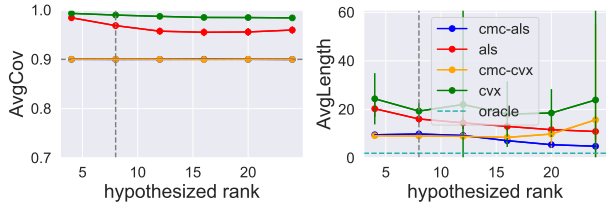
In Figure A.1b, we evaluate the local performance of cmc-als when conditioning on $\{p_{ij} = p_0\}$, i.e. a subpopulation determined by the value of sampling probability. We use the setting with heterogeneous missingness and Gaussian noise as shown in Section 2.4.1 with $k^* = 12$ and p_0 ranging from 0.1 to 0.9 with a step size of 0.1. The conditional coverage is approximated via kernel smoothing: with the indicators A_{ij} for coverage, we use the weight $K_{ij} = \phi_{p_0,h}(p_{ij})$ and calculate the conditional coverage by $(\sum A_{ij}K_{ij})/(\sum K_{ij})$. Here $\phi_{\mu,\sigma}$ is the density function of $\mathcal{N}(\mu, \sigma^2)$ and $h = 0.05$. When p_0 increases from 0.2 to 0.8, the conditional coverage increases and stays around 0.9. At $p_0 = 0.1,\ 0.9$, the coverage varies much, which is due to the small effective sample size for the edge values of p_{ij}. Moreover, for uniform sampling, since the rows and columns are generated i.i.d., there are no meaningful subgroups of the data to condition on.



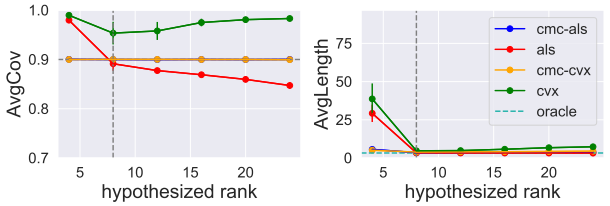
(a) Setting 1: Large sample size + Gaussian noise



(b) Setting 2: Small sample size+Gaussian noise



(c) Setting 3: Large sample size + heavy-tailed noise



(d) Setting 4: Violation of incoherence

Figure A.2: Comparison between conformalized and model-based matrix completion approaches.

A.2.3 Additional results for homogeneous misingness

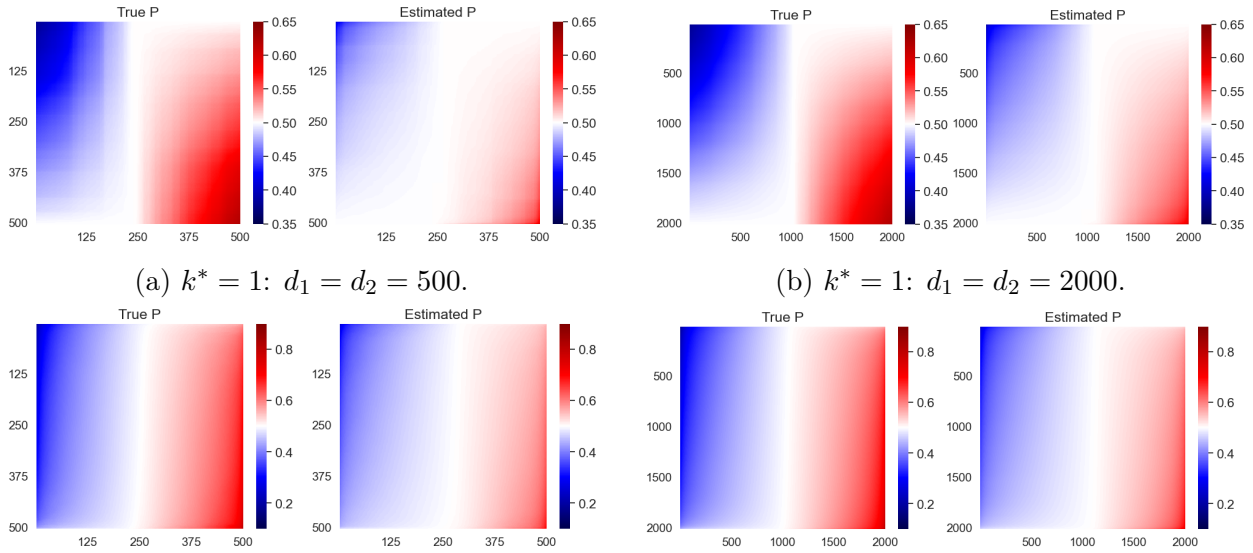
In this section, we present the results for synthetic simulation with the convex relaxation cvx and the conformalized convex matrix completion method cmc-cvx. Setting 1, 2, 3, 4 are the same as introduced in Section2.4.1. The true rank is $r^* = 8$ and the hypothesized rank varies from 4 to 24 with the stepsize 4.

The conformalized methods, regardless of the based algorithm adopted, have nearly exact coverage around $1-\alpha$. But we can observe different behaviors between als and cvx since the convex relaxation is free of the choice of r until the projection of $\widehat{\mathbf{M}}_{cvx}$ onto the rank-r subspace (Chen et al., 2020c). As a result, when $r > r^*$, cvx tends to overestimate the strength of the noise. In Figure A.2a, A.2b and A.2d, when $r > r^*$, cvx has coverage rate higher than the target level and the confidence interval is more conservative than conformalized methods. Since the accuracy of cvx is based on the large sample size, in Figure A.2b, when the effective sample size is insufficient with small p_{ij}, the residuals from cvx have a large deviation from the standard distribution and the intervals are much larger than the oracle

ones. Besides, in Figure A.2c, when the noise has heavy tails than Gaussian variables, cvx overestimates the noise strength similar to als and is conservative in coverage. When the incoherence condition is violated in Figure A.2d, if $r < r^*$, both cvx and als fit the missed factor by overestimating the noise strength and produce extremely large intervals.

A.2.4 Estimation error for one-bit matrix estimation

The estimation error in $\widehat{p}_{ij}$ can be visualized from the following heatmaps comparing $\mathbf{P}$ and $\widehat{\mathbf{P}}$. Here the entries are sorted by the order of p_{ij} for each row and each column.



(a) $k^* = 1$: $d_1 = d_2 = 500$.

(b) $k^* = 1$: $d_1 = d_2 = 2000$.

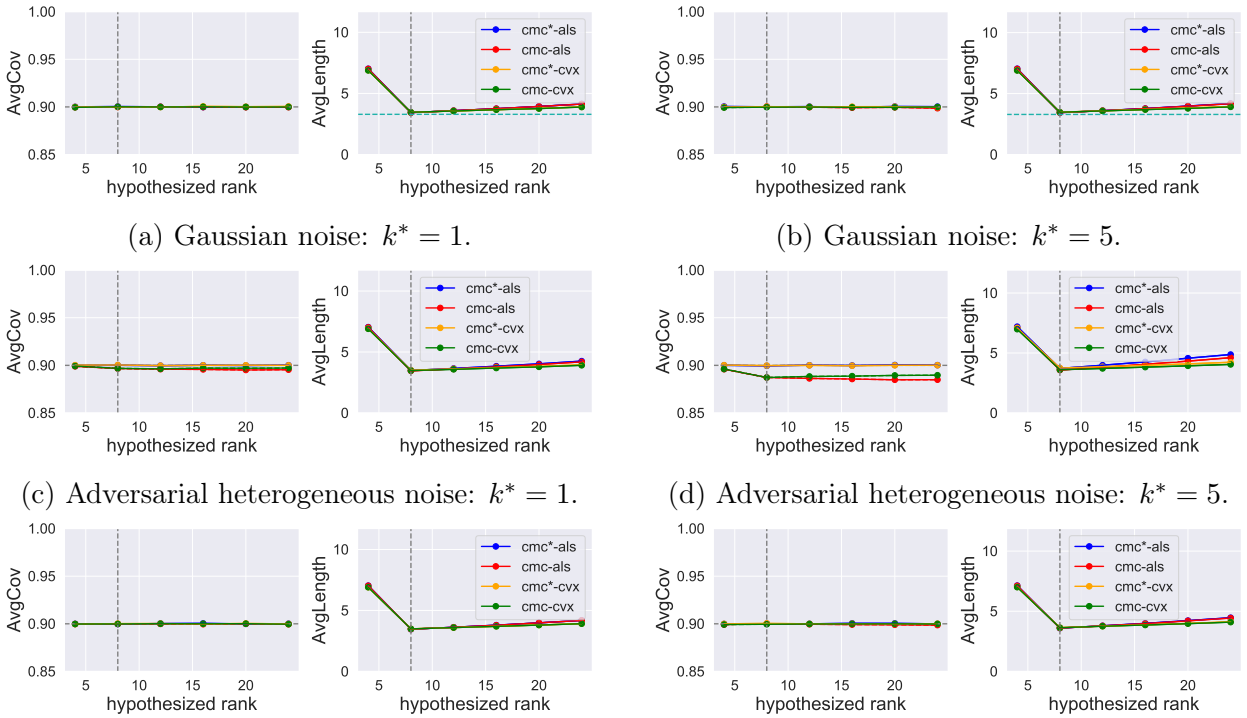
(c) $k^* = 5$: $d_1 = d_2 = 500$.

(d) $k^* = 5$: $d_1 = d_2 = 2000$.

Figure A.3: Heatmaps for $\mathbf{P}$ and $\widehat{\mathbf{P}}$.

A.2.5 Additional results for heterogeneous missingness

In Figure A.4, we present the results for synthetic simulation with the convex relaxation cvx as the base algorithm, where we denote cmc-cvx and cmc*-cvx as the conformalized matrix completion method with estimated weights and true weights, respectively. Three settings with heterogeneous missingness are the same as Figure 2.3.



(a) Gaussian noise: $k^* = 1$.

(b) Gaussian noise: $k^* = 5$.

(c) Adversarial heterogeneous noise: $k^* = 1$.

(d) Adversarial heterogeneous noise: $k^* = 5$.

(e) Random heterogeneous noise: $k^* = 1$.

(f) Random heterogeneous noise: $k^* = 5$.

Figure A.4: Comparison under heterogenous missingness.

A.2.6 Misspecified sampling probability

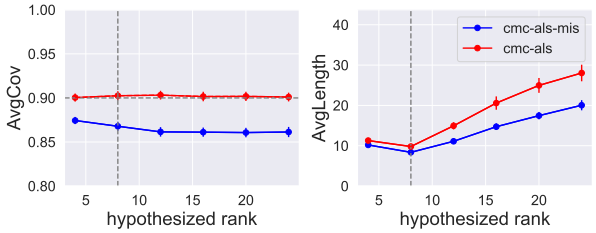
We consider the following four settings:

(a) The underlying missingness follows the rank-one model where $p_{ij} = a_i b_j$, both a_i and b_j are generated i.i.d. from $\mathrm{Unif}(0.2, 1)$. The noise is adversarial. But we estimate p_{ij} via the one-bit matrix completion based on the logistic model (working model with hypothesized rank $k = 5$).

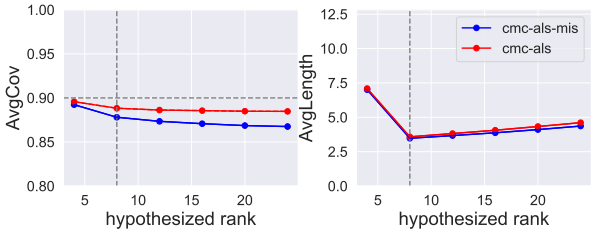
(b) The underlying missingness follows the logistic model as in Example 2.3.3 with $k^* = 5$ and we adopt the adversarial noise. But p_{ij} is estimated under the assumption of uniform sampling (working model), i.e. $\widehat{p}_{ij} = \widehat{p}$.

(c) Same as (a) except that we use the random heterogeneous noise.

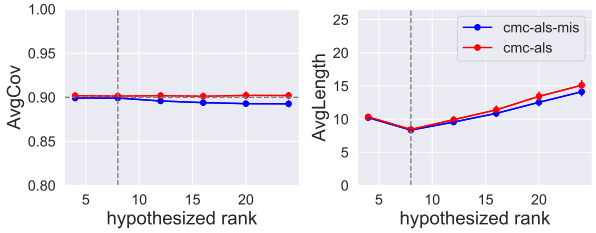
(d) Same as (b) except that we use the random heterogeneous noise.



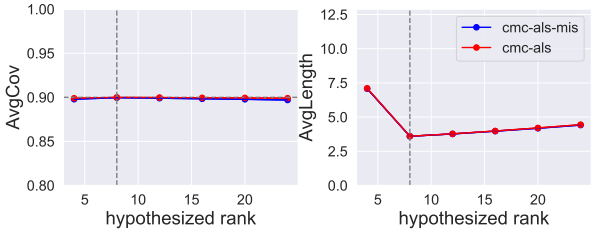
(a) Estimating rank-1 model with logistic model (adversarial noise).



(b) Estimating logistic model with uniform sampling (adversarial noise).



(c) Estimating rank-1 model with logistic model (random noise).

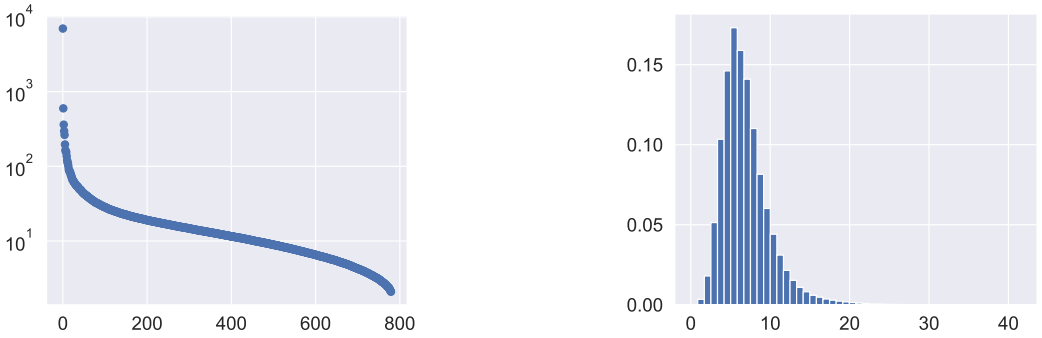


(d) Estimating logistic model with uniform sampling (random noise).

From the results, we can see that when the noise is generated following the random heterogeneous model, where the values of entries are independent of the sampling probabilities, the misspecification of the sampling model only slightly affects the coverage. Moreover, when the noise is generated in the adversarial way, where the values of entries depend on p_{ij}'s, we can see that the coverage with a misspecified sampling model is lower than the target level, but is above 0.85 in practice, which depends on the divergence between the true and the working sampling models.

A.2.7 Additional results for the sales dataset

Denote $\mathbf{M}$ the underlying matrix in the sales dataset. In Figure A.6a, we plot singular values of $\mathbf{M}$ and top-5 singular values contain a large proportion of the information. In Figure A.6b, we plot the histogram of entries M_{ij}'s of the underlying matrix, and the sales dataset has the range from 0 to over 20 thousand with a heavy tail.



(a) Singular values of the underlying matrix $\mathbf{M}$.

(b) Histogram of entries in $\mathbf{M}$.

Figure A.6: Descriptive plots for the underlying matrix $\mathbf{M}$.

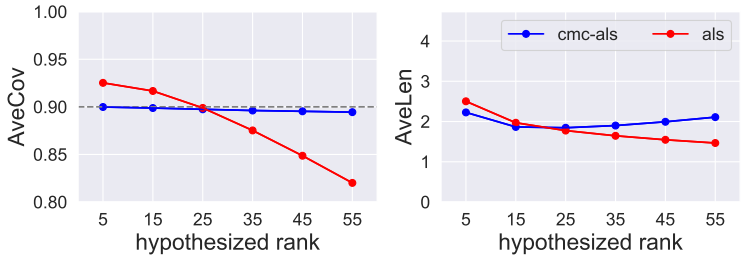


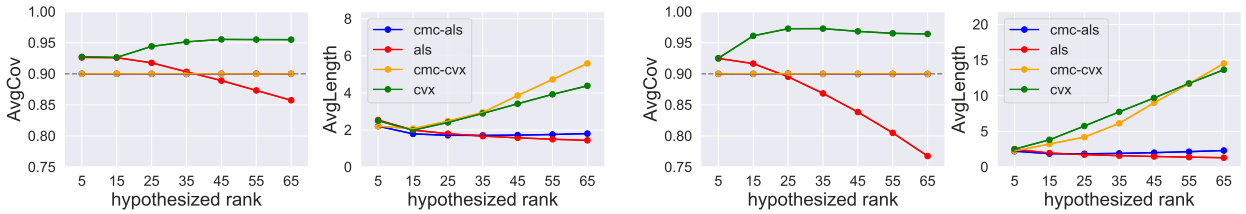
Figure A.7: Sales dataset with heterogeneous and misspecified missingness.

Comparison under misspecified missingness

We consider the sales dataset where the missingness occurs in the way that: for weekdays, each entry is observed with $p = 0.8$. On weekends, as the stores are likely to be operated by less experienced interns or are less frequent to report the sales data, each entry is observed with a lower probability, e.g. $0.8/3$. Moreover, as there could be a subgroup of stores that are less frequent in sales reporting, 200 stores are randomly sampled, for which $p = 0.8/3$. We use the logistic model with $k = 5$ to estimate p_{ij}. Comparison between als and cmc-als is shown in Figure A.7.

Results with convex optimization

In Figure A.8, cmc-cvx has nearly exact coverage at $1 - \alpha$, but cvx tends to have higher coverage than the target level. Besides, the convex approach has much larger intervals when



(a) Homogeneous missingness.

(b) Heterogeneous missingness with $k^* = 1$.

Figure A.8: Comparison between conformalized and model-based matrix completion with sales dataset.

r is large, which can be caused by the overfitting of the observed entries. As conformalized approach leaves out a proportion of observed entries as the training set, intervals produced by cmc-cvx are less accurate than cvx due to the poorly behaved base algorithm.

A.3 Additional details of algorithms and extensions

A.3.1 Extension to likelihood-based scores and categorical matrices

We will show in this section that cmc can also be applied to categorical matrix completion (Cao and Xie, 2015) or a more general setting in (A.6), the validity of which is also guaranteed by the presented theorem.

Setup To formulate the problem, consider an underlying parameter matrix $\mathbf{M}^* \in [d_1] \times [d_2]$ and the observations $\{M_{ij} : \ i \in [d_1], j \in [d_2]\}$ are drawn from the distribution

$$M_{ij} \mid M_{ij}^* \sim \mathcal{P}_{M_{ij}^*}, \tag{A.6}$$

where $\{\mathcal{P}_\theta\}_{\theta \in \Theta}$ can be a family of parametric distributions with probability density p_θ. The categorical matrix completion is a specific example where the support of M_{ij} is finite or countable. For example, a Poisson matrix is generated by $M_{ij} \sim \text{Pois}(M_{ij}^*)$, where M_{ij}^* is the Poisson mean. Similar to the previous setup, we treat $\mathbf{M}$ as deterministic, and entries in the subset $\mathcal{S} \subseteq [d_1] \times [d_2]$ are available. Here $\mathcal{S}$ is sampled in the same manner as before with the matrix $\mathbf{P} \in [d_1] \times [d_2]$.

Split conformal approach Consider the split approach with the partition $\mathcal{S} = \mathcal{S}_{\text{train}} \cup \mathcal{S}_{\text{cal}}$ as in Algorithm 1. With the training set $\mathbf{M}_{\text{train}}$, we obtain an estimated likelihood function $\widehat{\pi}(m; i, j)$ such that

$$\widehat{\pi}(m; i, j) = \widehat{p}_{M_{ij}^*}(m),$$

which is an estimate for the true likelihood of M_{ij} at m. The estimation can be feasible given certain low-complexity structures. For example, if a hypothesized distribution family $\{\mathcal{Q}_\theta\}_{\theta \in \Theta}$ with probability density q_θ is given and the underlying mean matrix $\mathbf{M}^*$ is assumed to be low-rank. Then, $\mathbf{M}$ can be viewed as a perturbation of $\mathbf{M}^*$ and we can estimate $\mathbf{M}^*$

via matrix completion algorithms with entries in $\mathbf{M}_{\text{train}}$. Denote $\widehat{\mathbf{M}}$ as the estimate for $\mathbf{M}^*$, then we have the estimated likelihood

$$\widehat{\pi}(m; i, j) = q_{\widehat{M}_{ij}}(m).$$

The odds ratios are also estimated from the training set, i.e. $\widehat{h}_{ij}$, from which we compute the weights

$$\widehat{w}_{ij} = \frac{\widehat{h}_{ij}}{\sum_{(i',j')\in\mathcal{S}_{\text{cal}}} \widehat{h}_{i'j'} + \max_{(i',j')\in\mathcal{S}^c} \widehat{h}_{i'j'}}, \ (i,j) \in \mathcal{S}_{\text{cal}}, \quad \widehat{w}_{\text{test}} = \frac{\widehat{h}_{i_*j_*}}{\sum_{(i',j')\in\mathcal{S}_{\text{cal}}} \widehat{h}_{i'j'} + \max_{(i',j')\in\mathcal{S}^c} \widehat{h}_{i'j'}}.$$

For each $(i,j) \in \mathcal{S}_{\text{cal}}$, calculate the likelihood-based nonconformity score

$$R_{ij} = -\widehat{\pi}(M_{ij}; i, j).$$

Then, for any test point $(i_*, j_*) \in \mathcal{S}^c$, we can construct the confidence interval

$$\widehat{C}(i_*, j_*) = \{m \in [K] : \ \widehat{\pi}(m; i_*, j_*) \leq \widehat{q}\},$$

where $\widehat{q}$ is the weighted quantile

$$\widehat{q} = \text{Quantile}_{1-\alpha}\left(\sum_{(i,j)\in\mathcal{S}_{\text{cal}}} \widehat{w}_{ij} \cdot \delta_{R_{ij}} + \widehat{w}_{\text{test}} \cdot \delta_{+\infty}\right).$$

More examples of conformal methods for classification are shown in Romano et al. (2020), Angelopoulos et al. (2020), etc.

A.3.2 Full conformalized matrix completion

In Algorithm 2, the procedure of the full conformalized matrix completion (full-cmc) is presented. This full conformal version of cmc offers the same coverage guarantee as given in Theorem 2.3.2 for the split version of cmc (except with the entire observed set $\mathcal{S}$ in place of $\mathcal{S}_{\text{cal}}$, when defining $\Delta(\widehat{\boldsymbol{w}})$; the formal proof of this bound for full conformal is very similar to the proof of Theorem 2.3.2, using an analogous weighted exchangeability argument as in Lemma 2.3.1, and so we omit it here.

To define this algorithm, we need some notation: given the observed data $\mathbf{M}_{\mathcal{S}}$, plus a test point location (i_*, j_*) and a hypothesized value m for the test point $M_{i_*j_*}$, define a matrix $\mathbf{M}^{(m)}$ with entries

$$M^{(m)}_{ij} = \begin{cases} M_{ij}, & (i,j) \in \mathcal{S}, \\ m, & (i,j) = (i_*, j_*), \\ \emptyset, & \text{otherwise}, \end{cases} \tag{A.7}$$

where, abusing notation, "$M_{ij} = \emptyset$" denotes that no information is observed in this entry.

We note that, when $\mathcal{M} = \mathbb{R}$ (or an infinite subset of $\mathbb{R}$), the computation of the prediction set is impossible in most of the cases. In that case, our algorithm can be modified via a trimmed or discretized approximation; these extensions are presented for the regression setting in the work of Chen et al. (2016, 2018), and can be extended to the matrix completion setting in a straightforward way.

A.3.3 Exact split conformalized matrix completion

In Algorithm 3, we present the exact split approach, which is less conservative than our one-shot approach given in Algorithm 1, but may be less computationally efficient. In this version of the algorithm, the quantile $\widehat{q} = \widehat{q}(i_*, j_*)$ needs to be computed for each missing entry since the weight vector $\widehat{\boldsymbol{w}}$ depends on the value of $\widehat{p}_{i_*,j_*}$.

Algorithm 2 full-cmc: full conformalized matrix completion

Input: target level $1-\alpha$; partially observed matrix $\mathbf{M}_{\mathcal{S}}$.

Using the training data $\mathbf{M}_{\mathcal{S}}$, compute an estimate $\widehat{\mathbf{P}}$ of the observation probabilities (with $\widehat{p}_{ij}$ estimating p_{ij}, the probability of entry (i,j) being observed).

for (i_*, j_*) in $\mathcal{S}^c$ **do**

  for $m \in \mathcal{M}$ **do**

    Augment $\mathbf{M}_{\mathcal{S}}$ with one additional hypothesized entry, $\{M_{i_*,j_*} = m\}$, to obtain $\mathbf{M}^{(m)}$ defined as in (A.7).

    Using the imputed matrix $\mathbf{M}_{\mathcal{S}}^{(m)}$, compute:

- An initial estimate $\widehat{\mathbf{M}}^{(m)}$ using any matrix completion algorithm (with $\widehat{M}_{ij}^{(m)}$ estimating the target M_{ij});
- Optionally, a local uncertainty estimate $\widehat{\mathbf{s}}^{(m)}$ (with $\widehat{s}_{ij}^{(m)}$ estimating our relative uncertainty in the estimate $\widehat{M}_{ij}^{(m)}$), or otherwise set $\widehat{s}_{ij}^{(m)} \equiv 1$;
- An estimate $\widehat{\mathbf{P}}$ of the observation probabilities (with $\widehat{p}_{ij}$ estimating p_{ij}, the probability of entry (i,j) being observed).

    Compute normalized residuals for $(i,j) \in \mathcal{S} \cup \{(i_*, j_*)\}$,

$$R_{ij}^{(m)} = \frac{|M_{ij} - \widehat{M}_{ij}^{(m)}|}{\widehat{s}_{ij}^{(m)}}.$$

    Compute weights

$$\widehat{w}_{ij} = \frac{\widehat{h}_{ij}}{\sum_{(i',j') \in \mathcal{S} \cup \{(i_*,j_*)\}} \widehat{h}_{i'j'}}, \qquad \widehat{h}_{ij} = \frac{1-\widehat{p}_{ij}}{\widehat{p}_{ij}}, \qquad (i,j) \in \mathcal{S} \cup \{(i_*, j_*)\}.$$

    Compute the weighted quantile

$$\widehat{q}^{(m)}(i_*, j_*) = \text{Quantile}_{1-\alpha}\left(\sum_{(i,j) \in \mathcal{S}} \widehat{w}_{ij} \delta_{R_{ij}^{(m)}} + \widehat{w}_{i_* j_*} \delta_{+\infty} \right)$$

  end for

end for

Output: $\left\{ \widehat{C}\,(i_*, j_*) = \left\{ m \in \mathcal{M} : \; R_{i_* j_*}^{(m)} \leq \widehat{q}^{(m)}(i_*, j_*) \right\} : \; (i_*, j_*) \in \mathcal{S}^c \right\}$

Algorithm 3 split-cmc: split conformalized matrix completion

Input: target coverage level $1-\alpha$; data splitting proportion $q \in (0,1)$; observed entries $\mathbf{M}_{\mathcal{S}}$.

Split the data: draw $W_{ij} \overset{\text{i.i.d.}}{\sim} \text{Bern}(q)$, and define training and calibration sets,

$$\mathcal{S}_{\text{train}} = \{(i,j) \in \mathcal{S} : W_{ij} = 1\}, \quad \mathcal{S}_{\text{cal}} = \{(i,j) \in \mathcal{S} : W_{ij} = 0\}.$$

Using the training data $\mathbf{M}_{\mathcal{S}_{\text{train}}}$ indexed by $\mathcal{S}_{\text{train}} \subseteq [d_1] \times [d_2]$, compute:

- An initial estimate $\widehat{\mathbf{M}}$ using any matrix completion algorithm (with $\widehat{M}_{ij}$ estimating the target M_{ij});
- Optionally, a local uncertainty estimate $\widehat{\mathbf{s}}$ (with $\widehat{s}_{ij}$ estimating our relative uncertainty in the estimate $\widehat{M}_{ij}$), or otherwise set $\widehat{s}_{ij} \equiv 1$;
- An estimate $\widehat{\mathbf{P}}$ of the observation probabilities (with $\widehat{p}_{ij}$ estimating p_{ij}, the probability of entry (i,j) being observed).

Compute normalized residuals on the calibration set,

$$R_{ij} = \frac{\left|M_{ij} - \widehat{M}_{ij}\right|}{\widehat{s}_{ij}}, \ (i,j) \in \mathcal{S}_{\text{cal}}.$$

Compute estimated odds ratios for the calibration set and test set,

$$\widehat{h}_{ij} = \frac{\widehat{p}_{ij}}{1-\widehat{p}_{ij}}, \ (i,j) \in \mathcal{S}_{\text{cal}} \cup \mathcal{S}^c,$$

for $(i_*, j_*) \in \mathcal{S}^c$ **do**

Compute weights for the calibration set and test point,

$$\widehat{w}_{ij} = \frac{\widehat{h}_{ij}}{\sum_{(i',j') \in \mathcal{S}_{\text{cal}}} \widehat{h}_{i'j'} + \widehat{h}_{i_*j_*}}, \ (i,j) \in \mathcal{S}_{\text{cal}}, \quad \widehat{w}_{\text{test}} = \frac{\widehat{h}_{i_*j_*}}{\sum_{(i',j') \in \mathcal{S}_{\text{cal}}} \widehat{h}_{i'j'} + \widehat{h}_{i_*j_*}}.$$

Compute threshold

$$\widehat{q}(i_*, j_*) = \text{Quantile}_{1-\alpha}\left(\sum_{(i,j) \in \mathcal{S}_{\text{cal}}} \widehat{w}_{ij} \cdot \delta_{R_{ij}} + \widehat{w}_{\text{test}} \cdot \delta_{+\infty}\right),$$

where δ_t denotes the point mass at t.

end for

Output: confidence intervals

$$\widehat{C}(i_*, j_*) = \widehat{M}_{i_*j_*} \pm \widehat{q}(i_*, j_*) \cdot \widehat{s}_{i_*j_*}$$

for each unobserved entry $(i_*, j_*) \in \mathcal{S}^c$. 139

APPENDIX B

APPENDIX FOR CHAPTER 3

B.1 Proofs of results in Section 3.2

B.1.1 Proof of Proposition 3.2.2

It is straightforward to check that $\Delta(R;\mathcal{B})$ is always an upper bound of the new formulation stated in Proposition 3.2.2, simply by taking $w = \phi \circ R$. Therefore, it remains to show the converse: under Condition 3.2.1, $\Delta(R;\mathcal{B})$ is also a lower bound of the new formulation stated in Proposition 3.2.2.

To this end, it suffices to prove that for any $w_{\#}P \in \mathcal{B}$, there exists a nondecreasing function ϕ such that $(\phi \circ R)_{\#}P \in \mathcal{B}$, and

$$\mathbb{E}_P\left[w(X)R(X)\right] \leq \mathbb{E}_P\left[\phi(R(X))R(X)\right].$$

We construct such a function ϕ in two steps.

Step 1: Conditioning. For any w such that $w_{\#}P \in \mathcal{B}$, we define g as a measurable function satisfying

$$g(R(X)) = \mathbb{E}\left[w(X) \mid R(X)\right], \quad P\text{-almost surely.}$$

(Note that g is not necessarily a monotone function.) As a result, by the tower law, we have

$$\mathbb{E}_P[w(X)R(X)] = \mathbb{E}_P[g(R(X))R(X)]. \tag{B.1}$$

Since $w_{\#}P \in \mathcal{B}$, by Jensen's inequality, for any convex function ψ, we have

$$\mathbb{E}_P\left[\psi\left(g(R(X))\right)\right] = \mathbb{E}_P\left[\psi\left(\mathbb{E}\left[w(X) \mid R(X)\right]\right)\right] \leq \mathbb{E}_P\left[\psi(w(X))\right],$$

which implies $(g \circ R)_{\#}P \in \mathcal{B}$ by Condition 3.2.1.

Step 2: Rearrangement. Denote F_1 and F_2 as the cumulative distribution functions of $g(R(X))$ and $R(X)$, respectively. Let $U \sim \text{Unif}([0,1])$. Then, we have $F_1^{-1}(U) \stackrel{d}{=} g(R(X))$ and $F_2^{-1}(U) \stackrel{d}{=} R(X)$, where F_k^{-1} is the generalized inverse of F_k for $k = 1, 2$, and where $\stackrel{d}{=}$ denotes equality in distribution. Moreover, F_1^{-1} is nondecreasing and

$$g(F_2^{-1}(U)) \stackrel{d}{=} g(R(X)) \stackrel{d}{=} F_1^{-1}(U),$$

which implies that F_1^{-1} is the monotone rearrangement of $g \circ F_2^{-1}$. By (Hardy et al., 1952, eqn. (378)), we have

$$\mathbb{E}_P\left[g(R(X))R(X)\right] = \mathbb{E}\left[g(F_2^{-1}(U))F_2^{-1}(U)\right] \leq \mathbb{E}\left[F_1^{-1}(U)F_2^{-1}(U)\right]. \tag{B.2}$$

Next, let ϕ be a measurable function satisfying

$$\phi(F_2^{-1}(U)) = \mathbb{E}\left[F_1^{-1}(U) \mid F_2^{-1}(U)\right],$$

almost surely with respect to the distribution $U \sim \text{Unif}([0,1])$. Since F_k^{-1} is the generalized inverse of a CDF F_k, for each $k = 1, 2$, it is therefore monotone nondecreasing. Therefore, we can choose ϕ to be a monotone nondecreasing function. Moreover, to verify that $(\phi \circ R)_{\#}P \in \mathcal{B}$, we will check that $\phi(R(X)) \stackrel{cvx}{\preceq} g(R(X))$ (and use Condition 3.2.1, along with the fact

that $(g \circ R)_{\#}P \in \mathcal{B}$ as established above): for any convex function ψ, we have

$$\begin{aligned}\mathbb{E}_P[\psi(\phi(R(X)))] &\stackrel{d}{=} \mathbb{E}[\psi(\phi(F_2^{-1}(U)))] \\ &= \mathbb{E}[\psi(\mathbb{E}\left[F_1^{-1}(U) \mid F_2^{-1}(U)\right])] \leq \mathbb{E}[\psi(F_1^{-1}(U))] = \mathbb{E}_P[\psi(g(R(X)))],\end{aligned}$$

where the inequality holds by Jensen's inequality.

We then have

$$\begin{aligned}\mathbb{E}\left[F_1^{-1}(U)F_2^{-1}(U)\right] &= \mathbb{E}\left[\mathbb{E}\left[F_1^{-1}(U) \mid F_2^{-1}(U)\right] F_2^{-1}(U)\right] \\ &= \mathbb{E}\left[\phi(F_2^{-1}(U))F_2^{-1}(U)\right] = \mathbb{E}_P\left[\phi(R(X))R(X)\right].\end{aligned}$$

This equality, combined with (B.1) and (B.2), yields the desired outcome: $\mathbb{E}_P\left[w(X)R(X)\right] \leq \mathbb{E}_P\left[\phi(R(X))R(X)\right]$. We hence complete the proof.

B.2 Proofs of results in Section 3.3

B.2.1 Preliminaries

Before we present the proof, we begin with some preliminaries: we introduce some notation, definitions, and facts that will aid in the proof below.

Adding an L_2 constraint

First, we will define a version of our optimization problem that defines $\Delta(R;\mathcal{B})$, by adding an L_2 constraint:

$$\begin{aligned}\Delta_2(R;\mathcal{B}) \quad = \quad & \sup_{w \geq 0,\, w \in L_2(P)} \quad \mathbb{E}_P\left[w(X)R(X)\right] - \mathbb{E}_P[R(X)] \\ & \text{subject to} \quad w_{\#}P \in \mathcal{B}.\end{aligned}$$

We can observe that, by construction,

$$\Delta_2(R;\mathcal{B}) = \Delta(R;\mathcal{B}\cap\mathcal{B}_{L_2}),$$

where $\mathcal{B}_{L_2}$ is the set of all distributions with finite second moment. The following result verifies that adding the L_2 constraint does not change the outcome of the optimization problem:

Proposition B.2.1. *Under the notation and definitions above, it holds that* $\Delta(R;\mathcal{B}) = \Delta_2(R;\mathcal{B})$.

We defer the proof of this proposition to Section B.2.5.

The isotonic projection

We next review some facts regarding the isotonic projection operator π. To ease notation, we denote $\langle a,b\rangle_P = \int_{\mathcal{X}} a(x)b(x)\mathsf{d}P(x)$ for any functions $a,b\in L_2(P)$.

The first property relates to the isotonic projection as a projection to a convex cone (Bauschke and Combettes (2019), Theorem 3.14; Edwards (2012), Proposition 1.12.4):

$$\text{For any } w\in L_2(P) \text{ and any } v\in \mathcal{C}^{\text{iso}}_{\preceq}\cap L_2(P),\ \langle v, w-\pi(w)\rangle_P \le 0. \tag{B.3}$$

Moreover, it holds that (Brunk (1963), Theorem 1; Brunk (1965), Corollary 3.1):

$$\text{For any } w\in L_2(P) \text{ and any } h:\mathbb{R}\to\mathbb{R},\ \langle h\circ\pi(w), w-\pi(w)\rangle_P = 0. \tag{B.4}$$

In particular, by chosing $h(t)\equiv 1$, we can see that isotonic projection preserves the mean,

$$\text{For any } w\in L_2(P),\ \ \mathbb{E}_P[w(X)] = \mathbb{E}_P[[\pi(w)](X)]. \tag{B.5}$$

Finally, we relate the isotonic projection to the convex ordering:

$$\text{For any } w \in L_2(P),\ \pi(w) \overset{cvx}{\preceq} w. \tag{B.6}$$

To see (B.6), for any convex function ψ, by the nonnegativity of Bregman divergence (Bregman, 1967), it holds that

$$\langle \psi(w) - \psi(\pi(w)), 1 \rangle_P \geq \langle \psi' \circ \pi(w), w - \pi(w) \rangle_P.$$

According to the property (B.4), we further obtain $\langle \psi' \circ \pi(w), w - \pi(w) \rangle_P$, which implies that $\pi(w) \overset{cvx}{\preceq} w$ by definition.

B.2.2 Proof of Theorem 3.3.1

We split the proof into three steps:

1. prove that $\Delta^{\text{iso}}(R; \mathcal{B}) \leq \Delta(\pi(R); \mathcal{B})$;
2. prove that $\Delta^{\text{iso}}(R; \mathcal{B}) = \Delta(\pi(R); \mathcal{B})$ provided that Condition 3.2.1 holds;
3. prove the claim on attainability of minimizers provided that Condition 3.2.1 holds.

Step 1: Prove $\Delta^{\text{iso}}(R; \mathcal{B}) \leq \Delta(\pi(R); \mathcal{B})$. By the definition of $\Delta^{\text{iso}}(R; \mathcal{B})$ as the supremum in the optimization problem (3.5), for any $\varepsilon > 0$, there exists a feasible w_ε such that

$$\mathbb{E}_P[w_\varepsilon(X) \cdot R(X)] - \mathbb{E}_P[R(X)] \geq \Delta^{\text{iso}}(R; \mathcal{B}) - \varepsilon.$$

Next, define a sequence of truncated functions, $w_{\varepsilon,n}(x) = \min\{w_\varepsilon(x), n\}$. Since $w_\varepsilon \in \mathcal{C}^{\text{iso}}_{\preceq}$, it holds that $w_{\varepsilon,n} \in \mathcal{C}^{\text{iso}}_{\preceq}$ as well, and moreover since the truncated function is bounded we

also have $w_{\varepsilon,n} \in L_2(P)$. By fact (B.3), it therefore holds that

$$\mathbb{E}_P[w_{\varepsilon,n}(X) \cdot (R - [\pi(R)])(X)] = \langle w_{\varepsilon,n}, R - \pi(R) \rangle_P \leq 0,$$

for each $n \geq 1$. Then, by the dominated convergence theorem, taking a limit as $n \to \infty$ we obtain

$$\mathbb{E}_P[w_{\varepsilon}(X) \cdot (R - [\pi(R)])(X)] \leq 0.$$

Moreover, $\mathbb{E}_P[[\pi(R)](X)] = \mathbb{E}_P[R(X)]$ by (B.5). Combining everything, then,

$$\begin{aligned}\Delta^{\text{iso}}(R; \mathcal{B}) - \varepsilon &\leq \mathbb{E}_P[w_{\varepsilon}(X) \cdot R(X)] - \mathbb{E}_P[R(X)] \\ &\leq \mathbb{E}_P[w_{\varepsilon}(X) \cdot [\pi(R)](X)] - \mathbb{E}_P[[\pi(R)](X)] \leq \Delta(\pi(R); \mathcal{B}),\end{aligned}$$

where the last step holds since, because w_ε is feasible for the optimization problem (3.5) that defines $\Delta^{\text{iso}}(R; \mathcal{B})$, it is also feasible for $\Delta(\pi(R); \mathcal{B})$ (i.e., $w \geq 0$ and $w_{\#}P \in \mathcal{B}$). Since $\varepsilon > 0$ is arbitrary, we obtain the desired result $\Delta^{\text{iso}}(R; \mathcal{B}) \leq \Delta(\pi(R); \mathcal{B})$.

Step 2: Prove $\Delta^{\text{iso}}(R; \mathcal{B}) = \Delta(\pi(R); \mathcal{B})$ under Condition 3.2.1. By Proposition B.2.1, we have $\Delta(\pi(R); \mathcal{B}) = \Delta_2(\pi(R); \mathcal{B}) = \Delta(\pi(R); \mathcal{B} \cap \mathcal{B}_{L_2})$. Next, note that if Condition 3.2.1 holds for $\mathcal{B}$, then this condition holds for $\mathcal{B} \cap \mathcal{B}_{L_2}$ as well (because for any $Q' \overset{cvx}{\preceq} Q$, we have $\mathbb{E}_{Q'}[X^2] \leq \mathbb{E}_Q[X^2]$ by definition of the convex ordering—and so if $Q \in \mathcal{B}_{L_2}$ then $Q' \in \mathcal{B}_{L_2}$ as well.) Therefore, we can apply Proposition 3.2.2 to the term $\Delta(\pi(R); \mathcal{B} \cap \mathcal{B}_{L_2})$, which

yields the following equivalent formulation:

$$\begin{aligned}\Delta(\pi(R);\mathcal{B}\cap\mathcal{B}_{L_2}) \quad = \quad & \sup_{\phi:\mathbb{R}\to\mathbb{R}_+} \quad \mathbb{E}_P\left[(\phi\circ\pi(R))(X)\cdot[\pi(R)](X)\right]-\mathbb{E}_P[[\pi(R)](X)] \\ & \text{subject to} \quad (\phi\circ\pi(R))_{\#}P\in\mathcal{B}, \quad \phi\circ\pi(R)\in L_2(P), \quad \phi \text{ is nondecreasing.}\end{aligned}$$

Then, for any $\varepsilon>0$, there exists some ϕ_ε satisfying the above constraints so that

$$\mathbb{E}_P\left[(\phi_\varepsilon\circ\pi(R))(X)\cdot[\pi(R)](X)\right]-\mathbb{E}_P[[\pi(R)](X)]\geq\Delta(\pi(R);\mathcal{B}\cap\mathcal{B}_{L_2})-\varepsilon=\Delta(\pi(R);\mathcal{B})-\varepsilon.$$

Now define $\widetilde{w}_\varepsilon=\phi_\varepsilon\circ\pi(R)$, i.e., we have

$$\mathbb{E}_P\left[\widetilde{w}_\varepsilon(X)\cdot[\pi(R)](X)\right]-\mathbb{E}_P[[\pi(R)](X)]\geq\Delta(\pi(R);\mathcal{B})-\varepsilon,$$

where $(\widetilde{w}_\varepsilon)_{\#}P\in\mathcal{B}$ and $\widetilde{w}_\varepsilon\in L_2(P)$, and also $\widetilde{w}_\varepsilon\in\mathcal{C}^{\text{iso}}_{\preceq}$, by construction and by feasibility of ϕ_ε. Moreover, by the facts (B.5) and (B.4),

$$\mathbb{E}_P[[\pi(R)](X)]=\mathbb{E}_P[R(X)], \quad \langle\widetilde{w}_\varepsilon,R-\pi(R)\rangle_P=\langle\phi_\varepsilon\circ\pi(R),R-\pi(R)\rangle_P=0,$$

and therefore,

$$\mathbb{E}_P\left[\widetilde{w}_\varepsilon(X)\cdot R(X)\right]-\mathbb{E}_P[R(X)]\geq\Delta(\pi(R);\mathcal{B})-\varepsilon.$$

But we have verified above that $\widetilde{w}_\varepsilon$ is feasible for the optimization problem (3.5) defining $\Delta^{\text{iso}}(R;\mathcal{B})$, i.e.,

$$\mathbb{E}_P\left[\widetilde{w}_\varepsilon(X)\cdot R(X)\right]-\mathbb{E}_P[R(X)]\leq\Delta^{\text{iso}}(R;\mathcal{B}).$$

Since $\varepsilon>0$ is arbitrary, this verifies that $\Delta(\pi(R);\mathcal{B})\leq\Delta^{\text{iso}}(R;\mathcal{B})$, and thus completes this step.

Step 3: attainability of minimizers under Condition 3.2.1. Suppose $\Delta(\pi(R); \mathcal{B})$ is attained at $\widetilde{w}$, i.e.,

$$\mathbb{E}_P[\widetilde{w}(X) \cdot [\pi(R)](X)] - \mathbb{E}_P[[\pi(R)](X)] = \Delta(\pi(R); \mathcal{B}).$$

By Proposition 3.2.2, we can construct some nondecreasing function $\widetilde{\phi}$, with $(\widetilde{\phi} \circ \pi(R))_{\#} P \in \mathcal{B}$, such that

$$\Delta(\pi(R); \mathcal{B}) = \mathbb{E}_P[\widetilde{\phi}([\pi(R)](X)) \cdot [\pi(R)](X)] - \mathbb{E}_P[[\pi(R)](X)].$$

Recalling that $\mathbb{E}_P[[\pi(R)](X)] = \mathbb{E}_P[R(X)]$ by (B.5), and $\Delta(\pi(R); \mathcal{B}) = \Delta^{\text{iso}}(R; \mathcal{B})$ by Steps 1 and 2, we now have

$$\Delta^{\text{iso}}(R; \mathcal{B}) = \mathbb{E}_P[\widetilde{\phi}([\pi(R)](X)) \cdot [\pi(R)](X)] - \mathbb{E}_P[R(X)].$$

Next, by fact (B.4),

$$\mathbb{E}_P[\widetilde{\phi}([\pi(R)](X)) \cdot (R(X) - [\pi(R)](X))] = \langle \widetilde{\phi} \circ \pi(R), R - \pi(R) \rangle_P = 0,$$

and so

$$\Delta^{\text{iso}}(R; \mathcal{B}) = \mathbb{E}_P[\widetilde{\phi}([\pi(R)](X)) \cdot R(X)] - \mathbb{E}_P[R(X)].$$

Therefore, $\Delta^{\text{iso}}(R; \mathcal{B})$ is attained at $\widetilde{\phi} \circ \pi(R)$ (which, by construction, satisfies $\widetilde{\phi} \circ \pi(R) \in \mathcal{C}^{\text{iso}}_{\preceq}$, as well as $(\phi \circ \pi(R))_{\#} P \in \mathcal{B}$ as above, and is therefore feasible).

Conversely, suppose that $\Delta^{\text{iso}}(R; \mathcal{B})$ is attained at $\widetilde{w}$, i.e.,

$$\mathbb{E}_P[\widetilde{w}(X) \cdot R(X)] - \mathbb{E}_P[R(X)] = \Delta^{\text{iso}}(R; \mathcal{B}).$$

Again applying (B.5), and the fact that $\Delta(\pi(R);\mathcal{B}) = \Delta^{\text{iso}}(R;\mathcal{B})$ by Steps 1 and 2,

$$\Delta(\pi(R);\mathcal{B}) = \mathbb{E}_P[\widetilde{w}(X)\cdot R(X)] - \mathbb{E}_P[[\pi(R)](X)] \leq \mathbb{E}_P[\widetilde{w}(X)\cdot [\pi(R)](X)] - \mathbb{E}_P[[\pi(R)](X)],$$

where for the last step, since $\widetilde{w} \in \mathcal{C}^{\text{iso}}_{\preceq}$ (because it is feasible for $\Delta^{\text{iso}}(R;\mathcal{B})$), we have

$$\mathbb{E}_P[\widetilde{w}(X)\cdot(R(X) - [\pi(R)](X))] = \langle \widetilde{w}, R - \pi(R)\rangle_P \leq 0,$$

by (B.3). But $\widetilde{w}$ is feasible for $\Delta(\pi(R);\mathcal{B})$ (since it is feasible for $\Delta^{\text{iso}}(R;\mathcal{B})$), and therefore, we also have

$$\Delta(\pi(R);\mathcal{B}) = \Delta^{\text{iso}}(R;\mathcal{B}) \leq \mathbb{E}_P[\widetilde{w}(X)\cdot [\pi(R)](X)] - \mathbb{E}_P[[\pi(R)](X)].$$

In other words, $\Delta(\pi(R);\mathcal{B})$ is attained at $\widetilde{w}$, which completes the proof.

B.2.3 Proof of Proposition 3.3.2

We formally define $\Delta^{\text{iso}}(R;\mathcal{B},w_0)$ as follows:

$$\begin{aligned}\Delta^{\text{iso}}(R;\mathcal{B},w_0) \quad = \quad & \sup_{w\geq 0} \quad \mathbb{E}_P\left[w(X)R(X)\right] - \mathbb{E}_P[R(X)] \\ & \text{subject to} \quad w_{\#}P \in \mathcal{B}, \quad w \in \mathcal{C}^{\text{iso}}_{w_0}.\end{aligned}$$

For comparison, we also consider the following optimization problem

$$\begin{aligned}\widetilde{\Delta}^{\text{iso}}(R;\mathcal{B},w_0) \quad = \quad & \sup_{h:\ h\circ w_0 \geq 0} \quad \mathbb{E}_P\left[(h\circ w_0)(X)R(X)\right] - \mathbb{E}_P[R(X)] \\ & \text{subject to} \quad (h\circ w_0)_{\#}P \in \mathcal{B}, \quad h \in \mathcal{C}^{\text{iso}}_1, \end{aligned} \tag{B.7}$$

where $\mathcal{C}_1^{\text{iso}}$ denotes the cone of isotonic functions defined on $\mathbb{R}$ equipped with the natural ordering. In fact, since $\mathcal{C}_{w_0}^{\text{iso}} = \{h \circ w_0 : h \in \mathcal{C}_1^{\text{iso}}\}$ by definition, we therefore have $\Delta^{\text{iso}}(R; \mathcal{B}, w_0) = \widetilde{\Delta}^{\text{iso}}(R; \mathcal{B}, w_0)$.

In addition, recalling that $\widetilde{R}(w_0(X)) = \mathbb{E}_P[R(X) \mid w_0(X)]$, by the change of measure, the optimization problem in (B.7) can be further rewritten as

$$\begin{aligned} \Delta^{\text{iso}}(R; \mathcal{B}, w_0) \quad = \quad & \sup_{h:\ h \geq 0} \quad \mathbb{E}_{(w_0)_{\#}P}\left[h(U)\widetilde{R}(U)\right] - \mathbb{E}_{(w_0)_{\#}P}\left[\widetilde{R}(U)\right] \\ & \text{subject to} \quad (h \circ w_0)_{\#}P \in \mathcal{B}, \quad h \in \mathcal{C}_1^{\text{iso}}. \end{aligned} \tag{B.8}$$

We observe that (B.8) has the same form with the definition of $\Delta^{\text{iso}}(R; \mathcal{B})$ in (3.5), where we consider the probability measure $(w_0)_{\#}P$ instead of P and $\widetilde{R}$ in place of R, and with the specific isotonic cone $\mathcal{C}_1^{\text{iso}}$ on $\mathbb{R}$.

Applying Theorem 3.3.1 to (B.8) yields

$$\begin{aligned} \Delta^{\text{iso}}(R; \mathcal{B}, w_0) \quad = \quad & \sup_{h:\ h \geq 0} \quad \mathbb{E}_{(w_0)_{\#}P}\left[h(U)[\pi_1(\widetilde{R})](U)\right] - \mathbb{E}_{(w_0)_{\#}P}\left[\widetilde{R}(U)\right] \\ & \text{subject to} \quad (h \circ w_0)_{\#}P \in \mathcal{B}, \end{aligned}$$

where π_1 is the projection onto $\mathcal{C}_1^{\text{iso}}$ under the measure $(w_0)_{\#}P$. By definition of $\widetilde{R}$, we can rewrite this as

$$\begin{aligned} \Delta^{\text{iso}}(R; \mathcal{B}, w_0) \quad = \quad & \sup_{h:\ h \geq 0} \quad \mathbb{E}_P\left[h(w_0(X))[\pi_1(\widetilde{R})](w_0(X))\right] - \mathbb{E}_P\left[\widetilde{R}(w_0(X))\right] \\ & \text{subject to} \quad (h \circ w_0)_{\#}P \in \mathcal{B}, \end{aligned}$$

which is equal to $\Delta(\pi_1(\widetilde{R}) \circ w_0; \mathcal{B}, w_0)$ as defined in (3.7) since we also have $\mathbb{E}_P\left[\widetilde{R}(w_0(X))\right] = \mathbb{E}_P\left[[\pi_1(\widetilde{R})](w_0(X))\right]$ by (B.5). We herein complete the proof.

B.2.4 A misspecified isotonic constraint

When the true distribution shift does not obey the isotonic constraint exactly, we can nonetheless provide a bound on the worst-case excess risk, which is tighter than the (non-iso) DRL bound whenever the isotonic constraint provides a reasonable approximation.

Denote $\widetilde{w}^*$ as the underlying density ratio $\mathsf{d}P_{\text{target}}/\mathsf{d}P$ and $\Delta^*(R) = \mathbb{E}_P[\widetilde{w}^*(X)R(X)] - \mathbb{E}_P[R(X)]$ as the true excess risk. Then, we have the following connections between $\Delta^*(R)$ and $\Delta^{\text{iso}}(R;\mathcal{B})$.

Proposition B.2.2. *Assume Condition 3.2.1 holds. If $\widetilde{w}^*_{\#}P \in \mathcal{B}$ and $\widetilde{w}^* \in L_2(P)$, then we have*

$$\Delta^*(R) \le \Delta^{\text{iso}}(R;\mathcal{B}) + \mathbb{E}_P\Big[[\widetilde{w}^* - \pi(\widetilde{w}^*)](X) \cdot [R - \pi(R)](X)\Big].$$

In particular, if either $\widetilde{w}^* \in \mathcal{C}^{\text{iso}}_{\preceq}$ or $R \in \mathcal{C}^{\text{iso}}_{\preceq}$, then $\Delta^*(R) \le \Delta^{\text{iso}}(R;\mathcal{B})$.

The result states that when the isotonic constraint is violated, the worst-case excess risk of iso-DRL will be no worse than the true excess risk plus a gap which can be controlled by the correlation between $[\widetilde{w}^* - \pi(\widetilde{w}^*)](X)$ and $[R - \pi(R)](X)$. In particular, if either the risk or the true density ratio is itself isotonic (or approximately isotonic), then the gap term must be zero (or approximately zero)—and so the excess risk calculation $\Delta^{\text{iso}}(R;\mathcal{B})$, which is tighter than the non-iso DRL bound $\Delta(R;\mathcal{B})$, will never underestimate the true risk $\Delta^*(R)$ (or will only be a mild underestimate).

Proof of Proposition B.2.2

Recall that $\widetilde{w}^*$ is the underlying density ratio $\mathsf{d}P_{\text{target}}/\mathsf{d}P$. Since $\pi(\widetilde{w}^*) \overset{cvx}{\preceq} w^*$ by (B.6), and $\mathcal{B}$ is closed under the convex ordering by Condition 3.2.1, we have $\pi(\widetilde{w}^*)_{\#}P \in \mathcal{B}$; of course, we also have $\pi(\widetilde{w}^*) \in \mathcal{C}^{\text{iso}}_{\preceq}$ by definition. Therefore, $\pi(\widetilde{w}^*)$ is feasible for the optimization

problem (3.5), and so we have

$$\Delta^{\text{iso}}(R;\mathcal{B}) \geq \mathbb{E}_P\Big[[\pi(\widetilde{w}^*)](X)R(X)\Big] - \mathbb{E}_P[R(X)].$$

We therefore have

$$\Delta^*(R) = \mathbb{E}_P[\widetilde{w}^*(X)R(X)] - \mathbb{E}_P[R(X)] \leq \Delta^{\text{iso}}(R;\mathcal{B}) + \mathbb{E}_P\left[\Big(\widetilde{w}^*(X) - [\pi(\widetilde{w}^*)](X)\Big)R(X)\right].$$

Moreover,

$$\mathbb{E}_P\left[\Big(\widetilde{w}^*(X) - [\pi(\widetilde{w}^*)](X)\Big)\cdot[\pi(R)](X)\right] = \langle \pi(R), \widetilde{w}^* - \pi(\widetilde{w}^*)\rangle_P \leq 0$$

by (B.3), and so we have

$$\mathbb{E}_P\left[\Big(\widetilde{w}^*(X) - [\pi(\widetilde{w}^*)](X)\Big)R(X)\right] \leq \mathbb{E}_P\left[\Big(\widetilde{w}^*(X) - [\pi(\widetilde{w}^*)](X)\Big)\cdot\big(R(X) - [\pi(R)](X)\big)\right].$$

This completes the proof.

B.2.5 Proof of Proposition B.2.1

For any $w \geq 0$, we define the sequence of truncated functions $\{w_n\}_{n\in\mathbb{N}}$ via

$$w_n(x) = w(x)\cdot\mathbb{1}\{w(x)\leq n\} + L_n\cdot\mathbb{1}\{w(x) > n\},$$

where $L_n = \mathbb{E}[w(X) \mid w(X) > n]$. By construction for each n, $\mathbb{E}_P[w_n(X)] = 1$ and, since $\max\{n, L_n\} = L_n < \infty$, $w_n \in L_2(P)$ for each $n \geq 1$.

Step 1: Feasibility of w_n. We first prove the feasibility of w_n. To see this, as $\mathbb{E}_P[w_n(X)] = 1$ by construction, we need to show that $(w_n)_{\#}P \in \mathcal{B}$. By Condition 3.2.1, since $\mathcal{B}$ is closed

under the convex ordering, it suffices to show that

$$\mathbb{E}_P\left[\psi(w_n(X))\right] \le \mathbb{E}_P\left[\psi(w(X))\right] \qquad \text{for any convex function } \psi.$$

This is true by Jensen's inequality, since, by construction, $\mathbb{E}_P[w(X) \mid w_n(X)] = w_n(X)$.

Step 2: Convergence of $\mathbb{E}_P[w_n(X)R(X)]$. To verify the convergence of $\mathbb{E}_P[w_n(X)R(X)]$, consider

$$\begin{aligned}
&\left|\mathbb{E}_P[w_n(X)R(X)] - \mathbb{E}_P[w(X)R(X)]\right| \\
&= \left| \int_{w(x)>n} (L_n - w(x))R(x)\mathsf{d}P(x) \right| \\
&\le B_R \int_{w(x)>n} \left|L_n - w(x)\right| \mathsf{d}P(x) \\
&\le B_R \left(\int_{w(x)>n} w(x)\mathsf{d}P(x) + L_n \mathbb{P}(w(X) > n) \right) \\
&= 2\mathbb{E}_P\left[w(X) \cdot \mathbb{1}\{w(X) > n\}\right].
\end{aligned}$$

Finally, since $\mathbb{E}_P[w(X)] = 1$ (i.e., we know that $w \in L_1(P)$), this means that

$$\lim_{n\to\infty} \mathbb{E}_P\left[w(X) \cdot \mathbb{1}\{w(X) > n\}\right] = 0.$$

Conclusion. For any $\varepsilon > 0$, there exists $w \ge 0$ such that $\mathbb{E}_P[w(X)] = 1$, $w_{\#}P \in \mathcal{B}$, and

$$\mathbb{E}_P[w(X)R(X)] - \mathbb{E}_P[R(X)] \ge \Delta(R; \mathcal{B}) - \varepsilon/2$$

Then, based on Step 2, for sufficiently large n it holds that $\mathbb{E}_P[w_n(X)R(X)] \geq \mathbb{E}_P[w(X)R(X)] - \varepsilon/2$. From Step 1, we know that w_n is feasible for $\Delta_2(R;\mathcal{B})$, i.e.,

$$\Delta_2(R;\mathcal{B}) \geq \mathbb{E}_P[w_n(X)R(X)] - \mathbb{E}_P[R(X)] \geq (\mathbb{E}_P[w(X)R(X)] - \varepsilon/2) - \mathbb{E}_P[R(X)] \geq \Delta(R;\mathcal{B}) - \varepsilon.$$

Since ε is arbitrary this verifies that $\Delta_2(R;\mathcal{B}) \geq \Delta(R;\mathcal{B})$, and clearly we must have $\Delta_2(R;\mathcal{B}) \leq \Delta(R;\mathcal{B})$ by construction, which completes the proof.

B.3 Proofs of results in Section 3.4

B.3.1 Proof of Proposition 3.4.2

To prove the proposition, it suffices to show that $\|w^{*\mathrm{iso}}_{f,\rho}\|_\infty < \infty$. Recall the dual formulation. There exists a pair (λ^*, ν^*) such that

$$w^{*\mathrm{iso}}_{f,\rho}(x) = \mathcal{P}_{[0,+\infty)}\left\{(f')^{-1}\left(\frac{[\pi(R)](x) - \nu^*}{\lambda^*}\right)\right\}.$$

Note that ν^* is the parameter for standardization, thus to guarantee $\mathbb{E}_P[w^{*\mathrm{iso}}_{f,\rho}(X)] = 1$, we have

$$(f')^{-1}\left(\frac{B_R - \nu^*}{\lambda^*}\right) \geq \sup_{x\in\mathcal{X}} \; w^{*\mathrm{iso}}_{f,\rho}(x) \geq 1.$$

Moreover, it holds that $(f')^{-1}(-\nu^*/\lambda^*) \leq \min_{x\in\mathcal{X}} \; w^{*\mathrm{iso}}_{f,\rho}(x) \leq 1$. Then, combining the inequalities yields

$$-\lambda^* f'(1) \leq \nu^* \leq B_R - \lambda^* f'(1).$$

If we further have $\lambda^* \geq \underline{\lambda} > 0$, it holds that

$$\|w^{*\mathrm{iso}}_{f,\rho}\|_\infty \leq (f')^{-1}\left(\frac{B_R + \lambda^* f'(1)}{\lambda^*}\right) \leq (f')^{-1}\left(f'(1) + \frac{B_R}{\underline{\lambda}}\right) < \infty.$$

Then, it remains to prove that $\lambda^* \neq 0$. To see this, consider the KKT condition:

$$\begin{aligned} -[\pi(R)](x) + \lambda^* f'(w^{*\mathrm{iso}}_{f,\rho}(x)) + \nu^* &= 0, \\ \lambda^* \left(\mathbb{E}_P[f(w^{*\mathrm{iso}}_{f,\rho}(X))] - \rho \right) &= 0, \\ \nu^* \left(\mathbb{E}_P[w^{*\mathrm{iso}}_{f,\rho}(X)] - 1 \right) &= 0. \end{aligned}$$

If $\lambda^* = 0$, we have $[\pi(R)](X) = \nu^*$ P-almost surely, which implies that $w^{*\mathrm{iso}}_{f,\rho}(X) = 1$ P-almost surely, in which case $w^{*\mathrm{iso}}_{f,\rho}$ is also bounded. Combining pieces above, we have shown that $\|w^{*\mathrm{iso}}_{f,\rho}\|_\infty < \infty$.

B.3.2 Proof of Theorem 3.4.4

We first fix any $w \in \mathcal{C}^{\mathrm{iso}}_{\preceq}$ with $w_{\#}P \in \mathcal{B}$. By Condition 3.4.1, it holds that $w(X) \leq \Omega$ P-almost surely, and therefore without loss of generality we can assume $w \in \mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}$. Then, by definition of $\varepsilon_{\mathcal{B}}$, for any $\delta > 0$, we can find some $s, t \geq 0$ with $s + t \leq \varepsilon_{\mathcal{B}} + \delta$, such that we have $w'_{\#}\widehat{P}_n \in \mathcal{B}$ by defining $w' = (1 - s) \cdot w + t \cdot \mathbf{1}$.

Moreover, by construction, we must have $w' \in \mathcal{C}^{\mathrm{iso}}_{\preceq}$. Therefore, by optimality, we have

$$\begin{aligned} \widehat{\Delta}^{\mathrm{iso}}(\mathcal{B}) &\geq \frac{1}{n} \sum_{i=1}^n w'(X_i) r(X_i, Y_i) - \frac{1}{n} \sum_{i=1}^n r(X_i, Y_i) \\ &= \mathbb{E}_{\widehat{P}_n}[(w'(X) - 1) r(X, Y)] \\ &\geq \mathbb{E}_P[(w'(X) - 1) R(X)] - \varepsilon_R, \end{aligned}$$

where the last inequality is by the definition of ε_R. Plugging in the definition of w', we

obtain that

$$\begin{aligned}
\widehat{\Delta}^{\text{iso}}(\mathcal{B}) &\geq \mathbb{E}_P[((1-s)w(X)+t)-1)\,R(X)] - \varepsilon_R \\
&= (1-s)\mathbb{E}_P[(w(X)-1)R(X)] + (t-s)\mathbb{E}_P[R(X)] - \varepsilon_R \\
&= \mathbb{E}_P[(w(X)-1)R(X)] - (s+t)\mathbb{E}_P[w(X)R(X)] + t\mathbb{E}_P[(w(X)+1)R(X)] - \varepsilon_R \\
&\geq \mathbb{E}_P[(w(X)-1)R(X)] - 2B_R\Omega\cdot\varepsilon_{\mathcal{B}} - \varepsilon_R,
\end{aligned}$$

where the last inequality is by the fact that $\|w\|_\infty \leq \Omega$ and R is B_R-bounded, and $\Omega \geq 1$. Since this holds for every $w \in \mathcal{C}^{\text{iso}}_{\preceq}$ with $w_\# P \in \mathcal{B}$, by definition of $\Delta^{\text{iso}}(R;\mathcal{B})$, we therefore have

$$\widehat{\Delta}^{\text{iso}}(\mathcal{B}) \geq \Delta^{\text{iso}}(R;\mathcal{B}) - \varepsilon_R - 2B_R\Omega\cdot\varepsilon_{\mathcal{B}}.$$

By identical arguments, with the roles of P and $\widehat{P}_n$ reversed, we can also show that

$$\Delta^{\text{iso}}(R;\mathcal{B}) \geq \widehat{\Delta}^{\text{iso}}(\mathcal{B}) - \varepsilon_R - 2B_R\Omega\cdot\varepsilon_{\mathcal{B}},$$

which completes the proof.

B.3.3 Proof of Lemma 3.4.5

Throughout this proof we will use the notation of supervised learning, since unsupervised learning can be viewed as a special case.

In the first step, we will bound $\mathbb{E}[\varepsilon_R]$. By symmetrization (Wellner et al. (2013) Theorem

2.3.1), we have

$$\begin{aligned}\mathbb{E}[\varepsilon_R] &= \mathbb{E}\left[\sup_{w\in\mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}}\left|\mathbb{E}_{\widehat{P}_n}[(w(X)-1)r(X,Y)] - \mathbb{E}_P[(w(X)-1)R(X)]\right|\right] \\ &\le 2\mathbb{E}\left[\sup_{w\in\mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i(w(X_i)-1)r(X_i,Y_i)\right|\right],\end{aligned}$$

where σ_i's are independent Unif$\{\pm 1\}$ random variables. Since risk is B_R-bounded, by the Ledoux-Talagrand contraction lemma (Ledoux and Talagrand (2013) Theorem 4.12) applied with functions $\phi_i(t) = (t-1)\cdot r(X_i,Y_i)$, we further have

$$\mathbb{E}\left[\sup_{w\in\mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i(w(X_i)-1)r(X_i,Y_i)\right|\right] \le 2B_R\mathbb{E}\left[\sup_{w\in\mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i w(X_i)\right|\right] = 2B_R\mathcal{R}_n(\mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}).$$

Now we bound ε_R with high probability. Since risk is B_R-bounded, and any function $w \in \mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}$ is Ω-bounded, we have $(w(X)-1)r(X,Y) \in [-B_R, (\Omega-1)B_R]$, and so resampling one data point can perturb ε_R by at most $\Omega B_R/n$. Therefore, by McDiarmid's inequality (McDiarmid et al., 1989), with probability at least $1-n^{-1}$, it holds that

$$\varepsilon_R \le \mathbb{E}[\varepsilon_R] + B_R\Omega\sqrt{\frac{\log n}{2n}}.$$

Combining all these calculations yields the desired bound.

B.3.4 Proof of Lemma 3.4.6

Recall that

$$\varepsilon_{\mathcal{B}} = \sup_{w\in\mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}} \max\left\{\varepsilon_{\mathcal{B}}\left(w; P, \widehat{P}_n\right), \varepsilon_{\mathcal{B}}\left(w; \widehat{P}_n, P\right)\right\},$$

where

$$\varepsilon_{\mathcal{B}}\left(w; P_0, P_1\right) = \inf\left\{s \ge 0 \ : \ \exists\, t \ge 0,\ \left((1-s)\cdot w + t\cdot \mathbf{1}\right)_{\#} P_1 \in \mathcal{B}\right\}.$$

First, following the exact same steps as in the proof of Lemma 3.4.5, with the notation $\delta_w = \mathbb{E}_P[w(X)] - \mathbb{E}_{\widehat{P}_n}[w(X)]$, we have

$$\sup_{w \in \mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}} |\delta_w| \leq 4\mathcal{R}_n(\mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}) + \Omega\sqrt{\frac{\log n}{2n}} =: \varepsilon' \tag{B.9}$$

with probability at least $1 - n^{-1}$.

Assume the event (B.9) holds. Fix any $w \in \mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}$ with $w_{\#}P \in \mathcal{B}_{a,b}$, and define

$$w' = (1-s) \cdot w + t \cdot \mathbf{1},$$

where $s, t \geq 0$ are chosen such that $\mathbb{E}_{\widehat{P}_n}[w'(X)] = 1$, indicating that $t = s + (1-s)\delta_w$.

If $\varepsilon' = 4\mathcal{R}_n(\mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}) + \Omega\sqrt{\frac{\log n}{2n}} > \frac{1}{2}\min\{1-a, b-1\}$, then since $\varepsilon_{\mathcal{B}} \leq 1$ holds by definition, the result of the lemma must hold trivially. Therefore we can restrict our attention to the case that

$$\varepsilon' \leq \frac{1}{2}\min\{1-a, b-1\}.$$

We can further choose

$$s = 2\max\left\{\frac{\varepsilon'}{b-1}, \frac{\varepsilon'}{1-a}\right\} \geq \max\left\{\frac{\varepsilon'}{b-1-\varepsilon'}, \frac{\varepsilon'}{1-a-\varepsilon'}\right\},$$

with which, we can verify that

$$w'(X) \leq (1-s)b + t = (1-s)(b+\delta_w) + s \leq (b+\varepsilon') + s(1-b+\varepsilon') \leq b,$$

and similarly, $w'(X) \geq a$. Therefore, we have $w'_{\#}\widehat{P}_n \in \mathcal{B}_{a,b}$.

The same construction holds with the roles of P and $\widehat{P}_n$ reversed. Therefore, we can take $\varepsilon_{\mathcal{B}} = s$, which completes the proof.

B.3.5 Proof of Lemma 3.4.7

First, following the same steps (i.e., symmetrization and contraction) as in the proof of Lemma 3.4.5, we have

$$\sup_{w\in\mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}}\left|\mathbb{E}_{\widehat{P}_n}[w(X)]-\mathbb{E}_P[w(X)]\right|\le 4\mathcal{R}_n(\mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega})+\Omega\sqrt{\frac{\log n}{2n}}=:\varepsilon' \tag{B.10}$$

with probability at least $1-n^{-1}$.

Moreover, denote $t^*_f=\operatorname*{argmin}_{t\in[0,\Omega]} f(t)$. We have the decomposition

$$f(t)=f(t)\cdot\mathbb{1}\{f(t)\ge t^*_f\}+f(t)\cdot\mathbb{1}\{f(t)<t^*_f\}=:f_1+f_2,$$

where both f_1 and $-f_2$ are nondecreasing. Then, for any $g=f\circ w$ with $w\in\mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}$, we have the decomposition $g=f_1\circ w+f_2\circ w$, where $f_1\circ w\in\mathcal{C}^{\mathrm{iso}}_{\preceq}$, $-f_2\circ w\in\mathcal{C}^{\mathrm{iso}}_{\preceq}$, and both functions f_1, f_2 are L_Ω-Lipschitz. Then, by the Ledoux-Talagrand contraction lemma (Ledoux and Talagrand (2013) Theorem 4.12) applied with functions $\phi_i(t)=f(t)$, we have

$$\begin{aligned}
&\mathcal{R}_n\left(\left\{g=f\circ w:\ w\in\mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}\right\}\right)\\
&\quad\le 2\mathcal{R}_n\left(\left\{g=f\circ w:\ w\in\mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega},\ f\text{ is nondecreasing and }L_\Omega\text{-Lipschitz}\right\}\right)\\
&\quad\le 8L_\Omega\mathcal{R}_n(\mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}).
\end{aligned}$$

Hence, similar to the proof of Lemma 3.4.5, we have

$$\sup_{w\in\mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}}\left|\mathbb{E}_{\widehat{P}_n}[f(w(X))]-\mathbb{E}_P[f(w(X))]\right|\le 8L_\Omega\mathcal{R}_n(\mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega})+L_\Omega\Omega\sqrt{\frac{\log n}{2n}}=:\varepsilon'' \tag{B.11}$$

with probability at least $1-n^{-1}$.

Assume events (B.10) and (B.11) both hold. Fix any $w\in\mathcal{C}^{\mathrm{iso}}_{\preceq,\Omega}$ with $w_\#P\in\mathcal{B}_{f,\rho}$, and

define

$$w' = (1-s)\cdot w + t\cdot \mathbf{1},$$

where $s, t \in (0,1)$ are chosen such that $\mathbb{E}_{\widehat{P}_n}[w'(X)] = 1$, which implies that $t = s + (1-s)\delta_w$.

Since f is L_Ω-Lipschitz on $[0,\Omega]$,

$$f(w'(x)) \le f\left((1-s)\cdot w(x) + s\right) + L_\Omega \cdot |t-s| \le f\left((1-s)\cdot w(x) + s\right) + L_\Omega(1-s)|\delta_w|.$$

And, since f is convex with $f(1) = 0$,

$$f\left((1-s)\cdot w(x) + s\right) \le (1-s)f(w(x)) + sf(1) = (1-s)f(w(x)).$$

Combining everything, for all x, it holds that

$$f(w'(x)) \le (1-s)f(w(x)) + L_\Omega(1-s)|\delta_w| \le (1-s)\left(f(w(x)) + L_\Omega\varepsilon'\right).$$

Hence, we have

$$\mathbb{E}_{\widehat{P}_n}[f(w'(X))] \le (1-s)\mathbb{E}_{\widehat{P}_n}[f(w(X))] + (1-s)L_\Omega\varepsilon'.$$

And by assumption, $\mathbb{E}_{\widehat{P}_n}[f(w(X))] \le \mathbb{E}_P[f(w(X))] + \varepsilon'' \le \rho + \varepsilon''$, so,

$$\mathbb{E}_{\widehat{P}_n}[f(w'(X))] \le (1-s)\cdot(\rho + \varepsilon'' + L_\Omega\varepsilon') \le \rho,$$

where the last step holds by choosing

$$s = \frac{1}{\rho}(\varepsilon'' + L_\Omega\varepsilon') \ge \frac{\varepsilon'' + L_\Omega\varepsilon'}{\rho + \varepsilon'' + L_\Omega\varepsilon'}.$$

This verifies that $w'_{\#}\widehat{P}_n \in \mathcal{B}_{f,\rho}$.

The same construction holds with the roles of P and $\widehat{P}_n$ reversed. Therefore, we can take $\varepsilon_{\mathcal{B}} = s$, which completes the proof.

B.3.6 The role of the isotonic constraint

The consistency bounds developed above show that, under appropriate conditions, the error in estimating $\Delta^{\text{iso}}(R; \mathcal{B})$ can be controlled whenever the appropriate Rademacher complexity terms are small. This suggests that the isotonic constraint plays an important role: essentially, the isotonic constraint induces a form of regularization, ensuring that we work with a low-complexity class of functions.

To verify this, we now present an example with the constraint set $\mathcal{B} = \mathcal{B}_{a,b}$, without an isotonic constraint, where the estimation error of the (non-iso) DRL risk does not converge to zero.

To make the question more concrete, we will work with the bound constraint $\mathcal{B}_{a,b}$ with $0 \leq a \leq 1 \leq b$, and consider the optimization problem

$$\widehat{\Delta}(r; \mathcal{B}_{a,b}) \quad = \quad \max_{w \geq 0} \quad \frac{1}{n} \sum_{i \leq n} w(X_i) r_i - \frac{1}{n} \sum_{i \leq n} r_i \qquad \text{subject to} \quad w_{\#} \widehat{P}_n \in \mathcal{B}_{a,b},$$

which estimates the excess risk without the isotonic constraint. In other words, using $\widehat{\Delta}(r; \mathcal{B}_{a,b})$ as an empirical estimate of $\Delta(R; \mathcal{B}_{a,b})$, is analogous to using $\widehat{\Delta}^{\text{iso}}(r; \mathcal{B}_{a,b})$ as an empirical estimate of $\Delta^{\text{iso}}(R; \mathcal{B}_{a,b})$ in the presence of an additional isotonic constraint.

The following result shows that, without an isotonic constraint, this empirical estimate is not a consistent estimator of the true excess risk.

Proposition B.3.1. *Assume $R(X) = 1/2$ holds P-almost surely. Then, $\Delta(R; \mathcal{B}_{a,b}) = 0$, but with probability at least $1 - 2e^{-n/24}$, it holds that $\widehat{\Delta}(r; \mathcal{B}_{a,b}) \geq \min\{1 - a, b - 1\}/16$.*

In other words, $\widehat{\Delta}(r; \mathcal{B}_{a,b})$ is not a consistent estimator of the true excess risk $\Delta(R; \mathcal{B}_{a,b})$, since the error in the estimate is bounded away from zero (as long as $a < 1 < b$). This means

that the constraint set $\mathcal{B}_{a,b}$, on its own, is not sufficiently constrained to enable consistent estimation—while in contrast, as we have seen in our theoretical guarantees for estimation for iso-DRL, adding an isotonic constraint enables the excess risk to be estimated consistently with an empirical sample.

Proof of Proposition B.3.1

By construction, we have $r_i \sim \text{Bern}(R(X_i)) = \text{Bern}(1/2)$ independently for $i = 1, \ldots, n$. According to Section 3.2, the worst-case weights take the form $w_i = w(X_i) = c_1 \cdot \mathbb{1}\{r_i = 0\} + c_2 \cdot \mathbb{1}\{r_i = 1\}$, where $a \le c_1 \le 1 \le c_2 \le b$. Moreover, by the KKT condition, at least one of $c_1 = a$ and $c_2 = b$ holds, which implies that $c_2 - c_1 \ge \min\{1 - a, b - 1\} =: \delta$. Then, the estimated excess risk can be expressed as

$$\widehat{\Delta}(r; \mathcal{B}_{a,b}) = \frac{c_2}{n} \sum_{i \le n} r_i - \frac{1}{n} \sum_{i \le n} r_i = \frac{c_2 - 1}{n} \sum_{i \le n} r_i.$$

Since $n^{-1} \sum_{i \le n} w_i = 1$, we have

$$\frac{1}{n} \sum_{i \le n} (1 - r_i) = \frac{c_2 - 1}{c_2 - c_1},$$

which implies

$$c_2 - 1 = \frac{c_2 - c_1}{n} \sum_{i \le n} (1 - r_i) \ge \frac{\delta}{n} \sum_{i \le n} (1 - r_i).$$

In the meantime, by Chernoff bounds, with probability at least $1 - 2e^{-n/24}$, it holds that

$$\left| \frac{1}{n} \sum_{i \le n} r_i - \frac{1}{2} \right| \le \frac{1}{4}.$$

Then, for the excess risk, with probability at least $1 - 2e^{-n/24}$, it holds that

$$\widehat{\Delta}(r; \mathcal{B}_{a,b}) = \frac{c_2 - 1}{n} \sum_{i \le n} r_i \ge \delta \left(\frac{1}{n} \sum_{i \le n} (1 - r_i) \right) \cdot \left(\frac{1}{n} \sum_{i \le n} r_i \right) \ge \frac{\delta}{16}.$$

B.4 Additional simulation results

B.4.1 Simulations for iso-DRL under componentwise order

In Section 3.5, we mainly focused on the partial order with respect to $w_0(x)$. In this section, to demonstrate the effect of various choices of the partial (pre)order, we further consider an alternative choice of the partial (pre)order: the componentwise order where

$$x \preceq x' \qquad \text{if and only if} \qquad x_j \le x'_j, \text{ for all } j \in [m],$$

where we set $m = 5 < d = 20$. Let iso-DRL-comp denote the CP interval with calibrated target level $\alpha'_{\text{iso}} = \max\{0, \alpha - \widetilde{\Delta}^{\text{iso}}\}$, where

$$\begin{aligned} \widetilde{\Delta}^{\text{iso}} = \max \quad & \frac{1}{n} \sum_{i \in \mathcal{D}_3} w_i \widetilde{r}_i^{\text{iso}} - \frac{1}{n} \sum_{i \in \mathcal{D}_3} r_i \\ \text{subject to} \quad & \frac{1}{n} \sum_{i \in \mathcal{D}_3} w_i = 1, \quad \frac{1}{n} \sum_{i \in \mathcal{D}_3} w_i \log w_i \le \rho, \quad 0 \le w_i \le \Omega, \end{aligned}$$

and $(\widetilde{r}_i)_{i \in \mathcal{D}_3}$ is the isotonic projection of $(r_i)_{i \in \mathcal{D}_3}$ with respect to the componentwise order.

We follow exactly the same settings with Section 3.5.1 with $n_{\text{pre}} = 50$ and vary ρ in $[0.002, 6]$. From Figure B.1 and B.2, each of the coverage rate and average interval width of iso-DRL-comp lies between that of DRL and iso-DRL-w_0, which indicates that additional constraints will relieve the conservativeness of DRL, but only a proper choice of the partial (pre)order will lead to desired performance close to the oracle weighted CP.

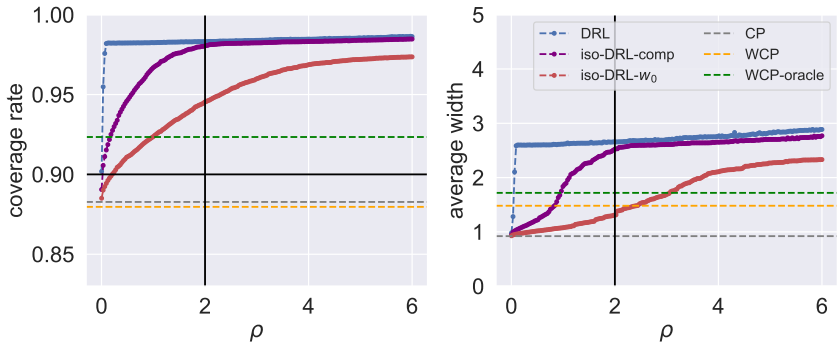


Figure B.1: Results with varying ρ in the well-specified setting. The solid vertical line denotes an estimate $\widehat{\rho}$ of the KL divergence, $D_{\mathrm{KL}}(P_{\mathrm{target}}\|P)$ (See Appendix B.4.2 for details). The solid horizontal line (in the left-hand plot) marks the nominal coverage level, $1-\alpha = 90\%$.

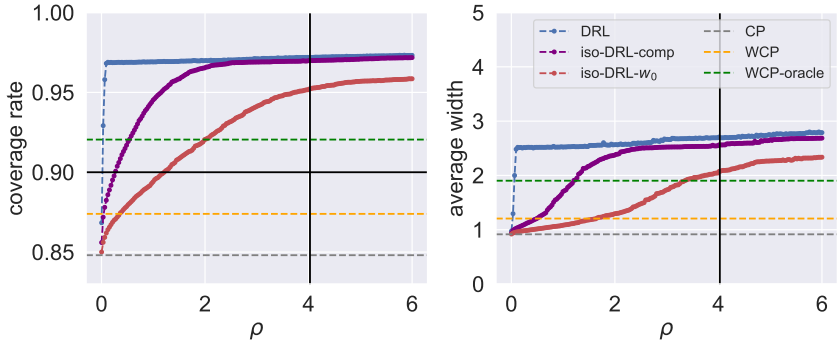


Figure B.2: Results with varying ρ in the misspecified setting. The solid vertical line denotes an estimate $\widehat{\rho}$ of the KL divergence, $D_{\mathrm{KL}}(P_{\mathrm{target}}\|P)$ (See Appendix B.4.2 for details). The solid horizontal line (in the left-hand plot) marks the nominal coverage level, $1-\alpha = 90\%$.

B.4.2 Details for the wine quality data set: a proxy of the oracle KL-divergence

In this section, we examine the choice of ρ in the wine quality data experiment from Section 3.5.2. In a real data setting, the true KL divergence, $D_{\mathrm{KL}}\left(P_{\mathrm{target}}\|P\right)$, is of course unknown, so we need to use a data-driven choice of ρ in order to implement a DRL procedure (with or without an isotonic constraint).

As is shown in Section 3.5.2, we denote $\widehat{w}_{\mathrm{kde}}$ as the density ratio obtained by kernel density estimation (Gaussian kernel with bandwidth 0.125). Accordingly, let $\mathsf{d}\widehat{Q}_{\mathrm{kde}} = \widehat{w}_{\mathrm{kde}} \cdot \mathsf{d}P$ be an estimate of P_{target}. With a subsample $\{X_i\}_{i\leq K}$ drawn the group of white wine

(data distribution P), a reasonable value for $\widehat{\rho}$ (i.e., an estimate of the true divergence ρ between the distributions P and P_{target}) can be calculated by

$$\begin{aligned}\widehat{\rho} &= \frac{1}{K}\sum_{i\le K}\widehat{w}_{\text{kde}}(X_i)\log\left(\widehat{w}_{\text{kde}}(X_i)\right)\\ &\approx \mathbb{E}_P\left\{\frac{\mathsf{d}\widehat{Q}_{\text{kde}}}{\mathsf{d}P}\log\left(\frac{\mathsf{d}\widehat{Q}_{\text{kde}}}{\mathsf{d}P}\right)\right\} = D_{\text{KL}}\left(\widehat{Q}_{\text{kde}}\|P\right).\end{aligned}$$

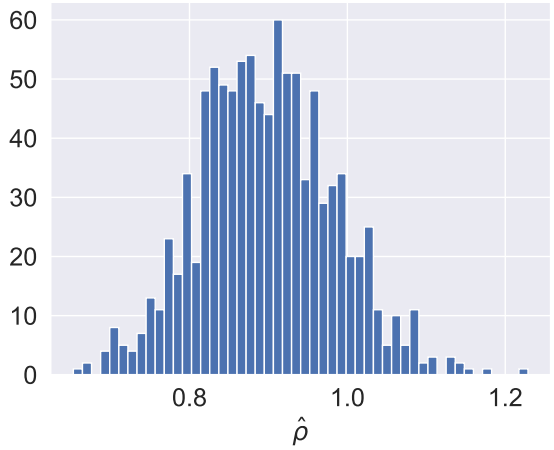


Figure B.3: Histogram of $\widehat{\rho}$. (See Appendix B.4.2 for details.)

To show the range for values of $\widehat{\rho}$, we repeatedly fit KDE on the 80% samples from each group (white and red wine groups respectively). Figure B.3 shows the histogram of $\widehat{\rho}$ with 1000 repetitions, of which the median is approximately 0.8950—this is the value of ρ used in our preview of the wine quality data experiment, shown in Figure 3.1.

APPENDIX C

APPENDIX FOR CHAPTER 4

C.1 Experiments: details and extensions

Reproducibility. Our code and data are included in the supplemental materials.

C.1.1 Experiment setup and details

Fixed encoder network. In Figure 4.3, we freeze the encoder network (ResNet-18) trained and saved in `https://github.com/sthalles/SimCLR` to focus on the behavior of projector under SimCLR loss. The pretrained architecture is trained with the default temperature $\tau = 0.07$ and the following composite augmentation

Listing C.1: Image augmentation

```
class TransformsSimCLR:
    """
    A stochastic data augmentation module that transforms any given data exam
    resulting in two correlated views of the same example,
    denoted x_i and x_j, which we consider as a positive pair.
    """

    def __init__(self, size, aug_str=1):
        color_jitter = torchvision.transforms.ColorJitter(
            0.8 * aug_str, 0.8 * aug_str, 0.8 * aug_str, 0.2 * aug_str
        )
        self.train_transform = torchvision.transforms.Compose(
            [
```

```
                torchvision.transforms.RandomResizedCrop(size=size),
                torchvision.transforms.RandomHorizontalFlip(),
                torchvision.transforms.RandomApply([color_jitter], p=0.8),
                torchvision.transforms.RandomGrayscale(p=0.2),
                torchvision.transforms.ToTensor(),
            ]
        )

        self.test_transform = torchvision.transforms.Compose(
            [
                torchvision.transforms.Resize(size=size),
                torchvision.transforms.ToTensor(),
            ]
        )

    def __call__(self, x):
        return self.train_transform(x), self.train_transform(x)
```

Training details. For simplicity, we extract and then center pretrained features of 10-class `STL`-10 images (which conforms to the zero-mean assumption in 2-GMM). We train the projector $\mathbf{W} \in \mathbb{R}^{512\times512}$ for 50 epochs with the batch size of 64.

In all experiments reported in the paper, we use one-layer linear projector $\mathbf{W} \in \mathbb{R}^{p\times p}$ without the bias term, and this matrix is initialized by a random orthogonal matrix (orthogonal initialization avoids potential optimization artifacts (Xiao et al., 2018; Arora et al., 2018). We train the linear projector using the standard SimCLR loss function `https://github.com/sthalles/SimCLR`. For downstream accuracy, we use "linear_model.LogisticRegression"

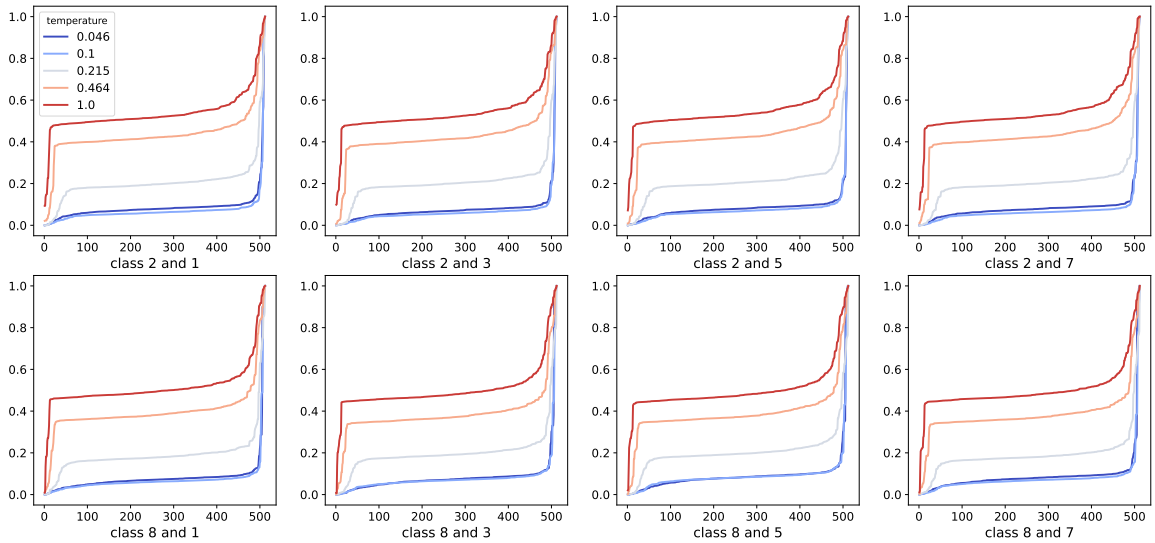


Figure C.1: Cumulative sums of alignment scores $\kappa_{j,c_1,c_2} = \langle \boldsymbol{v}_j, \boldsymbol{\mu}_{c_1,c_2} \rangle^2 / \|\boldsymbol{\mu}_{c_1,c_2}\|^2$ for different pairs with the pretrained encoder and a one-layer linear projector $\mathbf{W} \in \mathbb{R}^{512\times512}$ under the standard SimCLR loss (ResNet-18 and 10-class `STL`-10 dataset).

from `sklearn` with very small ℓ_2 regularization (choosing $C = 1000$), with the aim to approximate the max-margin classifier when data are linearly separable (Rosset et al., 2003).

To assess the behavior of $\mathbf{W}$ and the transition between expansion and shrinkage regimes, 10 values of $\log(\tau)$ are chosen from equi-spaced grids in $[\log(0.01), \log(10)]$ and two values of the augmentation strength (0.1 and 1.0) are chosen to represent small and moderate augmentation respectively.

Results for other pairs and deeper encoder. Since the projector $\mathbf{W}$ is trained with the 10-class `STL`-10 dataset, then the plots for singular values of $\mathbf{W}$ and downstream accuracy are the same as is shown in Figure 4.3. In Figure C.1, the cumulative sums of alignment scores are shown for different pairs which show similar patterns: as τ increases, the expansion effect is gradually gained.

In addition, we also experimented with ResNet-50 as the encoder instead of ResNet-18. Here we present results using ResNet-50 in Figure C.2a, which is similar to Figure 4.3. We also presented results for different pairs in C.2b.

We generate $n = 2000$ data $(\boldsymbol{x}_i)_{i\le n}$ from 2-GMM with $p = 100$ and $\boldsymbol{\mu} = 4\boldsymbol{e}_1$ where $(\boldsymbol{e}_k)_{k\le p}$ forms the canonical basis. The left plot shows the first two coordinates of these data points, which is equivalent to projecting data onto $\mathrm{span}(\boldsymbol{e}_1, \boldsymbol{e}_2)$.

We add random perturbation $\mathcal{N}(0, \sigma_{\mathrm{aug}}^2 \mathbf{I}_p)$ to form augmented data, and then we train a linear projector $\mathbf{W}$ and the standard SimCLR loss on $(\boldsymbol{x}_i)_{i\le n}$. After training, we calculate embeddings $\boldsymbol{z}_i = \mathbf{W}\boldsymbol{x}_i$, project the embeddings onto $\mathrm{span}(\mathbf{W}\boldsymbol{e}_1, \mathbf{W}\boldsymbol{e}_2)$ and visualize these 2D projections. We plot the embeddings under an archetypal expansion regime (middle plot, $\sigma_{\mathrm{aug}} = 0.1, \tau = 0.2$) and an archetypal shrinkage regime (right plot, $\sigma_{\mathrm{aug}} = 1, \tau = 20$).

To aid visualization, we add circles in each of the three plots. Note that the normalized embedding $\boldsymbol{z}_i/\|\boldsymbol{z}_i\|$ is used for calculating the cosine similarities in the SimCLR loss. We can interpret the plots using the alignment vs. uniformity perspective (Wang and Isola, 2020): in the expansion regime. the alignment loss is the dominant term and forces concentrations of normalized embeddings, whereas in the shrinkage regime, the uniformity loss is the dominant term and encourages normalized embeddings to be evenly spread.

C.1.2 Additional experiments

Beyond fixed encoders: training SimCLR from scratch. To explore the expansion/shrinkage phenomenon without freezing the encoder component, we train the entire architecture (e.g., ResNet-50 encoder and a one-layer linear projector) on `STL-10` train dataset with 400 epochs and the batch size of 256. We choose 10 values of $\log(\tau)$ from equi-spaced grids in $[\log(0.01), \log(10)]$ and we choose the augmentation as the default value 1.0.

As shown in Figure C.3, visibly, the dimensional collapse phenomenon is evident in both the feature space and the projector. Our GMM theory does not apply directly to this scenario since in Section 4.4 we assume that features are generated from a full-dimensional mixture model. Still, our expansion/shrinkage analysis provides partial explanations as summarized below.

1. When we train the encoder and projector simultaneously, their roles and effects are not distinctly separated. Indeed, there are many ways to express $\boldsymbol{g}_{\boldsymbol{\varphi}} \circ \boldsymbol{f}_{\boldsymbol{\theta}}$ as function compositions. Thus, the dimensional collapse in the feature space can be interpreted as a shrinkage effect induced by the last few layers in the encoder. Understanding how dimensional collapse emerges progressively across layers is an interesting research direction.

2. Our analysis still provides useful information about downstream accuracy when both the encoder and the projector are trained. For example, when we vary the temperature parameter, the severity of the collapse is correlated with the downstream accuracy; see Figure C.3.

3. Our theory matches the empirical results if we restrict the linear transform $\mathbf{W}$ on the subspace that the feature/embedding vectors span. Figure C.4 shows that the singular values/vectors of the restricted linear transform. Note that we recover similar cumulative score plots as in the fixed encoder scenario.

Below we provide more detailed explanations for point 2 and 3.

First, from Figure C.3 (left), we can see that the downstream accuracy using features is higher than that using embeddings, which validates the practice of using only the features before the projector for classification. When $\tau \leq 0.129$, the difference between two curves is decreasing and both achieve the highest value at $\tau = 0.129$. However, when τ further increases, which disagrees with our previous findings, both accuracy start to decrease. This phenomenon can be explained by the plots of singular values. As we can see from Figure C.3 (middle), the features already have dimensional collapse with the one-layer linear projector even when τ is small, but the collapse is relatively moderate when $\tau \leq 0.129$, which refers to the bunch of curves starting to drop after the index of 1000. When $\tau > 0.129$, the collapse becomes much more severe and the effective rank decreases fast below 200 when τ goes to

1.0. The singular values of embeddings change accordingly.

The trend in the downstream task accuracy together with the changes in singular values of $\mathbf{W}$ convey the message that

- when τ is moderate, the increase in τ, which enhances the expansion of signal (will be shown in the following figures), will improve the downstream task accuracy with embeddings, making it as good as the accuracy with features even when features are undergoing the dimensional collapse;

- when τ is large, it poses negative effects in downstream task accuracy in that the features are already low-rank as is shown in Tian (2022), which may lead to the information loss in the data and may further do harm to the training of projector. As a result, the accuracy with either embeddings or features decreases. Also, the benefits from the expansion of signal are surpassed and accuracy with embeddings can be worse than that with features. In contrast, with pretrained model and full-rank features, the accuracy with embeddings can be better than that with features with the benefits from expansion as is shown in Figure 4.3.

In Figure C.4, we first write the SVD of the feature matrix $\mathbf{H}_0$ as $\mathbf{H}_0 = \mathbf{U}(\mathbf{H}_0)\mathbf{D}(\mathbf{H}_0)\mathbf{V}^\top(\mathbf{H}_0)$, where $\mathbf{D}(\mathbf{H}_0)$ has singular values of $\mathbf{H}_0$ as diagonal elements: $\{s_k(\mathbf{H}_0) : \ 1 \leq k \leq p\}$. We consider projecting both $\mathbf{W}$ and $\boldsymbol{\mu}$ onto the features' top-r right singular subspace, that is we define $\mathbf{W}^{(r)} = \mathbf{W}\mathbf{V}_r(\mathbf{H}_0)$ and $\boldsymbol{\mu}^{(r)} = \mathbf{V}_r^\top(\mathbf{H}_0)\boldsymbol{\mu}$ where $\mathbf{V}_r(\mathbf{H}_0) \in \mathbb{R}^{p\times r}$ is the submatrix consisting of the first r columns of $\mathbf{V}(\mathbf{H}_0)$. We choose r by $r = \text{argmax}k : s_k(\mathbf{H}_0) \geq 10^{-3}\}$. {
Then, we can instead calculate the SVD of $\mathbf{W}^{(r)} = \mathbf{U}^{(r)}\mathbf{D}^{(r)}\mathbf{V}^{(r)\top}$. Denote $\boldsymbol{v}_j^{(r)}$ as the jth right singular vector of $\mathbf{W}^{(r)}$.

In the left column, we plot the unsquared alignment score $\langle \boldsymbol{v}_j^{(r)}, \boldsymbol{\mu}_{c_1,c_2}^{(r)} \rangle / \|\boldsymbol{\mu}_{c_1,c_2}^{(r)}\|$, where $(c_1, c_2) = (7, 2)$, and the cumulative scores are plotted in the middle column. With the truncated scores, we can see that when $\tau = 0.077$, the bottom singular vectors align much

better with the $\boldsymbol{\mu}_{c_1,c_2}^{(r)}$ than top ones (shrinkage regime). As τ increases, the expansion and shrinkage effects are comparable to each other with $\tau = 0.219$, but when τ further increases to 0.599, top singular vectors align better with $\boldsymbol{\mu}_{c_1,c_2}^{(r)}$ and the expansion effect is dominating. However, expansion effect directly enhances downstream accuracy with full-rank features as is shown in the pretrained experiments and the benefits can be hidden by the dimensional collapse of features as is shown in Figure C.3.

Convergence of the encoder Recall the decomposition

$$\mathcal{L}(\boldsymbol{\theta}^{(t)}, \mathbf{W}^{(t)}) = \mathcal{L}(\boldsymbol{\theta}^{(t)}, \widetilde{\mathbf{W}}^{(t)}) + \left(\mathcal{L}(\boldsymbol{\theta}^{(t)}, \mathbf{W}^{(t)}) - \mathcal{L}(\boldsymbol{\theta}^{(t)}, \widetilde{\mathbf{W}}^{(t)})\right),$$

where $(\boldsymbol{\theta}^{(t)}, \mathbf{W}^{(t)})$ are the parameters at epoch t, and $\widetilde{\mathbf{W}}^{(t)} \coloneqq \underset{\mathbf{W}}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{\theta}^{(t)}, \mathbf{W})$, i.e. the optimal projector with frozen encoder parameters $\boldsymbol{\theta}^{(t)}$ at epoch t. We observe that when $t \geq T_0$ for $T_0 \approx 50$, we observe $\|\widetilde{\mathbf{W}}^{(t)} - \mathbf{W}^{(t)}\| \ll \|\mathbf{W}^{(t)}\|$.

Feature-level vs. image-level augmentation Image-level augmentation results in feature-level perturbation in that with the composite augmentation and the fixed encoder $\boldsymbol{f}_{\boldsymbol{\theta}_n^*}$, for each image $\boldsymbol{x}_i$, we have the augmented image $\boldsymbol{x}_i^+$ and the mapped features $\boldsymbol{h}_i = \boldsymbol{f}_{\boldsymbol{\theta}_n^*}(\boldsymbol{x}_i)$, $\boldsymbol{h}_i^+ = \boldsymbol{f}_{\boldsymbol{\theta}_n^*}(\boldsymbol{x}_i^+)$. Then, the image-level augmentation is associated with the feature-level perturbation $\boldsymbol{e}_i = \boldsymbol{h}_i^+ - \boldsymbol{h}_i$. This perturbation is correlated with $\boldsymbol{h}_i$ and can be complicated. To make a connection with our 2-GMM theory, we also consider the homogeneous feature-level augmentation and its effect on the expansion/shrinkage phenomenon.

We freeze the encoder network (ResNet-50) as before, then the features are also fixed in this case. For each $\boldsymbol{h}_{0,i}$, we add Gaussian perturbation $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma_{\text{aug}}^2 \mathbf{I}_p)$ independently to obtain $\boldsymbol{h}_i^+ = \boldsymbol{h}_{0,i} + \boldsymbol{\varepsilon}_i$. Then, with the same training process, we train the projector $\mathbf{W} \in \mathbb{R}^{2048\times 2048}$ without the bias term under the standard SimCLR loss for 50 epochs with the batch size of 64. We choose 10 values of $\log(\tau)$ from equi-spaced grids in $[\log(0.01), \log(10)]$

and we choose the augmentation as the default value 1.0.

From Figure C.6 (middle), different from the results for image-level augmentation, we can see the dominating expansion effect when $\tau > 0.215$ where the bottom singular vectors of $\mathbf{W}$ is merely uncorrelated with $\boldsymbol{\mu}_{c_1,c_2}$ while the top singular values contribute over 90% of the correlation.

C.2 Proofs for Section 4.4

C.2.1 Proof of Proposition 4.4.1

Recall that the population loss $\mathcal{L}(\mathbf{W}) = \mathbb{E}[\mathcal{L}_n(\mathbf{W})]$ obeys

$$\mathcal{L}(\mathbf{W}) \coloneqq \mathcal{L}_{\text{align}}(\mathbf{W}) + \mathcal{L}_{\text{unif}}(\mathbf{W}),$$

where we denote

$$\mathcal{L}_{\text{align}}(\mathbf{W}) \coloneqq \frac{1}{\tau}\mathbb{E}_{\boldsymbol{h},\boldsymbol{h}^+}\left[\frac{1}{2}\frac{\|\mathbf{W}\boldsymbol{h} - \mathbf{W}\boldsymbol{h}^+\|_2^2}{\left(\mathbb{E}[\|\mathbf{W}\boldsymbol{h}\|_2^2]\cdot\mathbb{E}[\|\mathbf{W}\boldsymbol{h}^+\|_2^2]\right)^{1/2}}\right];$$
$$\mathcal{L}_{\text{unif}}(\mathbf{W}) \coloneqq \log\left(\mathbb{E}_{\boldsymbol{h}^-,\boldsymbol{h}}\left[\exp\left(-\frac{1}{2\tau}\frac{\|\mathbf{W}\boldsymbol{h} - \mathbf{W}\boldsymbol{h}^-\|_2^2}{\left(\mathbb{E}[\|\mathbf{W}\boldsymbol{h}\|_2^2]\cdot\mathbb{E}[\|\mathbf{W}\boldsymbol{h}^-\|_2^2]\right)^{1/2}}\right)\right]\right).$$

In what follows, we compute $\mathcal{L}_{\text{align}}(\mathbf{W})$ and $\mathcal{L}_{\text{unif}}(\mathbf{W})$.

Computing $\mathcal{L}_{\text{align}}(\mathbf{W})$. For the first term, we have

$$\mathbb{E}[\|\mathbf{W}\boldsymbol{h}\|_2^2] = \mathbb{E}[\|\mathbf{W}\boldsymbol{h}^+\|_2^2] = (1+\sigma_{\text{aug}}^2)\|\mathbf{W}\|_{\text{F}}^2 + \|\mathbf{W}\boldsymbol{\mu}\|_2^2 = \alpha,$$

where the last relation is the definition of α. Therefore, $\mathcal{L}_{\text{align}}$ can be simply written as

$$\mathcal{L}_{\text{align}}(\mathbf{W}) = \frac{1}{2\tau\alpha}\mathbb{E}_{\boldsymbol{h},\boldsymbol{h}^+}\left[\|\mathbf{W}\boldsymbol{h} - \mathbf{W}\boldsymbol{h}^+\|_2^2\right].$$

In addition, since $\boldsymbol{h} = \boldsymbol{h}_0 + \boldsymbol{\varepsilon}$, and $\boldsymbol{h}^+ = \boldsymbol{h}_0 + \boldsymbol{\varepsilon}^+$ with $\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}^+ \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\text{aug}}\mathbf{I})$, we obtain

$$\boldsymbol{h} - \boldsymbol{h}^+ = \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^+ \sim \mathcal{N}(0, 2\sigma_{\text{aug}}^2\mathbf{I}),$$

which yields

$$\mathbb{E}_{\boldsymbol{h},\boldsymbol{h}^+}\left[\|\mathbf{W}\boldsymbol{h} - \mathbf{W}\boldsymbol{h}^+\|_2^2\right] = 2\sigma_{\text{aug}}^2\|\mathbf{W}\|_{\text{F}}^2.$$

This further leads to the simplified expression for $\mathcal{L}_{\text{align}}(\mathbf{W})$:

$$\mathcal{L}_{\text{align}}(\mathbf{W}) = \frac{\sigma_{\text{aug}}^2\|\mathbf{W}\|_{\text{F}}^2}{\tau\alpha}.$$

Computing $\mathcal{L}_{\text{unif}}(\mathbf{W})$. For the second term $\mathcal{L}_{\text{unif}}(\mathbf{W})$, using the definition of α, we see that

$$\mathcal{L}_{\text{unif}}(\mathbf{W}) = \log\left(\mathbb{E}_{\boldsymbol{h}^-,\boldsymbol{h}}\left[\exp\left(-\frac{\|\mathbf{W}\boldsymbol{h} - \mathbf{W}\boldsymbol{h}^-\|_2^2}{2\tau\alpha}\right)\right]\right). \tag{C.1}$$

Recall that $\boldsymbol{h}$ and $\boldsymbol{h}^-$ are two independent draws from the Gaussian mixture model, i.e.,

$$\boldsymbol{h} = y\boldsymbol{\mu} + \sqrt{1+\sigma_{\text{aug}}^2}\boldsymbol{g}, \quad \boldsymbol{h}^- = y^-\boldsymbol{\mu} + \sqrt{1+\sigma_{\text{aug}}^2}\boldsymbol{g}^-,$$

where y and y^- are independent Rademacher random variables, $\boldsymbol{g}$ and $\boldsymbol{g}^-$ are two independent random vectors from $\mathcal{N}(0, \mathbf{I})$. This further yields

$$\begin{aligned}\boldsymbol{h} - \boldsymbol{h}^- &\overset{\mathsf{d}}{=} (y - y^-)\boldsymbol{\mu} + \sqrt{2(1+\sigma_{\text{aug}}^2)}\boldsymbol{g} \\ &\overset{\mathsf{d}}{=} \frac{1}{2}\delta_0 + \frac{1}{4}\delta_{2\boldsymbol{\mu}} + \frac{1}{4}\delta_{-2\boldsymbol{\mu}} + \sqrt{2(1+\sigma_{\text{aug}}^2)}\boldsymbol{g}.\end{aligned}$$

As a result, we obtain

$$\mathbb{E}_{\boldsymbol{h}^-,\boldsymbol{h}}\left[\exp\left(-\frac{\|\mathbf{W}\boldsymbol{h}-\mathbf{W}\boldsymbol{h}^-\|_2^2}{2\tau\alpha}\right)\right] = \mathbb{E}_{\boldsymbol{g}}\left[\frac{1}{2}\exp\left(-\frac{(1+\sigma_{\text{aug}}^2)\|\mathbf{W}\boldsymbol{g}\|_2^2}{\tau\alpha}\right)+\frac{1}{2}\exp\left(-\frac{\|\sqrt{2}\mathbf{W}\boldsymbol{\mu}+\sqrt{1+\sigma_{\text{aug}}^2}\mathbf{W}\boldsymbol{g}\|_2^2}{\tau\alpha}\right)\right] \tag{C.2}$$

Let $\sum_{j=1}^p \sigma_j \boldsymbol{u}_j\boldsymbol{v}_j^\top$ be the singular value decomposition of $\mathbf{W}$. Then one has

$$\mathbf{W}\boldsymbol{g}=\sum_{j=1}^p \sigma_j \boldsymbol{v}_j^\top \boldsymbol{g}\ \boldsymbol{u}_j, \qquad \|\mathbf{W}\boldsymbol{g}\|_2^2=\sum_{j=1}^p \sigma_j^2\left(\boldsymbol{v}_j^\top\boldsymbol{g}\right)^2,$$

and

$$\|\sqrt{2}\mathbf{W}\boldsymbol{\mu}+\sqrt{1+\sigma_{\text{aug}}^2}\mathbf{W}\boldsymbol{g}\|_2^2=\sum_{j=1}^p \sigma_j^2\left(\boldsymbol{v}_j^\top\left(\sqrt{2}\boldsymbol{\mu}+\sqrt{1+\sigma_{\text{aug}}^2}\boldsymbol{g}\right)\right)^2.$$

Taking expectations, we arrive at

$$\begin{aligned}\mathbb{E}_{\boldsymbol{g}}\left[\exp\left(-\frac{(1+\sigma_{\text{aug}}^2)\|\mathbf{W}\boldsymbol{g}\|_2^2}{\tau\alpha}\right)\right] &= \mathbb{E}_{\boldsymbol{g}}\left[\exp\left(-\sum_{j=1}^d \frac{(1+\sigma_{\text{aug}}^2)\sigma_j^2}{\tau\alpha}\left(\boldsymbol{v}_j^\top\boldsymbol{g}\right)^2\right)\right]\\ &= \prod_{j=1}^p \mathbb{E}_{\boldsymbol{g}}\left[\exp\left(-\frac{(1+\sigma_{\text{aug}}^2)\sigma_j^2}{\tau\alpha}\left(\boldsymbol{v}_j^\top\boldsymbol{g}\right)^2\right)\right],\end{aligned}$$

where the last equality uses the independence among $\{\boldsymbol{v}_j^\top\boldsymbol{g}\}_{1\le j\le p}$. Since $(\boldsymbol{v}_j^\top\boldsymbol{g})^2$ is a χ_1^2 random variable, we can use the moment generating function of χ_1^2 to obtain

$$\mathbb{E}_{\boldsymbol{g}}\left[\exp\left(-\frac{(1+\sigma_{\text{aug}}^2)\sigma_j^2}{\tau\alpha}\left(\boldsymbol{v}_j^\top\boldsymbol{g}\right)^2\right)\right]=\left(1+\frac{2(1+\sigma_{\text{aug}}^2)\sigma_j^2}{\tau\alpha}\right)^{-1/2}.$$

Combining the previous two relations yields

$$\mathbb{E}_{\boldsymbol{g}}\left[\exp\left(-\frac{(1+\sigma_{\text{aug}}^2)\|\mathbf{W}\boldsymbol{g}\|_2^2}{\tau\alpha}\right)\right]=\prod_{j=1}^p\left(1+\frac{2(1+\sigma_{\text{aug}}^2)\sigma_j^2}{\tau\alpha}\right)^{-1/2}. \tag{C.3}$$

Using a similar decomposition, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{g}}\left[\exp\left(-\frac{\|\sqrt{2}\mathbf{W}\boldsymbol{\mu}+\sqrt{1+\sigma_{\text{aug}}^2}\mathbf{W}\boldsymbol{g}\|_2^2}{\tau\alpha}\right)\right] \\
&\quad=\prod_{j=1}^{p}\mathbb{E}_{\boldsymbol{g}}\left[\exp\left(-\frac{\sigma_j^2}{\tau\alpha}\left(\sqrt{2}\boldsymbol{v}_j^\top\boldsymbol{\mu}+\sqrt{1+\sigma_{\text{aug}}^2}\boldsymbol{v}_j^\top\boldsymbol{g}\right)^2\right)\right]. \qquad \text{(C.4)}
\end{aligned}$$

Elementary calculations tell us that for any $t>0$ and any $a\in\mathbb{R}$, one has

$$\mathbb{E}_{u\sim N(0,1)}\left[\exp\left(-t(a+u)^2\right)\right]=\frac{1}{\sqrt{1+2t}}\exp\left(-\frac{a^2t}{2t+1}\right).$$

This allows us to simplify each term in (C.4) as

$$\begin{aligned}
&\mathbb{E}_{u\sim N(0,1)}\left[\exp\left(-\frac{\sigma_j^2}{\tau\alpha}\left(\sqrt{2}\boldsymbol{v}_j^\top\boldsymbol{\mu}+\sqrt{1+\sigma_{\text{aug}}^2}u\right)^2\right)\right] \\
&\quad=\frac{1}{\sqrt{1+\frac{2(1+\sigma_{\text{aug}}^2)\sigma_j^2}{\tau\alpha}}}\exp\left(-\frac{2\sigma_j^2}{2(1+\sigma_{\text{aug}}^2)\sigma_j^2+\tau\alpha}(\boldsymbol{v}_j^\top\boldsymbol{\mu})^2\right).
\end{aligned}$$

The previous two displays taken together lead to

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{g}}\left[\exp\left(-\frac{\|\sqrt{2}\mathbf{W}\boldsymbol{\mu}+\sqrt{1+\sigma_{\text{aug}}^2}\mathbf{W}\boldsymbol{g}\|_2^2}{\tau\alpha}\right)\right] \\
&=\prod_{j=1}^{p}\left(1+\frac{2(1+\sigma_{\text{aug}}^2)\sigma_j^2}{\tau\alpha}\right)^{-1/2}\exp\left(-\sum_{j=1}^{p}\frac{2\sigma_j^2}{2(1+\sigma_{\text{aug}}^2)\sigma_j^2+\tau\alpha}(\boldsymbol{v}_j^\top\boldsymbol{\mu})^2\right). \qquad \text{(C.5)}
\end{aligned}$$

Substitute (C.3) and (C.5) into the identities (C.2) and (C.1) to see that

$$\mathbb{E}_{\boldsymbol{h}^-,\boldsymbol{h}}\left[\exp\left(-\frac{\|\mathbf{W}\boldsymbol{h}-\mathbf{W}\boldsymbol{h}^-\|_2^2}{2\tau\alpha}\right)\right]$$
$$=\frac{1}{2}\prod_{j=1}^{p}\left(1+\frac{2(1+\sigma_{\text{aug}}^2)\sigma_j^2}{\tau\alpha}\right)^{-1/2}\left(1+\exp\left(-\sum_{j=1}^{p}\frac{2\sigma_j^2}{2(1+\sigma_{\text{aug}}^2)\sigma_j^2+\tau\alpha}(\boldsymbol{v}_j^\top\boldsymbol{\mu})^2\right)\right),$$

and

$$\mathcal{L}_{\text{unif}}(\mathbf{W})=-\frac{1}{2}\sum_{j=1}^{p}\log\left(1+\frac{2(1+\sigma_{\text{aug}}^2)\sigma_j^2}{\tau\alpha}\right)$$
$$+\log\left(1+\exp\left(-\sum_{j}\frac{2\sigma_j^2}{2(1+\sigma_{\text{aug}}^2)\sigma_j^2+\tau\alpha}(\boldsymbol{v}_j^\top\boldsymbol{\mu})^2\right)\right)-\log 2.$$

These complete the proof of Proposition 4.4.1.

C.2.2 Justification of approximate loss

Here we present justification underlying the approximation (4.3). In the regime where $\tau\alpha=\|\mathbf{W}\|_{\text{F}}^2+\|\mathbf{W}\boldsymbol{\mu}\|_2^2\gg\sigma_j^2$ for each j, we can use the approximations $\log(1+x)\approx x$ and $2(1+\sigma_{\text{aug}}^2)\sigma_j^2+\tau\alpha\approx\tau\alpha$ to obtain the approximate loss

$$\widetilde{\mathcal{L}}_{\text{unif}}(\mathbf{W})=-\sum_{j=1}^{p}\frac{(1+\sigma_{\text{aug}}^2)\sigma_j^2}{\tau\alpha}+\log\left(1+\exp\left(-\sum_{j}\frac{2\sigma_j^2}{\tau\alpha}(\mathbf{v}_j^\top\boldsymbol{\mu})^2\right)\right)-\log 2$$
$$=-\frac{(1+\sigma_{\text{aug}}^2)\|\mathbf{W}\|_{\text{F}}^2}{\tau\alpha}+\log\left(1+\exp\left(-\frac{2\|\mathbf{W}\boldsymbol{\mu}\|_2^2}{\tau\alpha}\right)\right)-\log 2.$$

This combined with the alignment term results in

$$\begin{aligned}\widetilde{\mathcal{L}}(\mathbf{W}) &= \mathcal{L}_{\text{align}}(\mathbf{W}) + \widetilde{\mathcal{L}}_{\text{unif}}(\mathbf{W}) \\ &= -\frac{\|\mathbf{W}\|_{\text{F}}^2}{\tau\alpha} + \log\left(1 + \exp\left(-\frac{2\|\mathbf{W}\boldsymbol{\mu}\|_2^2}{\tau\alpha}\right)\right) - \log 2.\end{aligned}$$

This loss function turns out to be a simple univariate function. To see this, we define $t = \|\mathbf{W}\|_{\text{F}}^2/\alpha$. As a result, we have $\|\mathbf{W}\boldsymbol{\mu}\|_2^2 = \alpha - (1+\sigma_{\text{aug}}^2)\alpha t$, and hence

$$\begin{aligned}\widetilde{\mathcal{L}}(\mathbf{W}) &= -\frac{t}{\tau} + \log\left(1 + \exp\left(-\frac{2\alpha\left[1-(1+\sigma_{\text{aug}}^2)t\right]}{\tau\alpha}\right)\right) - \log 2 \\ &= -\frac{t}{\tau} + \log\left(1 + \exp\left(\frac{2(1+\sigma_{\text{aug}}^2)t}{\tau} - \frac{2}{\tau}\right)\right) - \log 2.\end{aligned}$$

C.2.3 Proof of Theorem 4.4.3

Recall the approximate loss function

$$\widetilde{\mathcal{L}}(\mathbf{W}) = \ell(t) = -\frac{t}{\tau} + \log\left(\frac{1}{2} + \frac{1}{2}\exp\left(-\frac{2}{\tau} + \frac{2(1+\sigma_{\text{aug}}^2)}{\tau}t\right)\right),$$

where $t = \|\mathbf{W}\|_{\text{F}}^2/\alpha \in [1/(1+\sigma_{\text{aug}}^2+\|\boldsymbol{\mu}\|^2), 1/(1+\sigma_{\text{aug}}^2)]$ with $\alpha = (1+\sigma_{\text{aug}}^2)\|\mathbf{W}\|_{\text{F}}^2 + \|\mathbf{W}\boldsymbol{\mu}\|_2^2$.

Since $F(t;\sigma_{\text{aug}}^2,\tau) \coloneqq \frac{d}{dt}\widetilde{\mathcal{L}}(\mathbf{W})$, we have

$$F(t;\sigma_{\text{aug}}^2,\tau) = \frac{1}{\tau}\left\{1 + 2\sigma_{\text{aug}}^2 - \frac{2(1+\sigma_{\text{aug}}^2)}{1+\exp\left(-\frac{2}{\tau} + \frac{2(1+\sigma_{\text{aug}}^2)}{\tau}t\right)}\right\},$$

which is a strictly increasing function in t. In addition, $F\left((1+\sigma_{\text{aug}}^2)^{-1};\sigma_{\text{aug}}^2,\tau\right) = \sigma_{\text{aug}}^2/\tau > 0$. To determine the minimizer of $\ell(t)$, it suffices to check the sign of $F\left((1+\sigma_{\text{aug}}^2+\|\boldsymbol{\mu}\|)^{-1};\sigma_{\text{aug}}^2,\tau\right)$.

Case 1: expansion regime. If $F\left((1+\sigma_{\text{aug}}^2+\|\boldsymbol{\mu}\|)^{-1};\sigma_{\text{aug}}^2,\tau\right)>0$ (i.e., in the *expansion regime*), we have $F\left(t;\sigma_{\text{aug}}^2,\tau\right)>0$ for any $t\in[1/(1+\sigma_{\text{aug}}^2+\|\boldsymbol{\mu}\|^2),1/(1+\sigma_{\text{aug}}^2)]$. Thus, $\widetilde{\mathcal{L}}(W)$ is strictly increasing in t and is minimized when $t=1/(1+\sigma_{\text{aug}}^2+\|\boldsymbol{\mu}\|^2)$, which implies that

$$\frac{\|\mathbf{W}\|_{\text{F}}^2}{(1+\sigma_{\text{aug}}^2)\|\mathbf{W}\|_{\text{F}}^2+\|\mathbf{W}\boldsymbol{\mu}\|_2^2}=\frac{1}{1+\sigma_{\text{aug}}^2+\|\boldsymbol{\mu}\|^2}\quad\Longrightarrow\quad\frac{\|\mathbf{W}\boldsymbol{\mu}\|_2^2}{\|\mathbf{W}\|_{\text{F}}^2}=\|\boldsymbol{\mu}\|^2. \tag{C.6}$$

If we write the SVD of $\mathbf{W}$ as $\mathbf{W}=\sum_{j=1}^p\sigma_j\boldsymbol{u}_j\boldsymbol{v}_j^\top$, then (C.6) can be written as

$$\sum_{j=1}^p\langle\boldsymbol{v}_j,\boldsymbol{\mu}\rangle^2=\sum_{j=1}^p\frac{\sigma_j^2}{\sum_{k=1}^p\sigma_k^2}\langle\boldsymbol{v}_j,\boldsymbol{\mu}\rangle^2\le\frac{\sigma_1^2}{\sum_{k=1}^p\sigma_k^2}\sum_{j=1}^p\langle\boldsymbol{v}_j,\boldsymbol{\mu}\rangle^2\le\sum_{j=1}^p\langle\boldsymbol{v}_j,\boldsymbol{\mu}\rangle^2,$$

where equality holds if and only if $\sigma_j=0$ for all $j\ge2$ and $\sigma_1>0$. Then,

$$\|\boldsymbol{\mu}\|^2=\sum_{j=1}^p\langle\boldsymbol{v}_j,\boldsymbol{\mu}\rangle^2=\langle\boldsymbol{v}_1,\boldsymbol{\mu}\rangle^2.$$

Case 2: shrinkage regime. If $F\left((1+\sigma_{\text{aug}}^2+\|\boldsymbol{\mu}\|)^{-1};\sigma_{\text{aug}}^2,\tau\right)<0$ (*shrinkage regime*), then $\widetilde{\mathcal{L}}(W)$ is minimized at some t^* in the interior satisfying

$$F\left(t^*;\sigma_{\text{aug}}^2,\tau\right)=\frac{1}{\tau}\left\{1+2\sigma_{\text{aug}}^2-\frac{2(1+\sigma_{\text{aug}}^2)}{1+\exp\left(-\frac{2}{\tau}+\frac{2(1+\sigma_{\text{aug}}^2)}{\tau}t^*\right)}\right\}=0.$$

Solving this provides us with

$$t^*=\frac{1}{1+\sigma_{\text{aug}}^2}\left\{1-\frac{\tau}{2}\log\left(1+2\sigma_{\text{aug}}^2\right)\right\}.$$

When $\sigma_{\text{aug}} \to 0$, we have $t^* \to 1$, which implies that

$$\frac{\|\mathbf{W}\|_{\text{F}}^2}{\|\mathbf{W}\|_{\text{F}}^2 + \|\mathbf{W}\boldsymbol{\mu}\|_2^2} \to 1 \quad \Longrightarrow \quad \frac{\|\mathbf{W}\boldsymbol{\mu}\|_2^2}{\|\mathbf{W}\|_{\text{F}}^2} \to 0.$$

Note that for any fixed value of $\|\mathbf{W}\|_{\text{F}}$, $\|\mathbf{W}\boldsymbol{\mu}\|^2 \to 0$ is equivalent to

$$\sum_{j=1}^{p} \sigma_j^2 \langle \boldsymbol{v}_j, \boldsymbol{\mu} \rangle^2 \to 0$$

from which we have $\max_{j \le p} |\sigma_j \langle \boldsymbol{v}_j, \boldsymbol{\mu} \rangle| \to 0$.

C.2.4 Proof of Proposition 4.4.4

A sandwich formula

Define $\mathcal{L}_\ell(\mathbf{W}) \coloneqq \widetilde{\mathcal{L}}(\mathbf{W})$, and $\mathcal{L}_u(\mathbf{W}) \coloneqq \widetilde{\mathcal{L}}(\mathbf{W}) + \Delta(\mathbf{W})$, where we denote $r = 1 + \sigma_{\text{aug}}^2$ and

$$\Delta(\mathbf{W}) \coloneqq \sum_{j \le p} \frac{r^2 \sigma_j^4}{\tau^2 \alpha^2} + \exp\left(-\frac{\|\mathbf{W}\boldsymbol{\mu}\|^2}{\tau\alpha}\right) \cdot \left[\exp\Big(\sum_{j \le p} \frac{4r\sigma_j^4 \langle \boldsymbol{\mu}, \boldsymbol{v}_j \rangle^2}{\tau^2 \alpha^2}\Big) - 1\right].$$

We have the following sandwich-type result.

Lemma C.2.1. *We have*

$$\mathcal{L}_\ell(\mathbf{W}) \le \mathcal{L}(\mathbf{W}) \le \mathcal{L}_u(\mathbf{W}), \qquad \textit{for all } \mathbf{W} \in \mathbb{R}^{p \times p}.$$

Proof. The lower holds by the construction of $\widetilde{\mathcal{L}}(\mathbf{W})$. Hence we focus on the upper bound.

Since $\log(1+x) - x \ge -x^2/2$, we compare the difference between the first term in $\mathcal{L}_{\text{unif}}$ and that in $\widetilde{\mathcal{L}}_{\text{unif}}$ to obtain

$$-\frac{1}{2}\sum_{j=1}^{p} \log\left(1 + \frac{2r\sigma_j^2}{\tau\alpha}\right) + \sum_{j=1}^{p} \frac{r\sigma_j^2}{\tau\alpha} \le \sum_{j=1}^{p} \frac{r^2\sigma_j^4}{\tau^2\alpha^2}\,.$$

Write $\beta_j = \langle \boldsymbol{\mu}, \boldsymbol{v}_j \rangle^2$ for simplicity. We compare the difference between the second terms.

$$
\begin{aligned}
&\log\left(1+\exp\Big(-\sum_{j=1}^{p}\frac{2\sigma_j^2\beta_j}{2r\sigma_j^2+\tau\alpha}\Big)\right)-\log\left(1+\exp\Big(-\sum_{j=1}^{p}\frac{2\sigma_j^2\beta_j}{\tau\alpha}\Big)\right)\\
&=\log\left[1+\frac{\exp\big(-\sum_{j\le p}2\sigma_j^2\beta_j/(2r\sigma_j^2+\tau\alpha)\big)-\exp\big(-\sum_{j\le p}2\sigma_j^2\beta_j/(\tau\alpha)\big)}{1+\exp\big(-\sum_{j\le p}2\sigma_j^2\beta_j/(\tau\alpha)\big)}\right]\\
&\overset{(i)}{\le}\exp\Big(-\sum_{j\le p}\frac{2\sigma_j^2\beta_j}{2r\sigma_j^2+\tau\alpha}\Big)-\exp\Big(-\sum_{j\le p}\frac{2\sigma_j^2\beta_j}{\tau\alpha}\Big)\\
&=\exp\Big(-\frac{2\|\mathbf{W}\boldsymbol{\mu}\|^2}{\tau\alpha}\Big)\cdot\left[\exp\Big(-\sum_{j\le p}\frac{2\sigma_j^2\beta_j}{2r\sigma_j^2+\tau\alpha}+\sum_{j\le p}\frac{2\sigma_j^2\beta_j}{\tau\alpha}\Big)-1\right]\\
&\overset{(ii)}{\le}\exp\Big(-\frac{2\|\mathbf{W}\boldsymbol{\mu}\|^2}{\tau\alpha}\Big)\cdot\left[\exp\Big(\sum_{j\le p}\frac{4r\sigma_j^4\beta_j}{\tau^2\alpha^2}\Big)-1\right]
\end{aligned}
$$

where in (i) we used $\log(1+x)\le x$ and $\exp\big(-\sum_{j\le p}2\sigma_j^2\beta_j/(\tau\alpha)\big)\ge 0$, and in (ii) we used brute force computation. This completes the proof. □

We denote the minimizers of $\mathcal{L}_\ell$, $\mathcal{L}$, $\mathcal{L}_u$ by $\mathbf{W}_\ell^*$, $\mathbf{W}^*$, $\mathbf{W}_u^*$, respectively. We also denote the related *expansion measure*

$$t_\ell^* = t(\mathbf{W}_\ell^*), \quad t^* = t(\mathbf{W}^*), \quad t_u^* = t(\mathbf{W}_u^*),$$

where we recall

$$t = t(\mathbf{W}) = \frac{\|\mathbf{W}\|_{\mathrm{F}}^2}{r\|\mathbf{W}\|_{\mathrm{F}}^2 + \|\mathbf{W}\boldsymbol{\mu}\|^2}.$$

In what remains, we are mainly interested in showing that t^* and t_ℓ^* are close. This would suggest that our approximation has small effects on the expansion/shrinkage phenomenon, thus justifying our approximation.

Approximation bound in expansion regime

Recall the notation $r = 1 + \sigma_{\text{aug}}^2$, and define $\rho = \|\boldsymbol{\mu}\|$. We observe from Lemma C.2.1 that

$$\begin{aligned}\mathcal{L}_u(\mathbf{W}_\ell^*) - \mathcal{L}_\ell(\mathbf{W}_\ell^*) &\geq \mathcal{L}(\mathbf{W}_\ell^*) - \mathcal{L}_\ell(\mathbf{W}_\ell^*) \geq \mathcal{L}(\mathbf{W}^*) - \mathcal{L}_\ell(\mathbf{W}_\ell^*) \\ &\geq \mathcal{L}_\ell(\mathbf{W}^*) - \mathcal{L}_\ell(\mathbf{W}_\ell^*) \\ &\geq \min_{t \in [(r+\rho^2)^{-1}, r^{-1}]} F(t) \cdot \big(t^* - t_\ell^*\big).\end{aligned}$$

Here the last inequality uses the definition of $F(\cdot)$ as well as the fact that $\mathcal{L}_\ell$ is convex in t.

Note that in the expansion regime, t_ℓ^* achieves the smallest value, namely $t_\ell^* = 1/(r+\rho^2)$. Therefore we must have $0 \leq t^* - t_\ell^*$. In addition, we know that in the expansion regime, $F(t) > 0$ for all $t > t_\ell^*$. As a result, we obtain

$$0 \leq t^* - t_\ell^* \leq \left[\min_{t \in [(r+\rho^2)^{-1}, r^{-1}]} F(t)\right]^{-1} \big(\mathcal{L}_u(\mathbf{W}_\ell^*) - \mathcal{L}_\ell(\mathbf{W}_\ell^*)\big) \ . \tag{C.7}$$

Since $F(t)$ is increasing in t, we have

$$\min_{t \in [(r+\rho^2)^{-1}, r^{-1}]} F(t) = F(1/(r+\rho^2)) = -\frac{1}{\tau} + \frac{2r}{\tau} \cdot \left[1 + \exp\Big(\frac{2}{\tau} - \frac{2r}{\tau(r+\rho^2)}\Big)\right]^{-1} .$$

Denoting the following quantity

$$\widetilde{\theta}_{\rho,r,\tau} = \frac{1}{1+e^{2/\tau}} + \frac{2re^{2/\tau}}{(1+e^{2/\tau})^2\rho^2\tau},$$

then as $\rho \to \infty$,

$$\left| \left[1 + \exp\Big(\frac{2}{\tau} - \frac{2r}{\tau(r+\rho^2)}\Big)\right]^{-1} - \widetilde{\theta}_{\rho,r,\tau} \right| = O(\rho^{-4}).$$

Therefore,

$$\min_{t\in[(r+\rho^2)^{-1},r^{-1}]} F(t) = -\frac{1}{\tau} + \frac{2r}{(1+e^{2/\tau})\tau} + \frac{4r^2e^{2/\tau}}{(1+e^{2/\tau})^2\rho^2\tau^2} + O(\rho^{-4})$$
$$= \frac{1}{\tau}\Big(\frac{2r}{1+e^{2/\tau}} - 1\Big) + O(\rho^{-2}). \tag{C.8}$$

We view τ and σ_{aug}^2 as fixed parameters and let $\rho \to \infty$. Recall that the expansion regime occurs if $F((r+\rho^2)^{-1}) > 0$ and the shrinkage regime occurs if the reverse inequality is true. Under the asymptotics $\rho \to \infty$, the phase transition boundary simplifies to $2r > 1 + e^{2/\tau}$ for expansion regime and $2r < 1 + e^{2/\tau}$ for the shrinkage regime. Thus, the leading term C.8 is always positive in this asymptotics.

Since $\mathbf{W} = \mathbf{W}_\ell^*$ if and only if $\langle \boldsymbol{\mu}, \boldsymbol{v}_1\rangle^2 > 0$, $\langle \boldsymbol{\mu}, \boldsymbol{v}_j\rangle^2 = 0$ for $j > 1$ and $\sigma_2 = \ldots = \sigma_p = 0$, we can show that

$$\Delta(\mathbf{W}_\ell^*) = \mathcal{L}_u(\mathbf{W}_\ell^*) - \mathcal{L}_\ell(\mathbf{W}_\ell^*) = \frac{r^2}{\tau^2(r+\rho^2)^2} + \exp\left(-\frac{\rho^2}{\tau(r+\rho^2)}\right) \cdot \left[\exp\left(\frac{4r\rho^2}{\tau^2(r+\rho^2)^2}\right) - 1\right],$$

and as $\rho \to \infty$,

$$\left|\Delta(\mathbf{W}_\ell^*) - \frac{4r}{\tau^2 e^{1/\tau}}\frac{1}{\rho^2}\right| = O(\rho^{-4}). \tag{C.9}$$

From (C.8) and (C.9), we have the control for (C.7):

$$0 \le t^* - t_\ell^* \le \left(\frac{2r}{1+e^{2/\tau}} - 1\right)^{-1} \frac{4r}{\tau e^{1/\tau}}\frac{1}{\rho^2} + O(\rho^{-4})$$

This finishes the proof.

Approximation bound in shrinkage regime

We use the notation $\mathtt{C}^2$ to denote the space of twice continuously differentiable functions. We first present the general result as follows.

Lemma C.2.2. *Suppose that $f(\boldsymbol{x}), f_\ell(\boldsymbol{x}), f_u(\boldsymbol{x})$ are defined for $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^q$ with*

$$f_\ell(\boldsymbol{x}) \leq f(\boldsymbol{x}) \leq f_u(\boldsymbol{x}).$$

Assume that $f_\ell(\boldsymbol{x}) = h(\varphi(\boldsymbol{x}))$ where $\varphi : \mathcal{X} \to \mathcal{Y} \subset \mathbb{R}$ is a surjection satisfying $\varphi \in \mathtt{C}^2(\mathcal{X})$, and that $h : \mathcal{Y} \to \mathbb{R}$ satisfies $h \in \mathtt{C}^2(\mathcal{X})$ and is strongly convex. Suppose $\boldsymbol{x}_\ell^* \in \mathcal{X}$ and y_ℓ^* satisfy

$$\varphi(\boldsymbol{x}_\ell^*) = y_\ell^*, \qquad y_\ell^* = \operatorname*{argmin}_{y \in \mathcal{Y}} h(y).$$

Then, the following inequality holds.

$$\left\| y_\ell^* - \varphi(\boldsymbol{x}^*) \right\|^2 \leq \frac{2\big(f_u(\boldsymbol{x}_\ell^*) - f_\ell(\boldsymbol{x}_\ell^*)\big)}{\inf_{y \in \mathcal{Y}} h''(y)} \,. \tag{C.10}$$

Proof. By the definitions of f_u, f, f_ℓ and $\boldsymbol{x}_\ell^*, \boldsymbol{x}^*$, we have

$$f_u(\boldsymbol{x}_\ell^*) \geq f(\boldsymbol{x}_\ell^*) \geq f(\boldsymbol{x}^*) \geq f_\ell(\boldsymbol{x}^*) = h(\varphi(\boldsymbol{x}^*)) \,. \tag{C.11}$$

Since h is convex and $\varphi(\boldsymbol{x}_\ell^*)$ achieves the minimum, we have

$$h(\varphi(\boldsymbol{x}^*)) - h(\varphi(\boldsymbol{x}_\ell^*)) \geq \frac{1}{2} \inf_{y \in \mathcal{Y}} h''(y) \cdot \left\| \varphi(\boldsymbol{x}^*) - \varphi(\boldsymbol{x}_\ell^*) \right\|^2. \tag{C.12}$$

We combine (C.11)–(C.12) and use $h(\varphi(\boldsymbol{x}_\ell^*)) = f_\ell(\boldsymbol{x}_\ell^*)$, $\varphi(\boldsymbol{x}_\ell^*) = y_\ell^*$, which leads to the desired inequality (C.10). □

To use the above lemma, we set $f = \mathcal{L}, f_\ell = \mathcal{L}_\ell, f_u = \mathcal{L}_u$,

$$h(t) = -\frac{t}{\tau} + \log\left(1 + \exp\Big(-\frac{2}{\tau} + \frac{2rt}{\tau}\Big)\right) - \log 2,$$
$$\varphi(\mathbf{W}) = t(\mathbf{W}) = \frac{\|\mathbf{W}\|_{\mathrm{F}}^2}{r\|\mathbf{W}\|_{\mathrm{F}}^2 + \|\mathbf{W}\boldsymbol{\mu}\|^2},$$

where we identify y with t. We calculate the lower bound on $h''(t)$ as follows.

$$h'(t) = F(t) = -\frac{1}{\tau} + \frac{2r}{\tau}\left[1 + \exp\Big(\frac{2}{\tau} - \frac{2rt}{\tau}\Big)\right]^{-1},$$
$$h''(t) = F'(t) = \frac{4r^2}{\tau^2}\exp\Big(\frac{2}{\tau} - \frac{2rt}{\tau}\Big) \cdot \left[1 + \exp\Big(\frac{2}{\tau} - \frac{2rt}{\tau}\Big)\right]^{-2}.$$

We note that

$$\exp\left(\frac{2}{\tau}\right) \geq \exp\Big(\frac{2}{\tau} - \frac{2rt}{\tau}\Big) \geq \exp\left(\frac{2-2r}{\tau}\right).$$

So making use of the fact that $\frac{x}{(1+x)^2}$ is decreasing when $x \geq 1$, we deduce

$$\begin{aligned}\min_{t\in[(r+\rho)^{-1},1]} F'(t) &= \frac{4r^2}{\tau^2}\min\left\{\frac{e^{2\sigma_{\mathrm{aug}}^2/\tau r}}{(1+e^{2\sigma_{\mathrm{aug}}^2/\tau r})^2}, \frac{e^{2\rho/\tau(r+\rho)}}{(1+e^{2\rho/\tau(r+\rho)})^2}\right\} \\ &= \frac{4r^2}{\tau^2}\frac{e^{2\sigma_{\mathrm{aug}}^2/\tau r}}{(1+e^{2\sigma_{\mathrm{aug}}^2/\tau r})^2} \geq \frac{r^2}{\tau^2}e^{-2\sigma_{\mathrm{aug}}^2/\tau r}.\end{aligned}$$

Note that we have the freedom to choose any $\boldsymbol{x}$ (identified as $\mathbf{W}$) as long as $\varphi(\mathbf{W}) = t_\ell^*$. This leads to the error bound

$$\|t^* - t_\ell^*\|^2 \leq \frac{\tau^2}{r^2}e^{2\sigma_{\mathrm{aug}}^2/\tau} \cdot \Delta(\mathbf{W}).$$

It remains to control $\min_{\mathbf{W}:\varphi(\mathbf{W})=t_\ell^*}\Delta(\mathbf{W})$ which will be our target below.

Characterizing $\min_{\mathbf{W}:\varphi(\mathbf{W})=t_\ell^*}\Delta(\mathbf{W})$. It suffices to provide a feasible solution $\mathbf{W}$ such that $\varphi(\mathbf{W}) = t_\ell^*$. To do so, we construct $\mathbf{W}$ as follows. As usual, let $\sum_{j=1}^p \sigma_j \boldsymbol{u}_j\boldsymbol{v}_j^\top$ be

the singular value decomposition of $\mathbf{W}$. Here σ_j, $j \in [p]$, are not necessarily ordered. Let $\boldsymbol{v}_p = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$. As a result, we have $\|\mathbf{W}\|_{\mathrm{F}} = \sum_j \sigma_j^2$ and $\|\mathbf{W}\boldsymbol{\mu}\| = \sigma_p^2\|\boldsymbol{\mu}\|^2 = \sigma_p^2\rho^2$. Recall that in the shrinkage regime, we have

$$t_\ell^* = \frac{1}{r}\left(1 - \frac{\tau}{2}\log(2r-1)\right) = \frac{\|\mathbf{W}\|_{\mathrm{F}}^2}{r\|\mathbf{W}\|_{\mathrm{F}}^2 + \|\mathbf{W}\boldsymbol{\mu}\|^2} \in \left[\frac{1}{r+\rho^2}, \frac{1}{r}\right].$$

Consequently, one can verify that any sequence of singular values $\boldsymbol{\sigma}$ with $S_{p-1} = \sum_{j=1}^{p-1}\sigma_j^2$ that obeys

$$\frac{S_{p-1}}{\sigma_p^2} = \frac{t_\ell^*(\rho^2+r)-1}{1-t_\ell^* r} > 0 \tag{C.13}$$

will be a feasible solution. In particular,

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_{p-1}^2 = \frac{S_{p-1}}{p-1}.$$

Now we are ready to compute the quantity of interest

$$\Delta(\mathbf{W}) = \sum_{j\le p}\frac{r^2\sigma_j^4}{\tau^2\alpha^2} + \exp\left(-\frac{\|\mathbf{W}\boldsymbol{\mu}\|^2}{\tau\alpha}\right)\cdot\left[\exp\left(\sum_{j\le p}\frac{4r\sigma_j^4\langle\boldsymbol{\mu},\boldsymbol{v}_j\rangle^2}{\tau^2\alpha^2}\right) - 1\right].$$

1. First, we have

$$\begin{aligned}
\sum_{j\le p}\frac{r^2\sigma_j^4}{\tau^2\alpha^2} &= \sum_{j\le p}\frac{r^2\sigma_j^4}{\tau^2(r\|\boldsymbol{\sigma}\|^2+\sigma_p^2\rho^2)^2} = \frac{r^2(t_\ell^*)^2}{\tau^2}\cdot\frac{\sum_{j\le p}\sigma_j^4}{\|\boldsymbol{\sigma}\|^4} \\
&= \frac{r^2(t_\ell^*)^2}{\tau^2}\cdot\frac{(p-1)\frac{S_{p-1}^2}{(p-1)^2}+\sigma_p^4}{\left((p-1)\frac{S_{p-1}}{p-1}+\sigma_p^2\right)^2} \\
&= \frac{r^2(t_\ell^*)^2}{\tau^2(p-1)}\cdot\frac{\frac{S_{p-1}^2}{\sigma_p^2}+p-1}{\left(\frac{S_{p-1}}{\sigma_p^2}+1\right)^2}
\end{aligned}$$

Recall the identity (C.13) and the fact that $rt_\ell^* \le 1$, we can further write the first term as

$$\begin{aligned}
\sum_{j\le p} \frac{r^2\sigma_j^4}{\tau^2\alpha^2} &= \frac{r^2(t_\ell^*)^2}{\tau^2(p-1)} \frac{(\rho^2 t_\ell^* - 1 + rt_\ell^*)^2 + (p-1)(1-rt_\ell^*)^2}{\rho^4 (t_\ell^*)^2} \\
&\le \frac{r^2}{\tau^2\rho^4(p-1)} \cdot \left(\rho^4 (t_\ell^*)^2 + p(1-rt_\ell^*)^2\right) \\
&\le \frac{r^2}{\tau^2\rho^4(p-1)} \cdot \left(\frac{\rho^4}{r^2} + \frac{\tau^2 p}{4}\log^2(2r-1)\right) \\
&= \frac{1}{\tau^2(p-1)} + \frac{pr^2}{4\rho^4(p-1)}\log^2(2r-1) \\
&= \frac{1}{\tau^2(p-1)} + \frac{\sigma_{\text{aug}}^4}{\rho^4} + O\left(\frac{\sigma_{\text{aug}}^6}{\rho^4}\right) \qquad \sigma_{\text{aug}} \to 0.
\end{aligned}$$

2. Secondly, since $\exp\left(-\frac{\|\mathbf{W}\boldsymbol{\mu}\|^2}{\tau\alpha}\right) \le 1$ and $\sigma_p^2\rho^2/\alpha = 1 - rt_\ell^*$, then

$$\begin{aligned}
\exp\left(\sum_{j\le p} \frac{4r\sigma_j^4 \langle \boldsymbol{\mu}, \boldsymbol{v}_j\rangle^2}{\tau^2\alpha^2}\right) - 1 &= \exp\left(\frac{4r\sigma_p^4\rho^2}{\tau^2\alpha^2}\right) - 1 = \exp\left(\frac{4r}{\tau^2\rho^2}\left(1-rt_\ell^*\right)^2\right) - 1 \\
&= \exp\left(\frac{r}{\rho^2}\log^2(2r-1)\right) - 1 \\
&= \frac{4\sigma_{\text{aug}}^4}{\rho^2} + O\left(\frac{\sigma_{\text{aug}}^6}{\rho^2}\right) \qquad \sigma_{\text{aug}} \to 0.
\end{aligned}$$

Combining pieces above,

$$\Delta(\mathbf{W}) \le \frac{1}{\tau^2(p-1)} + O\left(\frac{\sigma_{\text{aug}}^4}{\|\boldsymbol{\mu}\|^2}\right).$$

As a result, we can bound the approximation error in the shrinkage regime by

$$\|t^* - t_\ell^*\|^2 \le \frac{1}{(1+\sigma_{\text{aug}}^2)^2(p-1)} e^{2\sigma_{\text{aug}}^2/\tau} + O\left(\frac{\sigma_{\text{aug}}^4}{\|\boldsymbol{\mu}\|^2}\right).$$

C.2.5 Proof of Proposition 4.4.5

Here we provide some informal analysis on the finite-sample scenario. We realize that a rigorous analysis is challenging and is left to future research.

Recall the contrastive loss

$$\begin{aligned}\mathcal{L}_n(\mathbf{W}) \coloneqq & -\frac{1}{\tau}\mathbb{E}_n\left\{\mathbb{E}_{\boldsymbol{h}^+,\boldsymbol{h}|\boldsymbol{h}_0}\left[\mathrm{sim}^*\left(\mathbf{W}\boldsymbol{h},\mathbf{W}\boldsymbol{h}^+\right)\mid \boldsymbol{h}_0\right]\right\} \\ & + \log\left(\mathbb{E}_n\left\{\mathbb{E}_{\boldsymbol{h}^-,\boldsymbol{h}|\boldsymbol{h}_0}\left[e^{\mathrm{sim}^*(\mathbf{W}\boldsymbol{h},\mathbf{W}\boldsymbol{h}^-)/\tau}\big|\boldsymbol{h}_0\right]\right\}\right),\end{aligned}$$

which can be decomposed as two terms

$$\begin{aligned}\mathcal{L}_{n,1}(\mathbf{W}) &= -\frac{1}{\tau}\mathbb{E}_n\left[\mathrm{sim}^*\left(\mathbf{W}\boldsymbol{h},\mathbf{W}\boldsymbol{h}^+\right)\right], \\ \mathcal{L}_{n,2}(\mathbf{W}) &= \log\left(\mathbb{E}_n\left\{\mathbb{E}_{\boldsymbol{h}^-,\boldsymbol{h}|\boldsymbol{h}_0}\left[e^{\mathrm{sim}^*(\mathbf{W}\boldsymbol{h},\mathbf{W}\boldsymbol{h}^-)/\tau}\big|\boldsymbol{h}_0\right]\right\}\right).\end{aligned}$$

Denote $\alpha = (1+\sigma_{\mathrm{aug}}^2)\|\mathbf{W}\|_{\mathrm{F}}^2 + \|\mathbf{W}\boldsymbol{\mu}\|_2^2$ and $\mathbf{A} = (2\tau\alpha)^{-1}\mathbf{W}^\top\mathbf{W}$, with the definition of sim^*, we have

$$\mathcal{L}_{n,1}(\mathbf{W}) = \frac{1}{2\tau\alpha}\mathbb{E}_n\left\{\mathbb{E}_{\boldsymbol{h}^+,\boldsymbol{h}|\boldsymbol{h}_0}\left[\|\mathbf{W}(\boldsymbol{h}-\boldsymbol{h}^+)\|_2^2\big|\boldsymbol{h}_0\right]\right\} = \frac{\sigma_{\mathrm{aug}}^2}{\tau\alpha}\|\mathbf{W}\|_{\mathrm{F}}^2.$$

For $\boldsymbol{h}^- \sim \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu},(1+\sigma_{\mathrm{aug}}^2)\mathbf{I}) + \frac{1}{2}\mathcal{N}(\boldsymbol{\mu},(1+\sigma_{\mathrm{aug}}^2)\mathbf{I})$ and $\boldsymbol{h}\mid\boldsymbol{h}_0 \sim \mathcal{N}(\boldsymbol{h}_0,\sigma_{\mathrm{aug}}^2\mathbf{I})$, we have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{h},\boldsymbol{h}^-|\boldsymbol{h}_0}\left[\exp\left(\frac{\mathrm{sim}^*(\mathbf{W}\boldsymbol{h},\mathbf{W}\boldsymbol{h}^-)}{\tau}\right)\middle|\boldsymbol{h}_0\right] = \mathbb{E}_{\boldsymbol{h},\boldsymbol{h}^-|\boldsymbol{h}_0}\left[\exp\left(-\frac{\|\mathbf{W}(\boldsymbol{h}-\boldsymbol{h}^-)\|_2^2}{2\tau\alpha}\right)\middle|\boldsymbol{h}_0\right] \\ &= \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{\mu},(1+2\sigma_{\mathrm{aug}}^2)\mathbf{I})}\left[\exp\left(-\frac{\|\mathbf{W}(\boldsymbol{h}_0-\boldsymbol{z})\|_2^2}{2\tau\alpha}\right)\right] \\ &\quad + \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(-\boldsymbol{\mu},(1+2\sigma_{\mathrm{aug}}^2)\mathbf{I})}\left[\exp\left(-\frac{\|\mathbf{W}(\boldsymbol{h}_0-\boldsymbol{z})\|_2^2}{2\tau\alpha}\right)\right].\end{aligned} \tag{C.14}$$

With $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, (1+2\sigma_{\text{aug}}^2)\mathbf{I})$, we have

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{z}|\boldsymbol{h}_0}\left[\exp\left(-\frac{\|\mathbf{W}(\boldsymbol{h}_0-\boldsymbol{z})\|_2^2}{2\tau\alpha}\right)\middle|\boldsymbol{h}_0\right] \\
=&\left(\frac{1}{2\pi(1+2\sigma_{\text{aug}}^2)}\right)^{d/2}\int \exp\left(-\frac{\|\mathbf{W}(\boldsymbol{h}_0-\boldsymbol{z})\|_2^2}{2\tau\alpha}\right)\exp\left(-\frac{1}{2(1+2\sigma_{\text{aug}}^2)}\|\boldsymbol{z}-\boldsymbol{\mu}\|_2^2\right)d\boldsymbol{z} \\
=&\left(\frac{1}{2\pi(1+2\sigma_{\text{aug}}^2)}\right)^{d/2}\int \exp\left(-\frac{1}{2(1+2\sigma_{\text{aug}}^2)}\left[(\boldsymbol{z}-\boldsymbol{\mu})^\top(\boldsymbol{z}-\boldsymbol{\mu})+(\boldsymbol{h}_0-\boldsymbol{z})^\top\widetilde{\mathbf{A}}(\boldsymbol{h}_0-\boldsymbol{z})\right]\right)d\boldsymbol{z},
\end{aligned}
\tag{C.15}
$$

where $\widetilde{\mathbf{A}} = 2(1+2\sigma_{\text{aug}}^2)\mathbf{A}$. Rearrange (C.15), we have

$$
\begin{aligned}
&(\boldsymbol{z}-\boldsymbol{\mu})^\top(\boldsymbol{z}-\boldsymbol{\mu})+(\boldsymbol{h}_0-\boldsymbol{z})^\top\widetilde{\mathbf{A}}(\boldsymbol{h}_0-\boldsymbol{z}) \\
=&(\boldsymbol{z}-\widetilde{\boldsymbol{a}})^\top\mathbf{M}^{-1}(\boldsymbol{z}-\widetilde{\boldsymbol{a}})-(\boldsymbol{\mu}+\widetilde{\mathbf{A}}\boldsymbol{h}_0)^\top\mathbf{M}(\boldsymbol{\mu}+\widetilde{\mathbf{A}}\boldsymbol{h}_0)+\boldsymbol{\mu}^\top\boldsymbol{\mu}+\boldsymbol{h}_0^\top\widetilde{\mathbf{A}}\boldsymbol{h}_0.
\end{aligned}
$$

where $\mathbf{M} = (\mathbf{I}+\widetilde{\mathbf{A}})^{-1}$ and $\widetilde{\boldsymbol{a}} = \mathbf{M}(\boldsymbol{\mu}+\widetilde{\mathbf{A}}\boldsymbol{h}_0)$. Then, the integral in (C.15) can be simplified as

$$
\text{(C.15)} = \sqrt{\texttt{det}(\mathbf{M})}\exp\left(\frac{1}{2(1+2\sigma_{\text{aug}}^2)}\left[(\boldsymbol{\mu}+\widetilde{\mathbf{A}}\boldsymbol{h}_0)^\top\mathbf{M}(\boldsymbol{\mu}+\widetilde{\mathbf{A}}\boldsymbol{h}_0)-\boldsymbol{\mu}^\top\boldsymbol{\mu}-\boldsymbol{h}_0^\top\widetilde{\mathbf{A}}\boldsymbol{h}_0\right]\right).
$$

Note that $\mathbf{M}\widetilde{\mathbf{A}} = \mathbf{I} - \mathbf{M} = \widetilde{\mathbf{A}}\mathbf{M}$, we can rewrite (C.14) as

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{h}^-,\boldsymbol{h}|\boldsymbol{h}_0}\left[\exp\left(\frac{\mathrm{sim}^*(\mathbf{W}\boldsymbol{h},\mathbf{W}\boldsymbol{h}^-)}{\tau}\right)\middle|\boldsymbol{h}_0\right] \\
=&\frac{1}{2}\sqrt{\mathtt{det}(\mathbf{M})}\exp\left(\frac{1}{2(1+2\sigma_{\mathrm{aug}}^2)}\left[(\boldsymbol{\mu}+\widetilde{\mathbf{A}}\boldsymbol{h}_0)^\top\mathbf{M}(\boldsymbol{\mu}+\widetilde{\mathbf{A}}\boldsymbol{h}_0)-\boldsymbol{\mu}^\top\boldsymbol{\mu}-\boldsymbol{h}_0^\top\widetilde{\mathbf{A}}\boldsymbol{h}_0\right]\right) \\
&\quad+\frac{1}{2}\sqrt{\mathtt{det}(\mathbf{M})}\exp\left(\frac{1}{2(1+2\sigma_{\mathrm{aug}}^2)}\left[(-\boldsymbol{\mu}+\widetilde{\mathbf{A}}\boldsymbol{h}_0)^\top\mathbf{M}(-\boldsymbol{\mu}+\widetilde{\mathbf{A}}\boldsymbol{h}_0)-\boldsymbol{\mu}^\top\boldsymbol{\mu}-\boldsymbol{h}_0^\top\widetilde{\mathbf{A}}\boldsymbol{h}_0\right]\right) \\
=&\frac{1}{2}\sqrt{\mathtt{det}(\mathbf{M})}\exp\left(\frac{1}{2(1+2\sigma_{\mathrm{aug}}^2)}\left[\boldsymbol{\mu}^\top(\mathbf{M}-\mathbf{I})\boldsymbol{\mu}+\boldsymbol{h}_0^\top(\mathbf{M}-\mathbf{I})\boldsymbol{h}_0-2\boldsymbol{\mu}^\top(\mathbf{M}-\mathbf{I})\boldsymbol{h}_0\right]\right) \\
&\quad\cdot\left(1+\exp\left(\frac{1}{2(1+2\sigma_{\mathrm{aug}}^2)}\left[4\boldsymbol{\mu}^\top(\mathbf{M}-\mathbf{I})\boldsymbol{h}_0\right]\right)\right).
\end{aligned}
$$

Further, as $\mathbf{I}-\mathbf{M}\succcurlyeq 0$,

$$
\begin{aligned}
&\mathbb{E}_n\left\{\mathbb{E}_{\boldsymbol{h}^-,\boldsymbol{h}|\boldsymbol{h}_0}\left[\exp\left(\frac{\mathrm{sim}^*(\mathbf{W}\boldsymbol{h},\mathbf{W}\boldsymbol{h}^-)}{\tau}\right)\middle|\boldsymbol{h}_0\right]\right\} \\
=&\frac{1}{2}\sqrt{\mathtt{det}(\mathbf{M})}\mathbb{E}_n\Bigg\{\exp\left(-\frac{1}{2(1+2\sigma_{\mathrm{aug}}^2)}(\boldsymbol{h}_0-\boldsymbol{\mu})^\top(\mathbf{I}-\mathbf{M})(\boldsymbol{h}_0-\boldsymbol{\mu})\right) \\
&\qquad+\exp\left(-\frac{1}{2(1+2\sigma_{\mathrm{aug}}^2)}(\boldsymbol{h}_0+\boldsymbol{\mu})^\top(\mathbf{I}-\mathbf{M})(\boldsymbol{h}_0+\boldsymbol{\mu})\right)\Bigg\}.
\end{aligned}
$$

Taking the logarithm, the uniformity loss can be simplified as

$$
\begin{aligned}
\mathcal{L}_{n,2}(\mathbf{W}) &= \log\left(\mathbb{E}_n\mathbb{E}_{\boldsymbol{h}^-,\boldsymbol{h}|\boldsymbol{h}_0}\left[\exp\left(\frac{\mathrm{sim}^*(\mathbf{W}\boldsymbol{h},\mathbf{W}\boldsymbol{h}^-)}{\tau}\right)\right]\right) \\
&= -\log 2+\frac{1}{2}\log\det[\mathbf{M}]+\log\left(\mathbb{E}_n S_{\boldsymbol{h}_0}\right),
\end{aligned}
$$

where

$$S_{\boldsymbol{h}_0} \coloneqq \exp\left(-\frac{1}{2(1+2\sigma_{\text{aug}}^2)}(\boldsymbol{h}_0-\boldsymbol{\mu})^\top(\mathbf{I}-\mathbf{M})(\boldsymbol{h}_0-\boldsymbol{\mu})\right) + \exp\left(-\frac{1}{2(1+2\sigma_{\text{aug}}^2)}(\boldsymbol{h}_0+\boldsymbol{\mu})^\top(\mathbf{I}-\mathbf{M})(\boldsymbol{h}_0+\boldsymbol{\mu})\right).$$

Organizing the terms above, we have

$$\mathcal{L}_n(\mathbf{W}) = -\log 2 + \frac{\sigma_{\text{aug}}^2}{\tau\alpha}\|\mathbf{W}\|_{\text{F}}^2 + \frac{1}{2}\log\det[\mathbf{M}] + \log\left(\mathbb{E}_n S_{\boldsymbol{h}_0}\right).$$

When $\alpha = (1+\sigma_{\text{aug}}^2)\|\mathbf{W}\|_{\text{F}}^2 + \|\mathbf{W}\boldsymbol{\mu}\|_2^2 \gg \sigma_j^2$ for any singular value σ_j of $\mathbf{W}$, we have $\mathbf{M} \approx \mathbf{I} - (\tau\alpha)^{-1}(1+2\sigma_{\text{aug}}^2)\mathbf{W}^\top\mathbf{W}$ and

$$\log\Big(\det[\mathbf{M}]\Big) = \log\left(\prod_{j=1}^{p}\left(1+\frac{1+2\sigma_{\text{aug}}^2}{\tau\alpha}\sigma_j^2\right)^{-1}\right) \approx -\frac{1+2\sigma_{\text{aug}}^2}{\tau\alpha}\|\mathbf{W}\|_{\text{F}}^2.$$

We can then approximate S_n by

$$\widetilde{S}_{\boldsymbol{h}_0} \coloneqq \exp\left(-\frac{1}{2\tau\alpha}\|\mathbf{W}(\boldsymbol{h}_0-\boldsymbol{\mu})\|^2\right) + \exp\left(-\frac{1}{2\tau\alpha}\|\mathbf{W}(\boldsymbol{h}_0+\boldsymbol{\mu})\|^2\right).$$

As a result, we have the approximation

$$\mathcal{L}_{n,2}(\mathbf{W}) \approx -\log 2 - \frac{1+2\sigma_{\text{aug}}^2}{2\tau\alpha}\|\mathbf{W}\|_{\text{F}}^2 + \log\left(\mathbb{E}_n\widetilde{S}_{\boldsymbol{h}_0}\right) =: \widetilde{\mathcal{L}}_{n,2}(\mathbf{W}).$$

Therefore, recall that $\alpha = (1+\sigma_{\text{aug}}^2)\|\mathbf{W}\|_{\text{F}}^2 + \|\mathbf{W}\boldsymbol{\mu}\|^2$, we have the following approximation for $\mathcal{L}_n(\mathbf{W})$:

$$\begin{aligned}\widetilde{\mathcal{L}}_n(\mathbf{W}) &= \mathcal{L}_{n,1}(\mathbf{W}) + \widetilde{\mathcal{L}}_{n,2}(\mathbf{W}) \\ &= -\log 2 - \frac{1}{2\tau\alpha}\|\mathbf{W}\|_{\text{F}}^2 + \log\left(\mathbb{E}_n\widetilde{S}_{\boldsymbol{h}_0}\right).\end{aligned}$$

As $\boldsymbol{h}_0 \sim \mathcal{N}(y \cdot \boldsymbol{\mu}, \mathbf{I})$, then

$$\log\left(\mathbb{E}_{\boldsymbol{h}_0}\widetilde{S}_{\boldsymbol{h}_0}\right) = -\frac{1}{2}\log\left(1+\frac{\sigma_j^2}{\tau\alpha}\right) + \log\left(1+\exp\left(-\frac{2\sigma_j^2}{\sigma_j^2+\tau\alpha}(\boldsymbol{v}_j^\top\boldsymbol{\mu})\right)\right).$$

Under the assumption that $\alpha \gg \|\mathbf{W}\|_{\mathrm{F}}^2$, we further have

$$\log\left(\mathbb{E}_{\boldsymbol{h}_0}\widetilde{S}_{\boldsymbol{h}_0}\right) \approx -\frac{1}{2\tau\alpha}\|\mathbf{W}\|_{\mathrm{F}}^2 + \log\left(1+\exp\left(-\frac{2}{\tau\alpha}\|\mathbf{W}\boldsymbol{\mu}\|^2\right)\right).$$

Accordingly, if we replace $\mathbb{E}_n$ by $\mathbb{E}_{\boldsymbol{h}_0}$ in $\widetilde{\mathcal{L}}_n(\mathbf{W})$, then the loss can be approximated by

$$-\log 2 - \frac{1}{\tau\alpha}\|\mathbf{W}\|_{\mathrm{F}}^2 + \log\left(1+\exp\left(-\frac{2}{\tau\alpha}\|\mathbf{W}\boldsymbol{\mu}\|^2\right)\right),$$

which is exactly the loss function $\widetilde{\mathcal{L}}(\mathbf{W})$ defined in (4.3).

C.2.6 An example of expansion/shrinkage transition

Consider the setting where the distribution of features depends on temperature τ:

$$\boldsymbol{h}_{0,i} \mid y_i \sim y_i\sqrt{g(\tau)}\cdot\boldsymbol{\mu} + \mathcal{N}\left(\mathbf{0}, \mathbf{I} + \phi(\tau)\cdot\boldsymbol{\mu}\boldsymbol{\mu}^\top\right).$$

We obtain augmented pairs via

$$\boldsymbol{h}_i,\ \boldsymbol{h}_i^+ \mid \boldsymbol{h}_{0,i} \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{h}_{0,i}, \sigma_{\mathrm{aug}}^2\mathbf{I}).$$

Redefine $\alpha = \mathbb{E}\|\mathbf{W}\boldsymbol{h}^+\|^2 = (1+\sigma_{\mathrm{aug}}^2)\|\mathbf{W}\|_{\mathrm{F}} + (g(\tau)+\phi(\tau))\|\mathbf{W}\boldsymbol{\mu}\|^2$. Similar with the calculation in Theorem 4.4.3, in order to minimize $\widetilde{\mathcal{L}}(\mathbf{W})$, it suffices to minimize

$$\ell(t) = \frac{t}{\tau(\phi(\tau)+g(\tau))}\left\{\sigma_{\mathrm{aug}}^2\phi(\tau) - g(\tau)\right\} + \log\left\{1+\exp\left(-\frac{2g(\tau)}{\phi(\tau)+g(\tau)}\frac{1-(1+\sigma_{\mathrm{aug}}^2)t}{\tau}\right)\right\},$$

where $t = \|\mathbf{W}\|_{\mathrm{F}}/\alpha$. We further define

$$F(t) = \frac{d}{dt}\ell(t) = \frac{\sigma_{\mathrm{aug}}^2\phi(\tau) - g(\tau)}{\tau(\phi(\tau)+g(\tau))} + \frac{2(1+\sigma_{\mathrm{aug}}^2)g(\tau)}{\phi(\tau)+g(\tau)}\left\{1+\exp\left(\frac{2g(\tau)}{\phi(\tau)+g(\tau)}\frac{1-(1+\sigma_{\mathrm{aug}}^2)t}{\tau}\right)\right\}.$$

Recall that the solution lies in the expansion regime if

$$F\left(\frac{1}{1+\sigma_{\mathrm{aug}}^2+(\phi(\tau)+g(\tau))\|\boldsymbol{\mu}\|^2}\right) \geq 0,$$

which is equivalent to

$$\frac{2}{\tau}\frac{g(\tau)\|\boldsymbol{\mu}\|^2}{1+\sigma_{\mathrm{aug}}^2+(\phi(\tau)+g(\tau))\|\boldsymbol{\mu}\|^2} \leq \log\left(\frac{(1+2\sigma_{\mathrm{aug}}^2)g(\tau)+\sigma_{\mathrm{aug}}^2\phi(\tau)}{g(\tau)-\sigma_{\mathrm{aug}}^2\phi(\tau)}\right). \tag{C.16}$$

Consider the example where $\phi(\tau) = -(1-a(\tau))g(\tau)$ with $0 \leq a(\tau) \leq 1$. Then, the inequality (C.16) can be written as

$$\frac{2}{\tau}\frac{\|\boldsymbol{\mu}\|^2}{\frac{1+\sigma_{\mathrm{aug}}^2}{g(\tau)}+a(\tau)\|\boldsymbol{\mu}\|^2} \leq \log\left(\frac{1+\sigma_{\mathrm{aug}}^2+\sigma_{\mathrm{aug}}^2a(\tau)}{1+\sigma_{\mathrm{aug}}^2-\sigma_{\mathrm{aug}}^2a(\tau)}\right).$$

If $g(\tau) \to \widetilde{g} \neq 0$ and $a(\tau) \ll \tau^{-1}$ as $\tau \to +\infty$, then the right-hand goes to zero faster than the left-hand side, indicating that (C.16) does not hold with large temperature.

C.3 Proofs for Section 4.5

C.3.1 Proof of Theorem 4.5.1

Denote $\boldsymbol{\theta} = \mathbf{W}\boldsymbol{\beta}$, $\boldsymbol{\Omega} = \mathbf{W}^{-2}$, and $\boldsymbol{\xi} = (\gamma, \boldsymbol{\theta}^\top)^\top$. One can rewrite the logistic loss function as

$$\widetilde{\ell}_n(\boldsymbol{\xi}; \boldsymbol{h}_0) = \ell_{n,logis}(\boldsymbol{\xi}; \boldsymbol{h}_0) + \lambda_n\boldsymbol{\theta}^\top\boldsymbol{\Omega}\boldsymbol{\theta} = \mathbb{E}_n\left\{\ell(\boldsymbol{\xi}; \boldsymbol{h}_0, \lambda_n)\right\}.$$

where we define

$$\ell(\boldsymbol{\xi}; \boldsymbol{h}_0, \lambda_n) = \log\left[1 + \exp(-y(\gamma + \boldsymbol{h}_0^\top \boldsymbol{\theta}))\right] + \lambda_n \boldsymbol{\theta}^\top \boldsymbol{\Omega} \boldsymbol{\theta}$$

and $\ell_{n,logis}(\boldsymbol{\xi}; \boldsymbol{h}_0) = \mathbb{E}_n \log\left[1 + \exp(-y(\gamma + \boldsymbol{h}_0^\top \boldsymbol{\theta}))\right]$. We also denote

$$\ell_n^*(\boldsymbol{\xi}) = \mathbb{E}\left\{\ell(\boldsymbol{\xi}; \boldsymbol{h}_0, \lambda_n)\right\},$$

where the loss depends on n through λ_n.

Since both $\ell_n^*(\boldsymbol{\xi})$ and $\widetilde{\ell}_n(\boldsymbol{\xi}; \boldsymbol{h}_0)$ are strictly convex in $\boldsymbol{\xi}$ and are strongly convex when $\lambda_n \geq 0$, we denote

$$\widehat{\boldsymbol{\xi}}_n = (\widehat{\gamma}, \widehat{\boldsymbol{\theta}}^\top)^\top = \operatorname{argmin} \widetilde{\ell}_n(\boldsymbol{\xi}; \boldsymbol{h}_0), \quad \text{and} \quad \boldsymbol{\xi}_n^* = (\gamma_n^*, \boldsymbol{\theta}_n^{*\top})^\top = \operatorname{argmin} \ell_n^*(\boldsymbol{\xi}).$$

Step 1: characterizing $\boldsymbol{\xi}_n^*$. Under the 2-GMM model, we have the following characterization of the solution $\boldsymbol{\xi}_n^*$ to the population loss $\ell_n^*(\boldsymbol{\xi})$.

Lemma C.3.1. *There exists a constant $C > 0$ such that the following holds. For any given $(\lambda_n)_{n\geq 1}$ with $\lambda_n \geq 0$, the population loss minimizer $\boldsymbol{\xi}_n^*$ is unique and has the form*

$$\boldsymbol{\xi}_n^* = (0, \kappa_{\lambda_n} \boldsymbol{\mu}^\top)^\top$$

where $\kappa_{\lambda_n} \in (0, C)$ is the zero of the function

$$\psi(\kappa) := \mathbb{E}\left\{\frac{1 + (1-\kappa)\exp(\kappa\|\boldsymbol{\mu}\|^2 + \kappa \boldsymbol{\mu}^\top \mathbf{G})}{\left(1 + \exp(\kappa\|\boldsymbol{\mu}\|^2 + \kappa \boldsymbol{\mu}^\top \mathbf{G})\right)^2}\right\} - 2\lambda_n (1+\eta)^{-2}\kappa.$$

Moreover, if $\lim_{n\to\infty} \lambda_n = \infty$, then $\lim_{n\to\infty} \lambda_n \kappa_{\lambda_n} = \frac{1}{4}(1+\eta)^2$.

As a corollary, we see that $\boldsymbol{\xi}_n^*$ has bounded norm.

Step 2: characterizing $\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_n^*$. Denote the Fisher information matrix

$$\mathbf{I}_n(\boldsymbol{\xi}) = \mathbb{E}_n \left\{ \frac{\bar{\boldsymbol{h}}_0 \bar{\boldsymbol{h}}_0^\top \exp(y \bar{\boldsymbol{h}}_0^\top \boldsymbol{\xi})}{(1+\exp(y \bar{\boldsymbol{h}}_0^\top \boldsymbol{\xi}))^2} \right\}, \quad \mathbf{I}(\boldsymbol{\xi}) = \mathbb{E} \left\{ \frac{\bar{\boldsymbol{h}}_0 \bar{\boldsymbol{h}}_0^\top \exp(y \bar{\boldsymbol{h}}_0^\top \boldsymbol{\xi})}{(1+\exp(y \bar{\boldsymbol{h}}_0^\top \boldsymbol{\xi}))^2} \right\}.$$

We first have the following lemma for consistency.

Lemma C.3.2. *For any $(\lambda_n)_{n\geq 1}$ with $\lambda_n \geq 0$ and any constant $R > 0$,*

$$\sup_{\boldsymbol{\xi} \in \mathbb{R}^{p+1}:\ \|\boldsymbol{\xi}\| \leq R} \left| \mathbb{E}_n \ell(\boldsymbol{\xi}; \boldsymbol{h}_0, \lambda_n) - \mathbb{E} \ell(\boldsymbol{\xi}; \boldsymbol{h}_0, \lambda_n) \right| \xrightarrow{p} 0\,,$$
$$\sup_{\boldsymbol{\xi} \in \mathbb{R}^{p+1}:\ \|\boldsymbol{\xi}\| \leq R} \left| \mathbf{I}_n(\boldsymbol{\xi}) - \mathbf{I}(\boldsymbol{\xi}) \right| \xrightarrow{p} 0\,.$$

As a result, we have

$$\|\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_n^*\| \xrightarrow{p} 0, \qquad \text{as } n \to \infty$$

Based on the consistency result, we consider the score functions for $\widetilde{\ell}_n(\boldsymbol{\xi}; \boldsymbol{h}_0)$ and $\ell_n^*(\boldsymbol{\xi})$. Denoting $\bar{\boldsymbol{\Omega}} = \mathrm{diag}\{0, \boldsymbol{\Omega}\}$, we obtain

$$0 = \frac{\partial}{\partial \boldsymbol{\xi}} \widetilde{\ell}_n(\widehat{\boldsymbol{\xi}}_n; \boldsymbol{h}_0) = -\mathbb{E}_n \left\{ \frac{\bar{\boldsymbol{h}}_0 y}{1+\exp(y \bar{\boldsymbol{h}}_0^\top \widehat{\boldsymbol{\xi}}_n)} \right\} + 2\lambda_n \bar{\boldsymbol{\Omega}} \widehat{\boldsymbol{\xi}}_n. \tag{C.17}$$

$$0 = \frac{\partial}{\partial \boldsymbol{\xi}} \ell_n^*(\boldsymbol{\xi}_n^*) = -\mathbb{E} \left\{ \frac{\bar{\boldsymbol{h}}_0 y}{1+\exp(y \bar{\boldsymbol{h}}_0^\top \boldsymbol{\xi}_n^*)} \right\} + 2\lambda_n \bar{\boldsymbol{\Omega}} \boldsymbol{\xi}_n^*. \tag{C.18}$$

According to Lemma C.3.2, we have the first-order approximation

$$\frac{\partial}{\partial \boldsymbol{\xi}} \widetilde{\ell}_{n,logis}(\widehat{\boldsymbol{\xi}}_n; \boldsymbol{h}_0) - \frac{\partial}{\partial \boldsymbol{\xi}} \widetilde{\ell}_{n,logis}(\boldsymbol{\xi}_n^*; \boldsymbol{h}_0) = \mathbf{I}_n(\boldsymbol{\xi}_n^*)(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}^*) + o_{\mathbb{P}}(\|\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_n^*\|).$$

By calculating the difference between (C.17) and (C.18), we obtain

$$-\mathbf{I}_n(\boldsymbol{\xi}_n^*)(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_n^*) - 2\lambda_n\bar{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_n^*) + (\mathbb{E}_n - \mathbb{E})\left[\frac{\bar{\boldsymbol{h}}_0 y}{1+\exp(y\bar{\boldsymbol{h}}_0^\top \boldsymbol{\xi}_n^*)}\right] + o_{\mathbb{P}}(\|\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_n^*\|) = 0.$$

When p is fixed and $n \to +\infty$, since $\boldsymbol{\xi}_n^*$ has bounded norm, then by Lemma C.3.2, we have $\|\left(\mathbf{I}_n(\boldsymbol{\xi}_n^*) - \mathbf{I}(\boldsymbol{\xi}_n^*)\right)(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_n^*)\| = o_{\mathbb{P}}(\|\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_n^*\|)$. In addition, denote

$$\boldsymbol{z}_n = \frac{\bar{\boldsymbol{h}}_0 y}{1+\exp(y\bar{\boldsymbol{h}}_0^\top \boldsymbol{\xi}_n^*)},$$

where $\mathbb{E}\boldsymbol{z}_n = \mathbf{0}$ and $cov(\boldsymbol{z}_n) = \mathbf{I}(\boldsymbol{\xi}_n^*)$, then we have the following lemma.

Lemma C.3.3. *We have* $\mathbf{I}(\boldsymbol{\xi}_n^*) \succ \mathbf{0}$ *and by the Lindeberg-Feller central limit theorem,*

$$\sqrt{n}\left(\mathbf{I}(\boldsymbol{\xi}_n^*)\right)^{-1/2}(\mathbb{E}_n - \mathbb{E})\left[\frac{\bar{\boldsymbol{h}}_0 y}{1+\exp(y\bar{\boldsymbol{h}}_0^\top \boldsymbol{\xi}_n^*)}\right] \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Let $\bar{\boldsymbol{\zeta}} = \sqrt{n}(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_n^*)$, then

$$\left(\mathbf{I}(\boldsymbol{\xi}_n^*)\right)^{-1/2}\left(\mathbf{I}(\boldsymbol{\xi}_n^*) + 2\lambda_n\bar{\boldsymbol{\Omega}}\right)\bar{\boldsymbol{\zeta}} \xrightarrow{d} \mathcal{N}(0, \mathbf{I}). \tag{C.19}$$

Step 3: calculating misclassification error. Conditioning on $y = 1$,

$$\texttt{Err}_1 = \mathbb{P}(y(\langle \boldsymbol{h}_0, \widehat{\boldsymbol{\theta}}\rangle + \widehat{\gamma}) < 0 | y = 1) = \mathbb{P}(\langle \boldsymbol{\mu} + \mathbf{G}, \kappa_{\lambda_n}\boldsymbol{\mu} + \frac{1}{\sqrt{n}}\boldsymbol{\zeta}\rangle + \widehat{\gamma} < 0), \tag{C.20}$$

where $\mathbf{G} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then (C.20) can be written using the Gaussian cdf as

$$\begin{aligned}(\text{C.20}) &= \Phi\left(-\frac{\widehat{\gamma} + \langle \boldsymbol{\mu}, \kappa_{\lambda_n}\boldsymbol{\mu} + \frac{1}{\sqrt{n}}\boldsymbol{\zeta}\rangle}{\|\kappa_{\lambda_n}\boldsymbol{\mu} + \frac{1}{\sqrt{n}}\boldsymbol{\zeta}\|}\right) \\ &= \Phi\left(-\|\boldsymbol{\mu}\|\cos(\boldsymbol{\mu}, \kappa_{\lambda_n}\boldsymbol{\mu} + \frac{1}{\sqrt{n}}\boldsymbol{\zeta}) - \frac{\widehat{\gamma}}{\|\kappa_{\lambda_n}\boldsymbol{\mu} + \frac{1}{\sqrt{n}}\boldsymbol{\zeta}\|}\right).\end{aligned}$$

By the symmetry of the distribution of y and the fact that $y\boldsymbol{h}_0 \sim \boldsymbol{\mu} + \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, if we condition on $y = -1$,

$$\begin{aligned}\texttt{Err}_{-1} &= \mathbb{P}(y(\langle \boldsymbol{h}_0, \widehat{\boldsymbol{\theta}}\rangle + \widehat{\gamma}) < 0 | y = -1) = \mathbb{P}(\langle \boldsymbol{\mu} + \mathbf{G}, \kappa_{\lambda_n}\boldsymbol{\mu} + \frac{1}{\sqrt{n}}\boldsymbol{\zeta}\rangle - \widehat{\gamma} < 0) \\ &= \Phi\left(-\|\boldsymbol{\mu}\|\cos(\boldsymbol{\mu}, \kappa_{\lambda_n}\boldsymbol{\mu} + \frac{1}{\sqrt{n}}\boldsymbol{\zeta}) + \frac{\widehat{\gamma}}{\|\kappa_{\lambda_n}\boldsymbol{\mu} + \frac{1}{\sqrt{n}}\boldsymbol{\zeta}\|}\right).\end{aligned}$$

Then, we will next analyze the two terms

$$\cos(\boldsymbol{\mu}, \kappa_{\lambda_n}\boldsymbol{\mu} + \frac{1}{\sqrt{n}}\boldsymbol{\zeta}) \qquad \text{and} \qquad \frac{\widehat{\gamma}}{\|\kappa_{\lambda_n}\boldsymbol{\mu} + \frac{1}{\sqrt{n}}\boldsymbol{\zeta}\|}$$

to see how they will affect the misclassification error.

Step 4: finalizing conclusions. We consider the following regimes.

Case 1: $\limsup_{n\to+\infty} \lambda_n \leq \bar{\lambda} < +\infty$**.** In this case, by Lemma C.3.1, we have $\liminf_{n\to+\infty} \kappa_{\lambda_n} \geq c_\kappa > 0$. Then,

$$\begin{aligned}\cos(\boldsymbol{\mu}, \kappa_{\lambda_n}\boldsymbol{\mu} + \frac{1}{\sqrt{n}}\boldsymbol{\zeta}) &= 1 + o_{\mathbb{P}}(1) \\ \frac{\widehat{\gamma}}{\|\kappa_{\lambda_n}\boldsymbol{\mu} + \frac{1}{\sqrt{n}}\boldsymbol{\zeta}\|} &= \frac{\widehat{\gamma}}{\|\kappa_{\lambda_n}\boldsymbol{\mu}\|} + o_{\mathbb{P}}(1).\end{aligned}$$

Denote $q(t) = \frac{1}{1+e^t}$, since $\boldsymbol{h}_0 y \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_p)$, one can check that for any $\boldsymbol{\xi} = (\gamma, \boldsymbol{\theta}^\top)^\top$ and $2 \le j \le p+1$,

$$\begin{aligned}[\mathbf{I}(\boldsymbol{\xi})]_{j,1} =& \frac{1}{2}\mathbb{E}[q(\boldsymbol{h}_0^\top\boldsymbol{\theta}y + \gamma y)(1 - q(\boldsymbol{h}_0^\top\boldsymbol{\theta}y + \gamma y))(\boldsymbol{h}_0)_j | y = 1] \\ &+ \frac{1}{2}\mathbb{E}[q(\boldsymbol{h}_0^\top\boldsymbol{\theta}y + \gamma y)(1 - q(\boldsymbol{h}_0^\top\boldsymbol{\theta}y + \gamma y))(\boldsymbol{h}_0)_j | y = -1] = 0.\end{aligned}$$

Then, by (C.19), we have

$$\sqrt{ni(\boldsymbol{\theta}_n^*)}\widehat{\gamma} \xrightarrow{d} \mathcal{N}(0,1),$$

where $i(\boldsymbol{\theta}_n^*) := [\mathbf{I}(\boldsymbol{\xi}_n^*)]_{1,1}$ with $\gamma^* = 0$ and $[\mathbf{I}(\boldsymbol{\theta}_n^*)]_{1,1}$ denotes the submatrix consisting of the first column and the last row of $\mathbf{I}(\boldsymbol{\theta}_n^*)$. Since $\liminf_{n\to+\infty} i(\boldsymbol{\theta}_n^*) > 0$, we have $\widehat{\gamma} = [ni(\boldsymbol{\theta}_n^*)]^{-1/2}G' + o_{\mathbb{P}}(1)$ with $G' \sim \mathcal{N}(0,1)$, then

$$\texttt{Err} = \frac{1}{2}\texttt{Err}_1 + \frac{1}{2}\texttt{Err}_{-1} = \Phi\left(-\|\boldsymbol{\mu}\| + \frac{[i(\boldsymbol{\theta}_n^*)]^{-1/2}G'}{\sqrt{n}\kappa_{\lambda_n}\|\boldsymbol{\mu}\|}\right) + o_{\mathbb{P}}(1).$$

From the reasoning above, we have $\kappa_{\lambda_n}^{-1}[ni(\boldsymbol{\theta}_n^*)]^{-1/2} = o(1)$, then

$$\texttt{Err} = \Phi\left(-\|\boldsymbol{\mu}\|\right) + o_{\mathbb{P}}(1).$$

Case 2: $\limsup_{n\to+\infty} \lambda_n = +\infty$**.** Without loss of generality, consider $\lambda_n \to +\infty$ as $n \to +\infty$. By Lemma C.3.1, $\lambda_n\kappa_{\lambda_n} = \frac{1}{4}(1+\eta)^{-2} + o(1)$, thus we have $\mathbf{I}(\boldsymbol{\xi}_n^*) = \frac{1}{4}\mathbf{I}_{p+1} + o(1)$, particularly $i(\boldsymbol{\theta}_n^*) = \frac{1}{4} + o(1)$. By (C.19), we have $\boldsymbol{\zeta} = O_{\mathbb{P}}(\kappa_{\lambda_n})$, then similar with before

$$\begin{aligned}&\cos(\boldsymbol{\mu}, \kappa_{\lambda_n}\boldsymbol{\mu} + \frac{1}{\sqrt{n}}\boldsymbol{\zeta}) = 1 + o_{\mathbb{P}}(1) \\ &\frac{\widehat{\gamma}}{\|\kappa_{\lambda_n}\boldsymbol{\mu} + \frac{1}{\sqrt{n}}\boldsymbol{\zeta}\|} = \frac{2}{\sqrt{n}\kappa_{\lambda_n}\|\boldsymbol{\mu}\|} + o_{\mathbb{P}}(1).\end{aligned}$$

1. If $\sqrt{n}/\lambda_n = o(1)$, then $\sqrt{n}\kappa_{\lambda_n} = o(1)$. We have $\mathbb{E}[\texttt{Err}] \to 1$ as $\lambda_n \to +\infty$.

2. If $\lambda_n = a \cdot \sqrt{n}$ with a positive constant a, then $a\sqrt{n}\kappa_n = \frac{1}{4}(1+\eta)^2 + o(1)$ and

$$\texttt{Err} = \Phi\left(-\|\boldsymbol{\mu}\| + \frac{8G'}{a(1+\eta)^2\|\boldsymbol{\mu}\|}\right) + o_{\mathbb{P}}(1).$$

   We have the following lemma

   Lemma C.3.4. *Let $\tau,\ w > 0$ and $U \sim \mathcal{N}(0,1)$, then*

$$\frac{d}{dw}\mathbb{E}\Phi(-\tau + w \cdot U) = \frac{\tau w}{\sqrt{2\pi}} \frac{1}{(1+w^2)^{3/2}} \exp\left(-\frac{\tau^2}{2(1+w^2)}\right) > 0.$$

   Then, $\mathbb{E}[\texttt{Err}]$ is increasing in $\frac{8}{a(1+\eta)^2\|\boldsymbol{\mu}\|}$, thus is decreasing in η.

3. If $\lambda_n = a \cdot b_n$ with a constant $a > 0$ and $b_n/\sqrt{n} = o(1)$, then $ab_n\kappa_n = \frac{1}{4}(1+\eta)^2 + o(1)$ and

$$\texttt{Err} = \Phi\left(-\|\boldsymbol{\mu}\| + \frac{8G'}{a(1+\eta)^2\|\boldsymbol{\mu}\|}\frac{b_n}{\sqrt{n}}\right) + o_{\mathbb{P}}(1) = \Phi(-\|\boldsymbol{\mu}\|) + O_{\mathbb{P}}\left(\frac{b_n}{\sqrt{n}}\right).$$

   As a result, $\mathbb{E}[\texttt{Err}] = \Phi(-\|\boldsymbol{\mu}\|) + O(b_n/\sqrt{n})$.

Combining pieces above, low-dimensional logistic regression with ℓ_2 regularization has decreasing prediction error in the expansion parameter η only if $\lambda_n \asymp \sqrt{n}$. Otherwise, $\texttt{Err}$ asymptotically remains constant in η.

Proof of Lemma C.3.1

Since the population loss function $\widetilde{\ell}(\boldsymbol{\xi})$ is strongly convex in $\boldsymbol{\xi}$, it suffices to show that (i) there exists some $\kappa := \kappa_{\lambda_n} \in (0, C)$ such that $\boldsymbol{\xi} = (0, \kappa\boldsymbol{\mu}^\top)^\top$ satisfies the score equation

$$\mathbb{E}\left\{\frac{\bar{\boldsymbol{h}}_0 y}{1+\exp(y\bar{\boldsymbol{h}}_0^\top \boldsymbol{\xi})}\right\} = 2\lambda_n \bar{\boldsymbol{\Omega}}\boldsymbol{\xi}; \tag{C.21}$$

and that (ii) $\lim_{n\to\infty} \lambda_n \kappa_{\lambda_n} = \frac{1}{4}(1+\eta)^{-2}$. Due to the block structure of $\bar{\boldsymbol{\Omega}}$, we separate the verification process into two parts.

Step 1: check the first coordinate. Clearly, the first coordinate on the right-hand side of (C.21) is zero. When it comes to the left-hand side, we have

$$\mathbb{E}\left\{\frac{y}{1+\exp(y\bar{\boldsymbol{h}}_0^\top \boldsymbol{\xi})}\right\} = \frac{1}{2}\mathbb{E}\left\{\frac{1}{1+\exp(\kappa\boldsymbol{\mu}^\top(\boldsymbol{\mu}+\mathbf{G}))}\right\} + \frac{1}{2}\mathbb{E}\left\{\frac{-1}{1+\exp(\kappa\boldsymbol{\mu}^\top(\boldsymbol{\mu}+\mathbf{G}))}\right\} = 0. \tag{C.22}$$

Here, $\mathbf{G} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. Thus, (C.22) holds in the first coordinate for $\boldsymbol{\xi} = (0, \kappa\boldsymbol{\mu}^\top)^\top$.

Step 2: check the remaining coordinates. Now we move on to the remaining coordinates, i.e., the coordinates corresponding to $\boldsymbol{h}_0$.

The right-hand side of (C.21) obeys

$$2\lambda_n \boldsymbol{\Omega}\boldsymbol{\theta} = 2\lambda_n (1+\eta)^{-2}\kappa\boldsymbol{\mu},$$

by the definition of $\mathbf{\Omega}$. For the left hand side, since $\boldsymbol{h}_0 y \overset{d}{=} \boldsymbol{\mu} + \mathbf{G} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_p)$, we have

$$\begin{aligned}\mathbb{E}\left\{\frac{\boldsymbol{h}_0 y}{1+\exp(y\kappa \boldsymbol{h}_0^\top \boldsymbol{\mu})}\right\} &= \mathbb{E}\left\{\frac{\boldsymbol{\mu}+\mathbf{G}}{1+\exp(\kappa\|\boldsymbol{\mu}\|^2+\kappa\boldsymbol{\mu}^\top\mathbf{G})}\right\} \\ &= \mathbb{E}\left\{\frac{1}{1+\exp(\kappa\|\boldsymbol{\mu}\|^2+\kappa\boldsymbol{\mu}^\top\mathbf{G})}\right\}\boldsymbol{\mu} + \mathbb{E}\left\{\frac{\mathbf{G}}{1+\exp(\kappa\|\boldsymbol{\mu}\|^2+\kappa\boldsymbol{\mu}^\top\mathbf{G})}\right\}.\end{aligned}$$

Apply Stein's identity to see that

$$\mathbb{E}\left\{\frac{\mathbf{G}}{1+\exp(\kappa\|\boldsymbol{\mu}\|^2+\kappa\boldsymbol{\mu}^\top\mathbf{G})}\right\} = -\kappa\mathbb{E}\left\{\frac{\exp(\kappa\|\boldsymbol{\mu}\|^2+\kappa\boldsymbol{\mu}^\top\mathbf{G})}{\left(1+\exp(\kappa\|\boldsymbol{\mu}\|^2+\kappa\boldsymbol{\mu}^\top\mathbf{G})\right)^2}\right\}\boldsymbol{\mu},$$

which leads to the conclusion

$$\mathbb{E}\left\{\frac{\boldsymbol{h}_0 y}{1+\exp(y\kappa \boldsymbol{h}_0^\top \boldsymbol{\mu})}\right\} = \mathbb{E}\left\{\frac{1+(1-\kappa)\exp(\kappa\|\boldsymbol{\mu}\|^2+\kappa\boldsymbol{\mu}^\top\mathbf{G})}{\left(1+\exp(\kappa\|\boldsymbol{\mu}\|^2+\kappa\boldsymbol{\mu}^\top\mathbf{G})\right)^2}\right\}\boldsymbol{\mu}. \tag{C.23}$$

Then it boils down to showing that there exists a constant $\kappa > 0$ such that

$$\mathbb{E}\left\{\frac{1+(1-\kappa)\exp(\kappa\|\boldsymbol{\mu}\|^2\kappa\boldsymbol{\mu}^\top\mathbf{G})}{\left(1+\exp(\kappa\|\boldsymbol{\mu}\|^2+\kappa\boldsymbol{\mu}^\top\mathbf{G})\right)^2}\right\} = 2\lambda_n(1+\eta)^{-2}\kappa.$$

Recall the function

$$\psi(\kappa) = \mathbb{E}\left\{\frac{1+(1-\kappa)\exp(\kappa\|\boldsymbol{\mu}\|^2+\kappa\boldsymbol{\mu}^\top\mathbf{G})}{\left(1+\exp(\kappa\|\boldsymbol{\mu}\|^2+\kappa\boldsymbol{\mu}^\top\mathbf{G})\right)^2}\right\} - 2\lambda_n(1+\eta)^{-2}\kappa.$$

We see that $\psi(0) = \frac{1}{2} > 0$.

(i) First, we claim that there exists $C > 0$ depending solely on $\|\boldsymbol{\mu}\|$ such that $\psi(C) < 0$. Once this is proved, since $\psi(\kappa)$ is a continuous function, by the intermediate value theorem, there must exist $\kappa \in (0, C)$ such that $\psi(\kappa) = 0$.

To prove this claim, we use the identity (C.23) to express $\psi(\kappa)$ into

$$\begin{aligned}\psi(\kappa) &= \|\boldsymbol{\mu}\|^{-2} \cdot \mathbb{E}\left\{\frac{\|\boldsymbol{\mu}\|^2 + \mathbf{G}^\top \boldsymbol{\mu}}{1+\exp(\kappa\|\boldsymbol{\mu}\|^2 + \kappa\boldsymbol{\mu}^\top \mathbf{G})}\right\} - 2\lambda_n(1+\eta)^{-2}\kappa \\ &\leq \|\boldsymbol{\mu}\|^{-2} \cdot \mathbb{E}\left\{\frac{\|\boldsymbol{\mu}\|^2 + \|\boldsymbol{\mu}\|G}{1+\exp(\kappa\|\boldsymbol{\mu}\|^2 + \kappa\|\boldsymbol{\mu}\|G)}\right\}\end{aligned}$$

where $G \sim \mathcal{N}(0,1)$. Denote an event $A = \{\|\boldsymbol{\mu}\|^2 + \|\boldsymbol{\mu}\|G < 0\}$. For any fixed $\|\boldsymbol{\mu}\| > 0$, the probability $\mathbb{P}(A) \in (0,1)$. By the dominated convergence theorem,

$$\begin{aligned}\lim_{\kappa\to\infty} \mathbb{E}\left\{\frac{(\|\boldsymbol{\mu}\|^2 + \|\boldsymbol{\mu}\|G)_+}{1+\exp(\kappa\|\boldsymbol{\mu}\|^2 + \kappa\|\boldsymbol{\mu}\|G)}\right\} &= 0, \\ \lim_{\kappa\to\infty} \mathbb{E}\left\{\frac{(\|\boldsymbol{\mu}\|^2 + \|\boldsymbol{\mu}\|G)_-}{1+\exp(\kappa\|\boldsymbol{\mu}\|^2 + \kappa\|\boldsymbol{\mu}\|G)}\right\} &= \mathbb{E}\big[(\|\boldsymbol{\mu}\|^2 + \|\boldsymbol{\mu}\|G)_-\big] > 0.\end{aligned}$$

Combining the two parts, we deduce that $\limsup_{\kappa\to\infty} \psi(\kappa) \leq -\mathbb{E}\big[(\|\boldsymbol{\mu}\|^2 + \|\boldsymbol{\mu}\|G)_-\big] < 0$.

(ii) Since for any $a \in \mathbb{R}$, $a < \kappa^{-1} + a \leq \kappa^{-1}\exp(\kappa a)$, we have

$$\mathbb{E}\left\{\frac{\|\boldsymbol{\mu}\|^2 + \|\boldsymbol{\mu}\|G}{1+\exp(\kappa\|\boldsymbol{\mu}\|^2 + \kappa\|\boldsymbol{\mu}\|G)}\right\} \leq \kappa^{-1}\mathbb{E}\left\{\frac{\exp(\kappa\|\boldsymbol{\mu}\|^2 + \kappa\|\boldsymbol{\mu}\|G)}{1+\exp(\kappa\|\boldsymbol{\mu}\|^2 + \kappa\|\boldsymbol{\mu}\|G)}\right\} < \kappa^{-1}.$$

The equation $\psi(\kappa_{\lambda_n}) = 0$ must imply

$$2\lambda_n(1+\eta)^{-2}\kappa_{\lambda_n} < \frac{1}{\kappa_{\lambda_n}\|\boldsymbol{\mu}\|^2}.$$

If $\lambda_n \to \infty$ as $n \to \infty$, then we must have $\kappa_{\lambda_n} = o(1)$. Taking the limit $\lim_{n\to\infty} \psi(\kappa_{\lambda_n})$, we get

$$\frac{1}{2} - 2(1+\eta)^{-2} \lim_{n\to\infty} \lambda_n\kappa_{\lambda_n} = 0$$

so $\lim_{n\to\infty} \lambda_n\kappa_{\lambda_n} = \frac{1}{4}(1+\eta)^2$.

Combining the two steps finishes the proof.

Proof of Lemma C.3.2

Recall that for any $\lambda_n \geq 0$ and any $\boldsymbol{\xi} = (\gamma, \boldsymbol{\theta}^\top)^\top \in \mathbb{R}^{p+1}$,

$$\ell(\boldsymbol{\xi}; \boldsymbol{h}_0, \lambda_n) = \log\left[1 + \exp(-y(\gamma + \boldsymbol{h}_0^\top \boldsymbol{\theta}))\right] + \lambda_n \boldsymbol{\theta}^\top \boldsymbol{\Omega} \boldsymbol{\theta}.$$

Step 1: Applying the uniform law of large numbers. Since the term $\lambda_n \boldsymbol{\theta}^\top \boldsymbol{\Omega} \boldsymbol{\theta}$ is canceled out, we need to show that for any R and the compact ball $B_R = \{\|\boldsymbol{\xi}\| \leq R\}$,

$$\sup_{\boldsymbol{\xi} \in \mathbf{B}_R} \left| (\mathbb{E}_n - \mathbb{E}) \left\{ \log\left[1 + \exp(-y(\gamma + \boldsymbol{h}_0^\top \boldsymbol{\theta}))\right] \right\} \right| \xrightarrow{p} 0. \tag{C.24}$$

We see that $\ell(\boldsymbol{\xi}; \boldsymbol{h}_0, \lambda_n)$ is continuous in $\boldsymbol{\xi}$ for any $\boldsymbol{h}_0$ and is measurable in $(\boldsymbol{h}_0, y)$ for any $\boldsymbol{\xi}$. In addition, for any R,

$$\log\left[1 + \exp(-y(\gamma + \boldsymbol{h}_0^\top \boldsymbol{\theta}))\right] \leq \log\left[1 + \exp(R\|\bar{\boldsymbol{h}}_0\|)\right] \quad \text{with} \quad \mathbb{E}\left\{\log\left[1 + \exp(R\|\bar{\boldsymbol{h}}_0\|)\right]\right\} < +\infty.$$

Then, we can apply the uniform law of large numbers (Newey and McFadden, 1994, Lemma 2.4) to show (C.24).

It is similar for the convergence of Fisher information as for any $\boldsymbol{\xi} \in \mathbb{R}^{p+1}$,

$$\frac{\bar{\boldsymbol{h}}_0 \bar{\boldsymbol{h}}_0^\top \exp(y \bar{\boldsymbol{h}}_0^\top \boldsymbol{\xi})}{(1 + \exp(y \bar{\boldsymbol{h}}_0^\top \boldsymbol{\xi}))^2} \leq \frac{1}{4} \bar{\boldsymbol{h}}_0 \bar{\boldsymbol{h}}_0^\top \quad \text{with} \quad \mathbb{E}\left\{\frac{1}{4} \bar{\boldsymbol{h}}_0 \bar{\boldsymbol{h}}_0^\top\right\} = \frac{1}{4} \mathbf{I}_{p+1}.$$

Then, we have

$$\sup_{\boldsymbol{\xi} \in \mathbf{B}_R} \left| \mathbf{I}_n(\boldsymbol{\xi}) - \mathbf{I}(\boldsymbol{\xi}) \right| \xrightarrow{p} 0.$$

Step 2: Proving the convergence of the minimizer. With $\widehat{\boldsymbol{\xi}}_n$ and $\boldsymbol{\xi}_n^*$ defined before, there exists a constant $R > 0$ such that $\mathbb{P}(\widehat{\boldsymbol{\xi}}_n \notin \mathbf{B}_R) = o(1)$. Therefore,

$$\ell_n^*(\widehat{\boldsymbol{\xi}}_n) - \ell_n^*(\boldsymbol{\xi}_n^*) = \ell_n^*(\widehat{\boldsymbol{\xi}}_n) - \widetilde{\ell}_n(\widehat{\boldsymbol{\xi}}_n; \boldsymbol{h}_0) + \widetilde{\ell}_n(\widehat{\boldsymbol{\xi}}_n; \boldsymbol{h}_0) - \widetilde{\ell}_n(\boldsymbol{\xi}_n^*; \boldsymbol{h}_0) + \widetilde{\ell}_n(\boldsymbol{\xi}_n^*; \boldsymbol{h}_0) - \ell_n^*(\boldsymbol{\xi}_n^*).$$

Here, we have $\ell_n^*(\widehat{\boldsymbol{\xi}}_n) - \widetilde{\ell}_n(\widehat{\boldsymbol{\xi}}_n; \boldsymbol{h}_0) \leq \sup_{\boldsymbol{\xi} \in \mathbf{B}_R} \left|(\mathbb{E}_n - \mathbb{E})\ell(\boldsymbol{\xi}; \boldsymbol{h}_0, \lambda_n)\right| \xrightarrow{p} 0$, $\widetilde{\ell}_n(\widehat{\boldsymbol{\xi}}_n; \boldsymbol{h}_0) - \widetilde{\ell}_n(\boldsymbol{\xi}_n^*; \boldsymbol{h}_0) \leq 0$ by the optimality of $\widehat{\boldsymbol{\xi}}_n$, and $\widetilde{\ell}_n(\boldsymbol{\xi}_n^*; \boldsymbol{h}_0) - \ell_n^*(\boldsymbol{\xi}_n^*) = o_{\mathbb{P}}(1)$ by the law of large numbers. Consequently, we have $\ell_n^*(\widehat{\boldsymbol{\xi}}_n) - \ell_n^*(\boldsymbol{\xi}_n^*) \leq o_{\mathbb{P}}(1)$. Since $\ell_n^*(\boldsymbol{\xi})$ is strictly convex and by definition $\boldsymbol{\xi}_n^* = \operatorname*{argmin}_{\boldsymbol{\xi} \in \mathbb{R}^{p+1}} \ell_n^*(\boldsymbol{\xi})$, we further have $\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_n^* = o_{\mathbb{P}}(1)$.

Proof of Lemma C.3.3

Denoting

$$\boldsymbol{z}_{n,i} = \frac{\bar{\boldsymbol{h}}_{0,i} y_i}{1 + \exp(y_i \bar{\boldsymbol{h}}_{0,i}^\top \boldsymbol{\xi}_n^*)},$$

then it suffices to show that for any $\varepsilon > 0$,

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}\left\{\|\boldsymbol{z}_{n,i}\|^2 \mathbb{1}\{\|\boldsymbol{z}_{n,i}\| \geq \varepsilon\sqrt{n}\}\right\} = \mathbb{E}\left\{\|\boldsymbol{z}_{n,1}\|^2 \mathbb{1}\{\|\boldsymbol{z}_{n,1}\| \geq \varepsilon\sqrt{n}\}\right\} \xrightarrow{p} 0.$$

Let $Z_{n,i} = \|\boldsymbol{z}_{n,i}\|^2$, then $\mathbb{E}Z_{n,i} = \operatorname{Tr}(\mathbb{E}(\boldsymbol{z}_{n,i}\boldsymbol{z}_{n,i}^\top)) = \operatorname{Tr}(\mathbf{I}(\boldsymbol{\xi}_n^*))$. Let $W_{n,i} = Z_{n,i}\mathbb{1}\{Z_{n,i} \geq \varepsilon^2 n\}$. We can verify that $0 \leq W_{n,i} \leq Z_{n,i}$ and for any $\delta > 0$, by Markov's inequality,

$$\mathbb{P}\left(W_{n,i} \geq \delta\right) = \mathbb{P}\left(Z_{n,i} \geq \epsilon^2 n\right) \leq \frac{\operatorname{Tr}(\mathbf{I}(\boldsymbol{\xi}_n^*))}{n\varepsilon^2} \to 0, \quad \text{as } n \to 0,$$

which shows that $W_{n,i} \xrightarrow{p} 0$. Then, by the dominated convergence theorem, we have

$$\lim_{n \to +\infty} \mathbb{E}W_{n,i} = \mathbb{E} \lim_{n \to +\infty} W_{n,i} = 0,$$

which verifies Lindeberg's condition. By the Lindeberg-Feller theorem for multivariate random variables, we have

$$\sqrt{n}\left[\mathbf{I}(\boldsymbol{\xi}_n^*)\right]^{-1/2}\left(\mathbb{E}_n \boldsymbol{z}_{n,i} - \mathbb{E}\boldsymbol{z}_{n,i}\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

C.3.2 High-dimensional regime: max-margin classifier

Written in terms of the input data $\{\boldsymbol{h}_{0,i}\}$, the max-margin classifier (4.5) is equivalent to

$$\begin{aligned} \max_{\boldsymbol{\beta}\in\mathbb{R}^p} \quad & \min_{i\le n} y_i\langle \boldsymbol{\beta}, \boldsymbol{h}_{0,i}\rangle \\ \text{subject to} \quad & \|\boldsymbol{\beta}\|_{\boldsymbol{\Omega}} = \sqrt{\boldsymbol{\beta}^\top \boldsymbol{\Omega}\boldsymbol{\beta}} \le 1, \end{aligned} \tag{C.25}$$

where $\boldsymbol{\Omega} \coloneqq (\mathbf{I} + \eta\boldsymbol{\mu}_0\boldsymbol{\mu}_0^\top)^{-2} = \mathbf{I} - \left(1 - \frac{1}{(1+\eta)^2}\right)\boldsymbol{\mu}_0\boldsymbol{\mu}_0^\top$, and $\boldsymbol{\mu}_0 = \frac{1}{\|\boldsymbol{\mu}\|}\boldsymbol{\mu}$.

Preliminaries

Before presenting the proof of Theorem 4.5.2, we first introduce important quantities and their equivalence.

Define a compact set $\Theta = \{\boldsymbol{\beta} \in \mathbb{R}^p : \boldsymbol{\beta}^\top\boldsymbol{\Omega}\boldsymbol{\beta} \le 1\}$. Similar to the development in Montanari et al. (2019), for any positive margin $\kappa > 0$, we have the equivalence

$$\begin{aligned} \mathcal{E}_{n,p,\kappa} &= \left\{\text{there exists } \boldsymbol{\beta}\in\Theta : y_i\langle \boldsymbol{h}_{0,i}, \boldsymbol{\beta}\rangle \ge \kappa \text{ for all } i \le n\right\} \\ &= \left\{ \min_{\boldsymbol{\beta}\in\Theta} \max_{\boldsymbol{\lambda}:\|\boldsymbol{\lambda}\|\le 1, \mathbf{y}\odot\boldsymbol{\lambda}\ge 0} \frac{1}{\sqrt{p}}\boldsymbol{\lambda}^\top(\kappa\mathbf{y} - \mathbf{H}_0\boldsymbol{\beta}) = 0 \right\}. \end{aligned}$$

To simplify the notation, for any set $\mathbf{B}_p \subseteq \mathbb{R}^p$, we define the quantity

$$\xi_{n,p,\kappa}(\mathbf{B}_p) = \min_{\boldsymbol{\beta}\in\mathbf{B}_p} \max_{\boldsymbol{\lambda}:\|\boldsymbol{\lambda}\|\le 1, \mathbf{y}\odot\boldsymbol{\lambda}\ge 0} \frac{1}{\sqrt{p}}\boldsymbol{\lambda}^\top(\kappa\mathbf{y} - \mathbf{H}_0\boldsymbol{\beta}).$$

In particular, we set $\xi_{n,p,\kappa} \coloneqq \xi_{n,p,\kappa}(\Theta)$, which obeys

$$\{\xi_{n,p,\kappa} > 0\} \Longleftrightarrow \mathcal{E}_{n,p,\kappa}^{c} \quad \text{and} \quad \{\xi_{n,p,\kappa} = 0\} \Longleftrightarrow \mathcal{E}_{n,p,\kappa}.$$

As a result, our goal is to analyze whether $\xi_{n,p,\kappa}$ is positive.

Step 1: Applying CGMT. Recall the Gaussian mixture model where $\mathbf{H}_0 = \mathbf{y}\boldsymbol{\mu}^\top + \mathbf{G}$ and each entry of $\mathbf{G} \in \mathbb{R}^{n\times p}$ is independently drawn from $\mathcal{N}(0,1)$. As a result, we have $\boldsymbol{\lambda}^\top(\kappa\mathbf{y} - \mathbf{H}_0\boldsymbol{\beta}) = \boldsymbol{\lambda}^\top(\kappa\mathbf{y} - \langle\boldsymbol{\mu},\boldsymbol{\beta}\rangle\mathbf{y} - \mathbf{G}\boldsymbol{\beta})$, which is a bilinear form of the Gaussian random matrix $\mathbf{G}$. We plan to use Gordan's comparison inequality to simplify the calculation of $\xi_{n,p,\kappa}(\mathbf{B}_p)$. To do so, we introduce another quantity for an arbitrary set $\mathbf{B}_p$:

$$\xi_{n,p,\kappa}^{(1)}(\mathbf{B}_p) = \min_{\boldsymbol{\beta}\in\mathbf{B}_p} \max_{\boldsymbol{\lambda}:\|\boldsymbol{\lambda}\|\le 1, \mathbf{y}\odot\boldsymbol{\lambda}\ge 0} \frac{1}{\sqrt{p}} \left\{ \|\boldsymbol{\lambda}\|\, \mathbf{g}^\top\boldsymbol{\beta} + \|\boldsymbol{\beta}\|\, \mathbf{h}^\top\boldsymbol{\lambda} + \kappa\boldsymbol{\lambda}^\top\mathbf{y} - \langle\boldsymbol{\mu},\boldsymbol{\beta}\rangle\boldsymbol{\lambda}^\top\mathbf{y} \right\},$$

where $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and $\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ are two independent Gaussian vectors. Similar to before, we denote $\xi_{n,p,\kappa}^{(1)} \coloneqq \xi_{n,p,\kappa}^{(1)}(\Theta)$. The following lemma connects $\xi_{n,p,\kappa}(\mathbf{B}_p)$ with $\xi_{n,p,\kappa}^{(1)}(\mathbf{B}_p)$, which is a simple corollary of Gordon's comparison inequality (Gordon, 1988; Thrampoulidis et al., 2015; Montanari et al., 2019).

Lemma C.3.5. *For any $t \in \mathbb{R}$ and any compact set $\mathbf{B}_p$,*

$$\mathbb{P}\left(\xi_{n,p,\kappa} \le t\right) \le 2\mathbb{P}\left(\xi_{n,p,\kappa}^{(1)} \le t\right) \quad \textit{and} \quad \mathbb{P}\left(\xi_{n,p,\kappa} \ge t\right) \le 2\mathbb{P}\left(\xi_{n,p,\kappa}^{(1)} \ge t\right).$$
$$\mathbb{P}\left(\xi_{n,p,\kappa}(\mathbf{B}_p) \le t\right) \le 2\mathbb{P}\left(\xi_{n,p,\kappa}^{(1)}(\mathbf{B}_p) \le t\right).$$

If in addition $\mathbf{B}_p$ is convex, then

$$\mathbb{P}\left(\xi_{n,p,\kappa}(\mathbf{B}_p) \ge t\right) \le 2\mathbb{P}\left(\xi_{n,p,\kappa}^{(1)}(\mathbf{B}_p) \ge t\right).$$

Step 2: Connecting $\xi^{(1)}_{n,p,\kappa}$ with $\widetilde{\xi}^{(1)}_{n,p,\kappa}$. We then move on to characterizing $\xi^{(1)}_{n,p,\kappa}$ by translating the min-max problem into a single minimization problem. For any set $\mathbf{B} \subseteq \mathbb{R} \times \mathbb{R}_+$, we define a quantity

$$\widetilde{\xi}^{(1)}_{n,p,\kappa}(\mathbf{B}) = \min_{(\gamma,z)\in\mathbf{B}} \frac{1}{\sqrt{p}} \left\{ -z\|\mathbf{g}\| + \left\| \left(\sqrt{\gamma^2 + z^2}\, \mathbf{h} \odot \mathbf{y} + (\kappa - \gamma\|\boldsymbol{\mu}\|)\mathbf{1}_n \right)_+ \right\| \right\}.$$

In particular, we set $\widetilde{\xi}^{(1)}_{n,p,\kappa} \coloneqq \widetilde{\xi}^{(1)}_{n,p,\kappa}(\widetilde{\Theta}_\eta)$, where $\widetilde{\Theta}_\eta \coloneqq \{(\gamma, z) \in \mathbb{R}\times\mathbb{R}_+ :\ (1+\eta)^{-2}\gamma^2 + z^2 \leq 1\}$. We have the following relation between $\widetilde{\xi}^{(1)}_{n,p,\kappa}$ and $\xi^{(1)}_{n,p,\kappa}$.

Lemma C.3.6. *With the definition above, we have*

$$\left| \left(\widetilde{\xi}^{(1)}_{n,p,\kappa}\right)_+ - \xi^{(1)}_{n,p,\kappa} \right| \xrightarrow{p} 0.$$

See Appendix C.3.2 for the proof.

Step 3: Connecting $\widetilde{\xi}^{(1)}_{n,p,\kappa}$ with $\xi^{(2)}_{n,p,\kappa}$. It turns out that $\widetilde{\xi}^{(1)}_{n,p,\kappa}$ can be further simplified for analytical purposes. For any set $\mathbf{B} \subseteq \mathbb{R} \times \mathbb{R}_+$, we define a new quantity

$$\xi^{(2)}_{n,p,\kappa}(\mathbf{B}) = \min_{(\gamma,z)\in\mathbf{B}} \left\{ -z + \sqrt{\delta} \sqrt{\mathbb{E}\left(\sqrt{\gamma^2 + z^2} G + \kappa - \gamma\|\boldsymbol{\mu}\| \right)_+^2} \right\}.$$

Similar to before, we simply denote $\xi^{(2)}_{n,p,\kappa} \coloneqq \xi^{(2)}_{n,p,\kappa}(\widetilde{\Theta}_\eta)$. The two quantities of interest can be related via the uniform law of large numbers as is shown in the following lemma.

Lemma C.3.7. *With the definition of $\xi^{(2)}_{n,p,\kappa}$, we have*

$$\left| \xi^{(1)}_{n,p,\kappa} - \left(\xi^{(2)}_{n,p,\kappa}\right)_+ \right| \xrightarrow{p} 0.$$

See Appendix C.3.2 for the proof.

Following the chain of equivalence, we switch our goal to study $\xi^{(2)}_{n,p,\kappa}$.

Step 4: Analyzing $\xi^{(2)}_{n,p,\kappa}$. In view of the chain of arguments above, analyzing the margin boils down to analyzing the positivity of $\xi^{(2)}_{n,p,\kappa}$, which is equivalent to the positivity of the following function

$$F_\delta(\gamma, z, \kappa) \coloneqq -z^2 + \delta \mathbb{E}\left(\sqrt{\gamma^2 + z^2} G + \kappa - \gamma \|\boldsymbol{\mu}\|\right)_+^2. \tag{C.26}$$

To do so, the following lemma is useful in which $\mathbf{B}_v \coloneqq \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\mathbf{P}_{\boldsymbol{\mu}}^\perp \boldsymbol{\beta}\| \le v \|\mathbf{P}_{\boldsymbol{\mu}} \boldsymbol{\beta}\|\}$ is a convex cone for each choice of $v \ge 0$.

Lemma C.3.8. *Define the event*

$$\mathcal{E}_{n,p,\kappa,v} \coloneqq \left\{\text{there exists } \boldsymbol{\beta} \in \Theta \cap \mathbf{B}_v : y_i \langle \boldsymbol{h}_{0,i}, \boldsymbol{\beta} \rangle \ge \kappa \text{ for all } i \le n\right\}.$$

1. *If there exists $(\gamma, z) \in \widetilde{\Theta}_\eta \cap \{(\gamma, z) : z \le v\gamma\}$ such that $F_\delta(\gamma, z, \kappa) < 0$, then*

$$\mathbb{P}\left(\mathcal{E}_{n,p,\kappa,v}\right) = 1 - o(1).$$

2. *If for any $(\gamma, z) \in \widetilde{\Theta}_\eta \cap \{(\gamma, z) : z \le v\gamma\}$, we have $F_\delta(\gamma, z, \kappa) > 0$, then*

$$\mathbb{P}\left(\mathcal{E}^c_{n,p,\kappa,v}\right) = 1 - o(1).$$

See Appendix C.3.2 for the proof.

Remark C.3.9. This proposition still holds when $v = +\infty$. This will be useful for proving Theorem 4.5.2 (1).

Step 5: Analyzing $F_\delta(\gamma, z, \kappa)$ via an equivalent form. Lemma C.3.8 motivates us to focus on the positivity of function F_δ (C.26). To this end, we define a helpful function

$f_\delta : \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}$

$$f_\delta(u, \kappa, c) = -u^2 + \delta \mathbb{E}\left[\left(\sqrt{1+u^2}G + \kappa\sqrt{u^2+c} - \|\mu\|\right)_+^2\right].$$

Fix any $c > 0$, we consider the optimization problem

$$\begin{aligned} &\sup_{\kappa \geq 0} \quad \kappa \\ &\text{subject to} \quad \inf_{u \geq 0} \; f_\delta(u, \kappa, c) \leq 0. \end{aligned} \tag{C.27}$$

In fact, $c = \frac{1}{(1+\eta)^2}$ is just a reparametrization of η. Denote by $\kappa^* = \kappa^*(c)$ the corresponding maximizer and $u^* = u^*(c)$, the smallest minimizer of $\inf_{u \geq 0} \; f_\delta(u, \kappa^*(c), c)$. The following lemma demonstrates that both are well defined.

Lemma C.3.10. *For every $c > 0$, one has $0 \leq \kappa^*(c) < +\infty$, and $0 < u^*(c) < +\infty$. And they attain the corresponding maximum and minimum.*

See Appendix C.3.2 for the proof.

Both κ^* and u^* play an essential role in our later developments.

Proof of Theorem 4.5.2 (1).

We plan to invoke Lemma C.3.8 part 2 to prove this claim. More precisely, we aim to show that for any fixed $\kappa > 0$, when $\delta > \delta^*(\rho)$ with[1]

$$\delta^*(\rho) = \left(\inf_{u>0} \frac{1}{u^2} \mathbb{E}\left\{ \left(\sqrt{1+u^2} G - \|\boldsymbol{\mu}\| \right)_+^2 \right\} \right)^{-1},$$

one has $\inf_{(\gamma,z)\in\widetilde{\Theta}_\eta} F_\delta(\gamma, z, \kappa) > 0$.

Case 1: $\gamma = 0$**.** In this case, we have $F_\delta(0, z, \kappa) = -z^2 + \delta\mathbb{E}\left[(zG+\kappa)_+^2\right]$. If in addition $z = 0$, one has $F_\delta(0, 0, \kappa) > 0$. Otherwise for any $0 < z \le 1$, one has

$$F_\delta(0, z, \kappa) = z^2 \left(-1 + \delta\mathbb{E}\left[\left(G + \frac{\kappa}{z} \right)_+^2 \right] \right).$$

Note that since $z \le 1$, we have

$$\begin{aligned} -1 + \delta\mathbb{E}\left[\left(G + \frac{\kappa}{z} \right)_+^2 \right] &\ge -1 + \delta\mathbb{E}\left[(G+\kappa)_+^2 \right] \\ &> -1 + \delta\mathbb{E}\left[(G)_+^2 \right] \qquad \text{since } \kappa > 0 \\ &> -1 + \delta^*(\rho)\frac{1}{2} \ge 0, \end{aligned}$$

where the last relation uses the fact that $\delta^*(\rho) \ge 2$.

1. To see that $\delta^*(\rho)$ is well-defined, let $\alpha = 1/u$. Then we need to check if

$$\inf_{\alpha\ge 0} \mathbb{E}\left\{ \left(\sqrt{1+\alpha^2} G - \|\boldsymbol{\mu}\|\alpha \right)_+^2 \right\} > 0.$$

At $\alpha = 0$, $\mathbb{E}[(G)_+^2] > 0$ indicates that the infimum is strictly positive in the neighborhood of 0; for $\alpha \ge C_\alpha$, $\inf_{\alpha\ge 0} \mathbb{E}\{\left(\sqrt{1+\alpha^2} G - \|\boldsymbol{\mu}\|\alpha\right)_+^2\} \ge \inf_{\alpha\ge 0} \alpha^2 \mathbb{E}\{(G - \|\boldsymbol{\mu}\|)_+^2\} \ge C_\alpha^2 \mathbb{E}\{(G - \|\boldsymbol{\mu}\|)_+^2\} > 0$. Thus, $\delta^*(\rho) < +\infty$ is well-defined. As a final remark, when $\|\boldsymbol{\mu}\| = 0$ (i.e., the pure noise case), we have $\delta^*(0) = 2$. As a result, $\delta^*(\rho) \ge 2$ for $\rho \ge 0$.

Case 2: $\gamma > 0$**.** In this case, we have

$$F_\delta(\gamma, z, \kappa) = \gamma^2 \left\{ -u^2 + \delta \mathbb{E}\left[\left(\sqrt{1+u^2} G + \frac{\kappa}{\gamma} - \|\mu\| \right)_+^2 \right] \right\},$$

where we denote $u = z/\gamma$. It is easy to see that when $u = 0$, one has $F_\delta(\gamma, z, \kappa) > 0$. Therefore we focus on the case when $u > 0$.

By the definition of $\delta^*(\rho)$, we know that for any $u > 0$,

$$\frac{1}{\delta} < \frac{1}{\delta^*(\rho)} \leq \frac{1}{u^2} \mathbb{E}\left[\left(\sqrt{1+u^2} G - \|\mu\| \right)_+^2 \right].$$

As a result, for any $u > 0$, one has

$$-u^2 + \delta \mathbb{E}\left[\left(\sqrt{1+u^2} G - \|\mu\| \right)_+^2 \right] > 0,$$

which further implies

$$-u^2 + \delta \mathbb{E}\left[\left(\sqrt{1+u^2} G + \frac{\kappa}{\gamma} - \|\mu\| \right)_+^2 \right] > 0.$$

This proves that $F_\delta(\gamma, z, \kappa) > 0$ for all γ, z.

Proof of Theorem 4.5.2 (2).

In view of Part 1 of Lemma C.3.8, we only need to show that for some $\kappa > 0$, there exist some (γ, z) such that $F_\delta(\gamma, z, \kappa) < 0$. Since $\delta < \delta^*(\rho)$, we have

$$\frac{1}{\delta} > \frac{1}{u^2} \mathbb{E}\left[\left(\sqrt{1+u^2} G - \|\mu\| \right)_+^2 \right]$$

for some $u > 0$. As a result, we have

$$-u^2 + \delta \mathbb{E}\left[\left(\sqrt{1+u^2}G - \|\mu\|\right)_+^2\right] < 0.$$

Due to continuity, we know that there exists some $L > 0$, such that

$$-u^2 + \delta \mathbb{E}\left[\left(\sqrt{1+u^2}G + \kappa\sqrt{u^2+c} - \|\mu\|\right)_+^2\right] < 0, \qquad \text{for all } \kappa < L.$$

Let $\gamma = \frac{1}{\sqrt{u^2+c}}$, and $z = u \cdot \gamma$. We then have

$$F_\delta(\gamma, z, \kappa) = \gamma^2 \left\{-u^2 + \delta \mathbb{E}\left[\left(\sqrt{1+u^2}G + \frac{L}{2}\sqrt{u^2+c} - \|\mu\|\right)_+^2\right]\right\} < 0.$$

This finishes the proof.

We then turn to the claim regarding the margin. Recall that we work in the regime where $\delta < \delta^*(\rho)$. Fix any $\varepsilon > 0$. We aim to prove that $\mathcal{E}_{n,p,\kappa^*+\varepsilon}$ does not hold and that $\mathcal{E}_{n,p,\kappa^*-\varepsilon}$ holds.

Step 1: Quantifying $\mathcal{E}_{n,p,\kappa^*+\varepsilon}$. Similar to the proof of Theorem 4.5.2 (1), it suffices to prove that for all γ, z, $F_\delta(\gamma, z, \kappa^* + \varepsilon) > 0$.

Case 1: $\gamma = 0$. It is simple to see that $F_\delta(0, 0, \kappa^* + \varepsilon) > 0$. Therefore we focus on the case for any $0 < z \leq 1$. Using similar arguments as in Theorem 4.5.2 (1), we have

$$F_\delta(0, z, \kappa^* + \varepsilon) = z^2 \left(-1 + \delta \mathbb{E}\left[\left(G + \frac{\kappa^* + \varepsilon}{z}\right)_+^2\right]\right).$$

Note that since $z \leq 1$, we have

$$-1 + \delta \mathbb{E}\left[\left(G + \frac{\kappa^* + \varepsilon}{z}\right)_+^2\right] \geq -1 + \delta \mathbb{E}\left[(G + \kappa^*)_+^2\right].$$

When $2 \le \delta < \delta^*(\rho)$, we have $\kappa^* \ge 0$, and hence $F_\delta(0, z, \kappa^* + \varepsilon) > 0$. When $\delta < 2$, since $\kappa^* \ge \kappa(\delta)$, where $\kappa(\delta)$ is the unique solution to

$$\frac{1}{\delta} = \mathbb{E}\left[(G + \kappa)_+^2\right],$$

we again have $F_\delta(0, z, \kappa^* + \varepsilon) > 0$.

Case 2: $\gamma > 0$**.** In this case, we have

$$F_\delta(\gamma, z, \kappa^* + \varepsilon) = \gamma^2 \left\{ -u^2 + \delta \mathbb{E}\left[\left(\sqrt{1+u^2} G + \frac{\kappa^* + \varepsilon}{\gamma} - \|\mu\| \right)_+^2 \right] \right\},$$

where we denote $u = z/\gamma$. By the definition of κ^*, we know that for all $u \ge 0$,

$$f_\delta(u, \kappa^* + \varepsilon, c) = -u^2 + \delta \mathbb{E}\left[\left(\sqrt{1+u^2} G + (\kappa^* + \varepsilon)\sqrt{u^2 + c} - \|\mu\| \right)_+^2 \right] > 0.$$

For all $0 < \gamma \le \frac{1}{\sqrt{u^2+c}}$, we further have

$$-u^2 + \delta \mathbb{E}\left[\left(\sqrt{1+u^2} G + \frac{\kappa^* + \varepsilon}{\gamma} - \|\mu\| \right)_+^2 \right] > -u^2 + \delta \mathbb{E}\left[\left(\sqrt{1+u^2} G + (\kappa^* + \varepsilon)\sqrt{u^2 + c} - \|\mu\| \right)_+^2 \right] > 0.$$

This finishes the claim.

Step 2: Quantifying $\mathcal{E}_{n,p,\kappa^*-\varepsilon}$**.** Similarly, for $\kappa^* - \varepsilon$, since $f_\delta(u, \kappa, c)$ is increasing in κ and $\inf_{u \ge 0} f_\delta(u, \kappa^*, c) = 0$, there exists some $u > 0$, such that

$$-u^2 + \delta \mathbb{E}\left[\left(\sqrt{1+u^2} G + (\kappa^* - \varepsilon)\sqrt{u^2 + c} - \|\mu\| \right)_+^2 \right] < 0.$$

Going through the same argument as in part 1 completes the proof of this part.

Combine the arguments above, for any $\varepsilon > 0$, we see that with probability approaching

1, $\kappa^* - \varepsilon < \widehat{\kappa} < \kappa^* + \varepsilon$, which proves that $\widehat{\kappa} \xrightarrow{p} \kappa^*$.

Proof of Theorem 4.5.2 (3).

We start with Lemma C.3.11 showing that $u^* = u^*(c) = u^*(\rho, \eta)$, defined previously, is indeed the limit of $\widehat{u} := \dfrac{\|\mathbf{P}_{\boldsymbol{\mu}}^{\perp}\widehat{\boldsymbol{\beta}}\|}{\|\mathbf{P}_{\boldsymbol{\mu}}\widehat{\boldsymbol{\beta}}\|}$.

Lemma C.3.11. *Assume that $\delta < \delta^*(\rho)$. Fix any $c > 0$ and any $\varepsilon > 0$,*

$$\mathcal{E}_{n,p,\kappa^*,u^*+\varepsilon} \qquad \textit{and} \qquad \mathcal{E}^c_{n,p,\kappa^*,u^*-\varepsilon} \qquad \textit{hold.}$$

As a result, we have

$$\widehat{u} := \frac{\|\mathbf{P}_{\boldsymbol{\mu}}^{\perp}\widehat{\boldsymbol{\beta}}\|}{\|\mathbf{P}_{\boldsymbol{\mu}}\widehat{\boldsymbol{\beta}}\|} \xrightarrow{p} u^*.$$

See Appendix C.3.2 for the proof.

Based on the lemma, recall the definition $\rho = \|\boldsymbol{\mu}\|^2$, $u^*(\rho, \eta) = u^*(c)$, and

$$\texttt{Err}(\widehat{\boldsymbol{\beta}}; \eta) = \mathbb{P}\left(y'\langle \boldsymbol{h}_0', \widehat{\boldsymbol{\beta}}\rangle < 0\right) = \Phi\left(-\langle \boldsymbol{\mu}, \frac{\widehat{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|}\rangle\right), \qquad \texttt{Err}^*(\eta) = \Phi\left(-\|\boldsymbol{\mu}\|\sqrt{\frac{1}{1+u^{*2}(\rho,\eta)}}\right).$$

From Lemma C.3.11, we have

$$\left|\langle \boldsymbol{\mu}, \frac{\widehat{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|}\rangle - \|\boldsymbol{\mu}\|\sqrt{\frac{1}{1+u^{*2}(\rho,\eta)}}\right| = o_{\mathbb{P}}(1),$$

which implies that

$$\texttt{Err}(\widehat{\boldsymbol{\beta}}; \eta) \xrightarrow{p} \texttt{Err}^*(\eta) = \Phi\left(-\|\boldsymbol{\mu}\|\sqrt{\frac{1}{1+u^{*2}(\rho,\eta)}}\right).$$

Since $\texttt{Err}^*(\eta)$ is monotonically decreasing in $u^*(\rho,\eta)$, then asymptotically, the classification error of the original max-margin solution is monotonically decreasing in $u^*(\rho,\eta)$. It remains to show that $u^*(\rho,\eta)$ is monotonically decreasing in η, which is equivalent to showing that $u^*(c)$ is monotonically increasing in c. This is provided in the following lemma.

Lemma C.3.12. *With the definition of (u^*,κ^*,c), for any $c>0$,*

1. *$u^*=u^*(c)$ is a continuous function in c;*
2. *$u^*=u^*(c)$ is monotonically increasing in c.*

See Appendix C.3.2 for the proof.

These finish the proof of Theorem 4.5.2.

Proof of Lemma C.3.6

Recall that

$$\xi^{(1)}_{n,p,\kappa}(\mathbf{B}_p)=\min_{\boldsymbol{\beta}\in\mathbf{B}_p}\max_{\boldsymbol{\lambda}:\|\boldsymbol{\lambda}\|\le 1,\mathbf{y}\odot\boldsymbol{\lambda}\ge 0}\frac{1}{\sqrt{p}}\left\{\|\boldsymbol{\lambda}\|\boldsymbol{g}^\top\boldsymbol{\beta}+\|\boldsymbol{\beta}\|\mathbf{h}^\top\boldsymbol{\lambda}+\kappa\boldsymbol{\lambda}^\top\mathbf{y}-\langle\boldsymbol{\mu},\boldsymbol{\beta}\rangle\boldsymbol{\lambda}^\top\mathbf{y}\right\}.$$

Denote

$$\ell(r_{\boldsymbol{\lambda}},\boldsymbol{\lambda}_0)\coloneqq\frac{1}{\sqrt{p}}\left\{\|\boldsymbol{\lambda}\|\boldsymbol{g}^\top\boldsymbol{\beta}+\|\boldsymbol{\beta}\|\mathbf{h}^\top\boldsymbol{\lambda}+\kappa\boldsymbol{\lambda}^\top\mathbf{y}-\langle\boldsymbol{\mu},\boldsymbol{\beta}\rangle\boldsymbol{\lambda}^\top\mathbf{y}\right\}.$$

Step 1: inner maximization. We focus on the inner maximization problem, which is equivalent to

$$\max_{r_{\boldsymbol{\lambda}}\in[0,1],\|\boldsymbol{\lambda}_0\|=1,\mathbf{y}\odot\boldsymbol{\lambda}_0\ge 0}\quad \ell(r_{\boldsymbol{\lambda}},\boldsymbol{\lambda}_0)=\frac{1}{\sqrt{p}}r_{\boldsymbol{\lambda}}\mathbf{g}^\top\boldsymbol{\beta}+\frac{1}{\sqrt{p}}r_{\boldsymbol{\lambda}}\left[\|\boldsymbol{\beta}\|\mathbf{h}\odot\mathbf{y}+(\kappa-\gamma\|\boldsymbol{\mu}\|)\mathbf{1}_n\right]^\top(\mathbf{y}\odot\boldsymbol{\lambda}_0),$$

where $\gamma=\langle\boldsymbol{\mu},\boldsymbol{\beta}\rangle/\|\boldsymbol{\mu}\|$.

When $r_{\boldsymbol{\lambda}} = 0$, we have $\ell(r_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}_0) = 0$. When $r_{\boldsymbol{\lambda}} > 0$, since $\|\mathbf{y} \odot \boldsymbol{\lambda}_0\| = \|\boldsymbol{\lambda}_0\| = 1$ and $\mathbf{y} \odot \boldsymbol{\lambda}_0 \geq 0$, we have for any vector $\boldsymbol{\nu}$ with at least one nonnegative coordinate,

$$\max_{\|\boldsymbol{\lambda}_0\|=1, \mathbf{y}\odot\boldsymbol{\lambda}_0 \geq 0} \boldsymbol{\nu}^\top (\mathbf{y} \odot \boldsymbol{\lambda}_0) = \|\boldsymbol{\nu}_+\|. \tag{C.28}$$

In fact, letting $\boldsymbol{\nu} = \|\boldsymbol{\beta}\|\mathbf{h} \odot \mathbf{y} + (\kappa - \gamma\|\boldsymbol{\mu}\|)\mathbf{1}_n$, we see that $\boldsymbol{\nu}$ has i.i.d. coordinates, and for each coordinate, the probability for it to be negative is given by

$$\mathbb{P}\left(\|\boldsymbol{\beta}\| g + \kappa - \gamma\|\boldsymbol{\mu}\| < 0\right) = \Phi\left(-\frac{\kappa}{\|\boldsymbol{\beta}\|} + \frac{\gamma\|\boldsymbol{\mu}\|}{\|\boldsymbol{\beta}\|}\right) \leq \Phi\left(\frac{\gamma\|\boldsymbol{\mu}\|}{\|\boldsymbol{\beta}\|}\right) \leq \Phi(\|\boldsymbol{\mu}\|).$$

As a result, with probability at least $1 - \Phi(\|\boldsymbol{\mu}\|)^n$, $\boldsymbol{\nu}$ has at least one nonnegative entry, and hence (C.28) holds.

Combine the two cases to arrive at the conclusion that: with probability at least $1 - \Phi(\|\boldsymbol{\mu}\|)^n$,

$$\max_{r_{\boldsymbol{\lambda}} \in [0,1], \|\boldsymbol{\lambda}_0\|=1, \mathbf{y}\odot\boldsymbol{\lambda}_0 \geq 0} \quad \ell(r_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}_0) = \left(\frac{1}{\sqrt{p}} \left\{ \boldsymbol{g}^\top \boldsymbol{\beta} + \left\| \left(\|\boldsymbol{\beta}\|\mathbf{h} \odot \mathbf{y} + (\kappa - \gamma\|\boldsymbol{\mu}\|)\mathbf{1}_n\right)_+ \right\| \right\}\right)_+,$$

where we recall that $\gamma = \langle \boldsymbol{\mu}, \boldsymbol{\beta} \rangle / \|\boldsymbol{\mu}\|$.

Step 2: outer minimization. Now we turn to the outer minimization over $\boldsymbol{\beta} \in \Theta = \{\boldsymbol{\beta} \in \mathbb{R}^p : \boldsymbol{\beta}^\top \boldsymbol{\Omega} \boldsymbol{\beta} \leq 1\}$. Using a reparametrization $\boldsymbol{\beta} = \gamma \boldsymbol{\mu}_0 + \boldsymbol{\beta}_\perp$, where $\boldsymbol{\beta}_\perp = r \cdot \boldsymbol{w}$ with $z = \|\boldsymbol{\beta}_\perp\|$ and $\boldsymbol{w} \perp \boldsymbol{\mu}$, $\|\boldsymbol{w}\| = 1$, we see that with high probability $\xi^{(1)}_{n,p,\kappa}$ is equal to

$$\bar{\xi}^{(1)}_{n,p,\kappa} = \frac{1}{\sqrt{p}} \min_{(\gamma,z)\in\widetilde{\Theta}_\eta} \min_{\boldsymbol{w}\perp\boldsymbol{\mu}, \|\boldsymbol{w}\|=1} \left\{ \gamma g^\top \boldsymbol{\mu}_0 + z\boldsymbol{g}^\top \boldsymbol{w} + \left\| \left(\sqrt{\gamma^2 + z^2}\mathbf{h} \odot \mathbf{y} + (\kappa - \gamma\|\boldsymbol{\mu}\|)\mathbf{1}_n \right)_+ \right\| \right\}_+.$$

Here we recall that $\widetilde{\Theta}_\eta = \{(\gamma, z) \in \mathbb{R} \times \mathbb{R}_+ : \ (1+\eta)^{-2}\gamma^2 + z^2 \leq 1\}$.

We can solve the inner minimization, that is

$$\begin{aligned}&\min_{\boldsymbol{w}\perp\boldsymbol{\mu},\|\boldsymbol{w}\|=1}\left\{\gamma g^\top\boldsymbol{\mu}_0+z\boldsymbol{g}^\top\boldsymbol{w}+\left\|\left(\sqrt{\gamma^2+z^2}\mathbf{h}\odot\mathbf{y}+(\kappa-\gamma\|\boldsymbol{\mu}\|)\mathbf{1}_n\right)_+\right\|\right\}_+\\&=\left\{\gamma g^\top\boldsymbol{\mu}_0-z\|\mathbf{P}_{\boldsymbol{\mu}}^\perp\boldsymbol{g}\|+\left\|\left(\sqrt{\gamma^2+z^2}\mathbf{h}\odot\mathbf{y}+(\kappa-\gamma\|\boldsymbol{\mu}\|)\mathbf{1}_n\right)_+\right\|\right\}_+\end{aligned}$$

Note that $\gamma g^\top\boldsymbol{\mu}_0/\sqrt{p}=o_{\mathbb{P}}(1)$. This together with the boundedness of $\widetilde{\Theta}_\eta$ implies that

$$\bar{\xi}_{n,p,\kappa}^{(1)}-\min_{(\gamma,z)\in\widetilde{\Theta}_\eta}\frac{1}{\sqrt{p}}\left\{-r\|\mathbf{g}\|+\left\|\left(\sqrt{\gamma^2+r^2}\,\mathbf{h}\odot\mathbf{y}+(\kappa-\gamma\|\boldsymbol{\mu}\|)\mathbf{1}_n\right)_+\right\|\right\}_+=o_{\mathbb{P}}(1).$$

Note that $\min_{(\gamma,r)\in\widetilde{\Theta}_\eta}\frac{1}{\sqrt{p}}\left\{-z\|\mathbf{g}\|+\left\|\left(\sqrt{\gamma^2+z^2}\,\mathbf{h}\odot\mathbf{y}+(\kappa-\gamma\|\boldsymbol{\mu}\|)\mathbf{1}_n\right)_+\right\|\right\}_+=\left(\widetilde{\xi}_{n,p,\kappa}^{(1)}\right)_+$.

Combine all the pieces together to complete the proof.

Proof of Lemma C.3.7

Define two functions

$$\begin{aligned}\phi_{n,\boldsymbol{\mu},\delta,\kappa}^{(1)}(\gamma,z)&=\left[\mathbb{E}\left(\sqrt{\gamma^2+z^2}G+\kappa-\gamma\|\boldsymbol{\mu}\|\right)_+^2\right]^{1/2},\\\phi_{n,\boldsymbol{\mu},\delta,\kappa}^{(2)}(\gamma,z;\mathbf{h})&=\frac{1}{\sqrt{n}}\left\|\left(\sqrt{\gamma^2+z^2}\widetilde{\mathbf{h}}+(\kappa-\gamma\|\boldsymbol{\mu}\|)\mathbf{1}_n\right)_+\right\|,\end{aligned}$$

where $\widetilde{\boldsymbol{h}}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_n)$ and $G\sim\mathcal{N}(0,1)$. By the uniform law of large numbers (Jennrich, 1969) (see also (Montanari et al., 2019) (Lemma 6.2)), we have

$$\sup_{(\gamma,z)\in\widetilde{\Theta}_\eta}\left|\phi_{n,\boldsymbol{\mu},\delta,\kappa}^{(1)}(\gamma,z)-\phi_{n,\boldsymbol{\mu},\delta,\kappa}^{(2)}(\gamma,z;\boldsymbol{h})\right|=o_{\mathbb{P}}(1),$$

In addition, since $p^{-1/2}\|\mathbf{g}\| = 1 + o_{\mathbb{P}}(1)$, we have

$$\left|\widetilde{\xi}^{(1)}_{n,p,\kappa} - \xi^{(2)}_{n,p,\kappa}\right| \xrightarrow{p} 0,$$

which implies the desired claim.

Proof of Lemma C.3.8

Step 1: establish equivalence. We notice that for the convex set $\Theta \cap \mathbf{B}_v$,

$$\mathcal{E}_{n,p,\kappa,v} = \mathcal{E}_{n,p,\kappa}(\Theta \cap \mathbf{B}_v).$$

Denote the quantity

$$\xi_{n,p,\kappa,v} = \xi_{n,p,\kappa}(\mathbf{B}_v \cap \Theta) = \min_{\boldsymbol{\beta}\in\mathbf{B}_v\cap\Theta} \max_{\|\boldsymbol{\lambda}\|\leq 1, \mathbf{y}\odot\boldsymbol{\lambda}\geq 0} \frac{1}{\sqrt{p}}\boldsymbol{\lambda}^\top(\kappa\mathbf{y} - \mathbf{H}_0\boldsymbol{\beta}),$$

for which we have

$$\{\xi_{n,p,\kappa,v} > 0\} \Longleftrightarrow \mathcal{E}^c_{n,p,\kappa,v} \quad \text{and} \quad \{\xi_{n,p,\kappa,v} = 0\} \Longleftrightarrow \mathcal{E}_{n,p,\kappa,v}.$$

Following the equivalence established in Lemma C.3.5, C.3.6 and C.3.7, denote

$$\begin{aligned}
\xi^{(1)}_{n,p,\kappa,v} &= \xi^{(1)}_{n,p,\kappa}(\mathbf{B}_v \cap \Theta) \\
\xi^{(2)}_{n,p,\kappa,v} &= \xi^{(2)}_{n,p,\kappa}(\widetilde{\Theta}_\eta \cap \{(\gamma, z) : z \leq v\gamma\}).
\end{aligned}$$

Then, it is directly implied by Lemma C.3.5 that for any $t \in \mathbb{R}$,

$$\mathbb{P}(\xi_{n,p,\kappa,v} \leq t) \leq 2\mathbb{P}(\xi^{(1)}_{n,p,\kappa,v} \leq t), \qquad \mathbb{P}(\xi_{n,p,\kappa,v} \geq t) \leq 2\mathbb{P}(\xi^{(1)}_{n,p,\kappa,v} \geq t). \tag{C.29}$$

By Lemma C.3.6 and C.3.7, we further have

$$\left|\xi^{(1)}_{n,p,\kappa,v} - \left(\xi^{(2)}_{n,p,\kappa,v}\right)_+\right| \xrightarrow{p} 0. \tag{C.30}$$

Step 2: analyze $F_\delta(\gamma, z, \kappa)$. We consider the following two cases.

1. By the definition of $F_\delta(\gamma, z, \kappa)$, if there exists $(\gamma, z) \in \widetilde{\Theta} \cap \{(\gamma, z) : z \le v\gamma\}$ such that $F_\delta(\gamma, z, \kappa) < 0$, then

$$\xi^{(2)}_{n,p,\kappa,v} = \min_{(\gamma,z)\in\widetilde{\Theta}\cap\{(\gamma,z):z\le v\gamma\}} \left\{-z + \sqrt{\delta}\sqrt{\mathbb{E}\left(\sqrt{\gamma^2+z^2}G + \kappa - \gamma\|\boldsymbol{\mu}\|\right)_+^2}\right\} < 0,$$

which, according to (C.30), implies that $\mathbb{P}(\xi^{(1)}_{n,p,\kappa,v} > 0) = o(1)$ and for any $\varepsilon > 0$,

$$\mathbb{P}(\xi^{(1)}_{n,p,\kappa,v} \ge \varepsilon) = o(1).$$

As $\mathbb{P}(\xi_{n,p,\kappa,v} > t)$ is right-continuous in t, then

$$\mathbb{P}(\xi_{n,p,\kappa,v} > 0) = \lim_{t\to 0+} \mathbb{P}(\xi_{n,p,\kappa,v} > t).$$

Combining (C.29) with the definition of $\mathcal{E}^c_{n,p,\kappa,v}$, we have

$$\mathbb{P}(\mathcal{E}^c_{n,p,\kappa,v}) = \mathbb{P}(\xi_{n,p,\kappa,v} > 0) = \lim_{\varepsilon\to 0+} \mathbb{P}(\xi_{n,p,\kappa,v} > \varepsilon) \le \lim_{\varepsilon\to 0+} \mathbb{P}(\xi_{n,p,\kappa,v} \ge \varepsilon) \le 2 \lim_{\varepsilon\to 0+} \mathbb{P}(\xi^{(1)}_{n,p,\kappa,v} \ge \varepsilon).$$

In addition, we verify the validity of changing the order of $\lim_{n\to+\infty, n/p\to\delta}$ and $\lim_{\varepsilon\to 0+}$. Let $q_{n,p}(t) = \mathbb{P}(\xi^{(1)}_{n,p,\kappa,v} \ge t)$ and $\widetilde{q}_{n,p}(t) = \lim_{s\to t+} q_{n,p}(s) = \mathbb{P}(\xi^{(1)}_{n,p,\kappa,v} > t)$, which satisfies that $q_{n,p}(t)$ is left-continuous, $\widetilde{q}_{n,p}(t)$ is right-continuous, and for any $t > 0$

$$\lim_{n\to+\infty, n/p\to\delta} q_{n,p}(t) = \lim_{n\to+\infty, n/p\to\delta} \widetilde{q}_{n,p}(t) = 0.$$

Further, we have

$$\lim_{n\to+\infty,n/p\to\delta}\widetilde{q}_{n,p}(0)=\lim_{n\to+\infty,n/p\to\delta}\mathbb{P}(\xi^{(1)}_{n,p,\kappa,v}>0)=0.$$

By the fact that $\widetilde{q}_{n,p}(t)$ in non-increasing in t, we have

$$\lim_{n\to+\infty,n/p\to\delta}\sup_{t\geq 0}|\widetilde{q}_{n,p}(t)-0|,$$

which indicates that $\widetilde{q}_{n,p}(t)$ uniformly converges to 0. Therefore, by Moore-Osgood theorem,

$$\lim_{n\to+\infty,n/p\to\delta}\lim_{\varepsilon\to 0+}q_{n,p}(\varepsilon)=\lim_{n\to+\infty,n/p\to\delta}\lim_{\varepsilon\to 0}\widetilde{q}_{n,p}(\varepsilon)=\lim_{\varepsilon\to 0}\lim_{n\to+\infty,n/p\to\delta}\widetilde{q}_{n,p}(\varepsilon)=0,$$

thus, we have

$$\lim_{n\to+\infty,n/p\to\delta}\lim_{\varepsilon\to 0+}\mathbb{P}(\xi^{(1)}_{n,p,\kappa,v}\geq\varepsilon)=0\qquad\Longrightarrow\qquad\lim_{n\to+\infty,n/p\to\delta}\mathbb{P}(\mathcal{E}^{c}_{n,p,\kappa,v})=0.$$

2. If, instead, for any $(\gamma,z)\in\widetilde{\Theta}\cap\{(\gamma,z):z\leq v\gamma\}$, we have $F_\delta(\gamma,z,\kappa)>0$. Since the set $(\gamma,z)\in\widetilde{\Theta}\cap\{(\gamma,z):z\leq v\gamma\}$ is convex and compact, then

$$\xi^{(2)}_{n,p,\kappa,v}=\min_{(\gamma,z)\in\widetilde{\Theta}\cap\{(\gamma,z):z\leq v\gamma\}}\left\{-z+\sqrt{\delta}\sqrt{\mathbb{E}\left(\sqrt{\gamma^2+z^2}G+\kappa-\gamma\|\boldsymbol{\mu}\|\right)_+^2}\right\}>0,$$

which implies that

$$\mathbb{P}(\mathcal{E}_{n,p,\kappa,v})=\mathbb{P}(\xi_{n,p,\kappa,v}=0)\leq 2\mathbb{P}(\xi^{(1)}_{n,p,\kappa,v}\leq 0)=o(1).$$

Proof of Lemma C.3.10

In this section, we will first show the relationship between $F_\delta(\gamma, z, \kappa)$ and $f_\delta(u, \kappa, c)$. In the next step, we show the boundedness of $\kappa^*(c)$ and $u^*(c)$.

Step 1: Relationship between $F_\delta(\gamma, z, \kappa)$ and $f_\delta(u, \kappa, c)$. Consider the properties of $F_\delta(\gamma, z, \kappa)$. Since $F_\delta(\gamma, z, \kappa) < F_\delta(-\gamma, z, \kappa)$ for $\gamma > 0$, in the following we only consider $\gamma \in \mathbb{R}_+$. Denote $u = z/\gamma$, then $F_\delta(\gamma, z, \kappa)$ can be written as a function of (γ, u, κ):

$$\widetilde{F}_\delta(\gamma, u, \kappa) = \gamma^2 \left(-u^2 + \delta \mathbb{E} \left\{ \left(\sqrt{1+u^2} G + \frac{\kappa}{\gamma} - \|\boldsymbol{\mu}\| \right)_+^2 \right\} \right),$$

where $(\gamma, u) \in \widetilde{\Omega} = \{(\gamma, u) \in \mathbb{R}_+ \times \bar{\mathbb{R}}_+ : \gamma^2 \left(u^2 + c\right) \leq 1\}$ with $c = (1+\eta)^{-2}$ and $\bar{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{\infty\}$. We have the following problem

$$\begin{cases} \sup & \kappa \\ \text{subject to} & \exists (\gamma, u) \in \widetilde{\Omega},\ \widetilde{F}_\delta(\gamma, u, \kappa) < 0. \end{cases} \tag{C.31}$$

Note that when κ approaches the supremum, for any $(\gamma', u) \notin \partial\widetilde{\Omega}$ such that $\widetilde{F}_\delta(\gamma', u, \kappa) < 0$, there exists $(\gamma, u) \in \partial\widetilde{\Omega}$ with $\gamma > \gamma'$ at which $\widetilde{F}_\delta(\gamma, u, \kappa) < \widetilde{F}_\delta(\gamma', u, \kappa)$. The reason is that when $\gamma^2 > (\gamma')^2$ and $\gamma^{-2}\widetilde{F}_\delta(\gamma, u, \kappa) < (\gamma')^{-2}\widetilde{F}_\delta(\gamma', u, \kappa) < 0$, we have $\widetilde{F}_\delta(\gamma, u, \kappa) < \widetilde{F}_\delta(\gamma', u, \kappa)$. Then, the above shows that the minimizer (γ, u) must be located on the boundary $\partial\widetilde{\Omega}$, that is $\gamma^{-2} = u^2 + c$. If we also view c as a variable, recall the definition $f_\delta(u, \kappa, c)$ with

$$\begin{aligned} g_\delta(u, \kappa, c) &= \widetilde{F}_\delta((u^2+c)^{-1/2}, u, \kappa) \\ &= \frac{1}{u^2+c} \cdot \left\{ -u^2 + \delta \mathbb{E} \left\{ \left(\sqrt{1+u^2} G + \kappa \sqrt{u^2+c} - \|\boldsymbol{\mu}\| \right)_+^2 \right\} \right\} \end{aligned}$$

and the optimization problem in (C.27)

$$\begin{aligned} \max \quad & \kappa \\ \text{subject to} \quad & \inf_{u \geq 0} \ f_\delta(u, \kappa, c) \leq 0, \end{aligned}$$

where $\kappa^* = \kappa^*(c)$ is the corresponding maximizer and $u^* = u^*(c)$, short for $u^*(\rho, \eta)$, is the smallest minimizer of $\min_{u \geq 0} \ f_\delta(u, \kappa, c)$.

We have the following lemma.

Lemma C.3.13. *If $\widetilde{\kappa}$ is the maximizer to* (C.31) *and $(\widetilde{u}, \widetilde{\gamma})$ is the minimizer to $F_\delta(\gamma, u\gamma, \widetilde{\kappa})$ such that $\widetilde{u}$ is the smallest minimizer for u, then*

$$\widetilde{\kappa} = \kappa^*, \quad \widetilde{u} = u^*, \quad \widetilde{\gamma} = \frac{1}{\sqrt{u^{*2} + c}}.$$

Proof. Fix $c \in \mathbb{R}_+$ and for any $\kappa \in \mathbb{R}_+$, we obtain from the reasoning above that for any (γ, u) such that $\widetilde{F}_\delta(\gamma, u, \kappa) \leq 0$,

$$\widetilde{F}_\delta(\gamma, u, \kappa) \geq \widetilde{F}_\delta((u^2 + c)^{-1/2}, u, \kappa) = g_\delta(u, \kappa, c).$$

As a result, when $\inf_{(\gamma,u) \in \widetilde{\Omega}} \widetilde{F}_\delta(\gamma, u, \kappa) \leq 0$,

$$\inf_{(\gamma,u) \in \widetilde{\Omega}} \widetilde{F}_\delta(\gamma, u, \kappa) \geq \inf_{u \geq 0} g_\delta(u, \kappa, c).$$

In addition, for the other direction,

$$\inf_{(\gamma,u) \in \widetilde{\Omega}} \widetilde{F}_\delta(\gamma, u, \kappa) \leq \inf_{(\gamma,u) \in \widetilde{\Omega}, \gamma = (u^2+c)^{-2}} \widetilde{F}_\delta(\gamma, u, \kappa) = \inf_{u \geq 0} g_\delta(u, \kappa, c).$$

Combining the two inequalities, we have

$$\inf_{(\gamma,u)\in\widetilde{\Omega}} \widetilde{F}_\delta(\gamma,u,\kappa) = \inf_{u\geq 0} g_\delta(u,\kappa,c),$$

which implies that the following feasible sets are the same

$$\left\{\kappa>0: \inf_{(\gamma,u)\in\widetilde{\Omega}} \widetilde{F}_\delta(\gamma,u,\kappa)<0\right\} = \left\{\kappa>0: \inf_{u\geq 0} g_\delta(u,\kappa,c)<0\right\}.$$

Then, it is equivalent to show that (C.27) has the same solution with the following problem

$$\begin{cases} \sup & \kappa \\ \text{subject to} & \inf_{u\geq 0} g_\delta(u,\kappa,c)<0. \end{cases}$$

We denote $\widetilde{u}$ and $\widetilde{\kappa}$ as the solution to (C.3.2) such that $\widetilde{u}$ is the smallest one if the minimizer is not unique. Since $\bar{\mathbb{R}}_+$ is a closed set, then $\inf_{u\geq 0} g_\delta(u,\kappa,c) = \min_{u\geq 0} g_\delta(u,\kappa,c)$. We further have

$$\left\{\kappa>0: \inf_{u\geq 0} g_\delta(u,\kappa,c)<0\right\} \subseteq \left\{\kappa>0: \inf_{u\geq 0} g_\delta(u,\kappa,c)\leq 0\right\},$$

indicating that

$$\kappa^* = \sup\left\{\kappa>0: \inf_{u\geq 0} g_\delta(u,\kappa,c)\leq 0\right\} \geq \sup\left\{\kappa>0: \inf_{u\geq 0} g_\delta(u,\kappa,c)<0\right\} = \widetilde{\kappa}.$$

Recall that the solution (u^*,κ^*) for (C.27) has the property that u^* is the smallest minimizer. Since $g_\delta(u,\kappa,c)$ is strictly increasing in κ, then for any $\varepsilon>0$, there exists $\kappa_\varepsilon \geq \kappa^*-\varepsilon$ such that

$$\inf_{u\geq 0} g_\delta(u,\kappa_\varepsilon,c) < \inf_{u\geq 0} g_\delta(u,\kappa^*,c) \leq 0,$$

which means

$$\kappa_\varepsilon \in \left\{ \kappa > 0 : \inf_{u \geq 0} g_\delta(u, \kappa, c) < 0 \right\}. \tag{C.32}$$

If $\widetilde{\kappa} < \kappa^*$, let $\epsilon = (\kappa^* - \widetilde{\kappa})/2$. Then, $\kappa_\varepsilon > \widetilde{\kappa}$ and (C.32) shows that there is a feasible κ for that is larger than $\widetilde{\kappa}$. This draws the contradiction and we have

$$\widetilde{\kappa} = \sup \left\{ \kappa > 0 : \inf_{u \geq 0} g_\delta(u, \kappa, c) < 0 \right\} = \kappa^*.$$

By definition, $(\widetilde{u}, \widetilde{\gamma})$ is the minimizer to $F_\delta(\gamma, u\gamma, \kappa^*)$ where $\widetilde{u}$ is the smallest one if the minimizer is not unique, then by the reasoning above that the minimizer locates on the boundary of $\widetilde{\Theta}_\eta$, it is equivalent that $\widetilde{u}$ is the minimizer to $g_\delta(u, \kappa^*, c)$. By the uniqueness of the smallest minimizer to $g_\delta(u, \kappa^*, c)$, we have $\widetilde{u} = u^*$. By the boundary equation, we have $\widetilde{\gamma} = 1/\sqrt{u^{*2} + c}$. □

Step 2: Proving $\kappa^*(c) < +\infty$. It suffices to provide some $\bar{\kappa} < +\infty$ such that $\inf_{u \geq 0} f_\delta(u, \bar{\kappa}, c) > 0$. Recall that

$$\begin{aligned}
f_\delta(u, \kappa, c) &= -u^2 + \delta \mathbb{E}\left[\left(\sqrt{1+u^2} G + \kappa \sqrt{u^2 + c} - \|\boldsymbol{\mu}\| \right)_+^2 \right] \\
&= u^2 \left\{ -1 + \delta \frac{1+u^2}{u^2} \mathbb{E}\left[\left(G + \kappa \sqrt{\frac{u^2+c}{u^2+1}} - \sqrt{\frac{1}{1+u^2}} \|\boldsymbol{\mu}\| \right)_+^2 \right] \right\} \\
&\geq u^2 \left\{ -1 + \delta \mathbb{E}\left[\left(G + \kappa \sqrt{\frac{u^2+c}{u^2+1}} - \sqrt{\frac{1}{1+u^2}} \|\boldsymbol{\mu}\| \right)_+^2 \right] \right\}.
\end{aligned}$$

Since $\mathbb{E}[(G+a)_+^2]$ is increasing in a and

$$\kappa \sqrt{\frac{u^2+c}{u^2+1}} - \sqrt{\frac{1}{1+u^2}} \|\boldsymbol{\mu}\| \geq (c \wedge 1)\kappa - \sqrt{\frac{1}{1+u^2}} \|\boldsymbol{\mu}\|.$$

Then, it suffices to show that there exists some $0 \le \bar{\kappa} < +\infty$ such that

$$\inf_{u \ge 0} \left\{ -1 + \delta \mathbb{E} \left[\left(G + (c \wedge 1)\bar{\kappa} - \sqrt{\frac{1}{1+u^2}} \|\boldsymbol{\mu}\| \right)_+^2 \right] \right\} = -1 + \delta \mathbb{E} \left[(G + (c \wedge 1)\bar{\kappa} - \|\boldsymbol{\mu}\|)_+^2 \right] > 0.$$

Since $c \wedge 1 > 0$, then $\mathbb{E}[(G + (c \wedge 1)\kappa - \|\boldsymbol{\mu}\|)_+^2]$ is an increasing and continuous function in κ, and

$$\lim_{\kappa \to +\infty} \left\{ -1 + \delta \mathbb{E}[(G + (c \wedge 1)\kappa - \|\boldsymbol{\mu}\|)_+^2] \right\} = +\infty.$$

Consequently, there exists $\bar{\kappa} < +\infty$ such that $-1 + \delta \mathbb{E}[(G + (c \wedge 1)\bar{\kappa} - \|\boldsymbol{\mu}\|)_+^2] > 0$, which proves that $\kappa^*(c) < +\infty$.

Step 3: Proving $u^*(c) < +\infty$. For a given c, denote

$$\mathcal{A} = \{(u, \kappa) \in \mathbb{R}_+ \times \mathbb{R}_+ : f_\delta(u, \kappa, c) \le 0\}, \qquad \mathcal{A}_\kappa = \{u \in \mathbb{R}_+ : (u, \kappa) \in \mathcal{A}\}.$$

Then, to show that $u^* < +\infty$, it is equivalent to show that $\mathcal{A}_{\kappa^*} \setminus \{+\infty\} \neq \varnothing$.

(1) $2 < \delta < \delta^*(\|\boldsymbol{\mu}\|)$. We first consider $\delta \in (2, \delta^*(\rho))$. As $u \to +\infty$, we have for any $\kappa \ge 0$,

$$\lim_{u \to +\infty} f_\delta(u, \kappa, c) = \lim_{u \to +\infty} \left\{ \left(\frac{\delta}{2} - 1 \right) u^2 \right\} = +\infty,$$

which implies that $+\infty \notin \mathcal{A}_\kappa$ for any $\kappa \ge 0$. Since $\mathcal{A}_{\kappa^*} \neq \varnothing$ by the existence of the max-margin solution, then $\mathcal{A}_{\kappa^*} \setminus \{+\infty\} \neq \varnothing$.

(2) $\delta \le 2$. When $\delta \le 2$, let κ_0 be the unique solution to

$$\frac{1}{\delta} = \mathbb{E}\left[(G + \kappa)_+^2 \right].$$

Since $f_\delta(u,\kappa,c)$ is increasing in κ, then for $\kappa > \kappa'$, $\mathcal{A}_\kappa \subseteq \mathcal{A}_{\kappa'}$. Also, for $\kappa < \kappa_0$, we have $\delta\mathbb{E}\left[(G+\kappa)_+^2\right] < 1$,

$$\lim_{u\to+\infty} f_\delta(u,\kappa,c) = \lim_{u\to+\infty} \left\{ \left(\delta\mathbb{E}\left[(G+\kappa)_+^2\right] - 1 \right) u^2 \right\} = -\infty,$$

which implies that $\mathcal{A}_\kappa \neq \varnothing$ and $+\infty \in \mathcal{A}_\kappa$ for $\kappa < \kappa_0$.

Further, if $\kappa > \kappa_0$, we have

$$\delta\mathbb{E}\left[(G+\kappa)_+^2\right] > 1,$$

which implies that

$$\lim_{u\to+\infty} f_\delta(u,\kappa,c) = \lim_{u\to+\infty} \left\{ -u^2 + \left(\delta\mathbb{E}\left[(G+\kappa)_+^2\right] \right) u^2 \right\} = +\infty.$$

Thus, $+\infty \notin \mathcal{A}_\kappa$ for $\kappa > \kappa_0$. Then, to complete the proof that $u^* < +\infty$, we need to show that $\mathcal{A}_{\kappa_0} \neq \varnothing$. Consider

$$\sqrt{1+u^2}G + \kappa_0\sqrt{u^2+c} = Gu + \kappa_0 u + \frac{G}{\sqrt{1+u^2}+u} + \frac{c}{\sqrt{u^2+c}+u},$$

there exists $C > 0$ such that for $u \geq C$, we have

$$\frac{c}{\sqrt{u^2+c}+u} < \frac{1}{2}\|\boldsymbol{\mu}\|.$$

Then, we have

$$\mathbb{E}\left\{\left(\sqrt{1+u^2}G+\kappa_0\sqrt{u^2+c}-\|\boldsymbol{\mu}\|\right)_+^2\right\} \le \mathbb{E}\left\{\left(Gu+\kappa_0 u+\frac{G}{\sqrt{1+u^2}+u}-\frac{1}{2}\|\boldsymbol{\mu}\|\right)_+^2\right\}$$
$$\le \mathbb{E}\left\{\left(Gu+\kappa_0 u-\frac{1}{4}\|\boldsymbol{\mu}\|\right)_+^2 \mathbb{1}\left\{\frac{G}{\sqrt{1+u^2}+u}<\frac{1}{4}\|\boldsymbol{\mu}\|\right\}\right\}$$
$$+\mathbb{E}\left\{\left(Gu+\kappa_0 u+\frac{G}{\sqrt{1+u^2}+u}\right)_+^2 \mathbb{1}\left\{\frac{G}{\sqrt{1+u^2}+u}>\frac{1}{4}\|\boldsymbol{\mu}\|\right\}\right\}$$
$$\le \mathbb{E}\left\{\left(Gu+\kappa_0 u-\frac{1}{4}\|\boldsymbol{\mu}\|\right)_+^2\right\}+\mathbb{E}\left\{\left(Gu+\kappa_0 u+\frac{G}{\sqrt{1+u^2}+u}\right)_+^2 \mathbb{1}\left\{\frac{G}{\sqrt{1+u^2}+u}>\frac{1}{4}\|\boldsymbol{\mu}\|\right\}\right\}$$
$$:= I_1(u)+I_2(u),$$

where $\lim_{u\to+\infty} I_2(u)=0$. For $I_1(u)$, consider the function

$$q(t)=\delta\mathbb{E}\left\{(G+\kappa_0-t)_+^2\right\}-1,$$

then $q(0)=1$ and $q'(t)=-2\delta\mathbb{E}\left\{(G+\kappa_0-t)_+\right\}<0$. At $t\approx 0$, we have the approximation $q(t)=q'(0)t+o(t)$. If we take $t=\|\boldsymbol{\mu}\|/(4u)$, $-u^2+\delta I_1(u)$ can be approximated by

$$-u^2+\delta I_1(u)=\left\{\delta\mathbb{E}\left\{\left(G+\kappa_0-\frac{\|\boldsymbol{\mu}\|}{4u}\right)_+^2\right\}-1\right\}u^2=\frac{\|\boldsymbol{\mu}\|}{4}uq'(0)+o(u),$$

thus there exists sufficiently large $u\ge C$ such that $-u^2+\delta(I_1(u)+I_2(u))<0$, which implies that $\mathcal{A}_{\kappa_0}\neq\varnothing$.

We consider the constrained problem with arbitrarily small $\varepsilon>0$,

$$\begin{cases}\max & \kappa \\ \text{subject to} & \inf_{u\ge 0,\kappa\ge\kappa_0+\varepsilon} f_\delta(u,\kappa,c)\le 0,\end{cases} \tag{C.33}$$

Let $(\widetilde{u},\widetilde{\kappa})$ be any solution to (C.33), where the existence is obtained by the continuity of

$g_\delta(u,\kappa,c)$ in κ. On one hand, since (C.33) is a constrained version of (C.27), then we have $\widetilde{\kappa} \le \kappa^*$; on the other hand, when $\widetilde{\kappa} < \kappa^*$, then (u^*, κ^*) is a feasible point for the constrained problem, thus $\kappa^* \le \widetilde{\kappa}$. Then, we have $\kappa^* = \widetilde{\kappa} > \kappa_0$, indicating that $0 < u^* < +\infty$ since $+\infty \notin \mathcal{A}_\kappa$ for $\kappa > \kappa_0$.

Combining pieces above, we have shown that in either regime $\delta \in (2, \delta^*(\rho))$ or $\delta \le 2$, the solution to (C.27) satisfies that $0 < u^* < +\infty$, which completes the proof.

Proof of Lemma C.3.11

Recall the notation

$$\begin{aligned}
\mathcal{E}_{n,p,\kappa,v} &\coloneqq \left\{\text{there exists } \boldsymbol{\beta} \in \Theta \cap \mathbf{B}_v : y_i \langle \boldsymbol{h}_{0,i}, \boldsymbol{\beta}\rangle \ge \kappa \text{ for all } i \le n\right\}, \\
\xi_{n,p,\kappa,v} &\coloneqq \xi_{n,p,\kappa}(\mathbf{B}_v \cap \Theta) = \min_{\boldsymbol{\beta}\in\mathbf{B}_v\cap\Theta} \max_{\boldsymbol{\lambda}:\|\boldsymbol{\lambda}\|\le 1, \mathbf{y}\odot\boldsymbol{\lambda}\ge 0} \frac{1}{\sqrt{p}} \boldsymbol{\lambda}^\top (\kappa \mathbf{y} - \mathbf{H}_0 \boldsymbol{\beta}).
\end{aligned}$$

where $\mathbf{B}_v = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\mathbf{P}_{\boldsymbol{\mu}}^\perp \boldsymbol{\beta}\| \le v \|\mathbf{P}_{\boldsymbol{\mu}} \boldsymbol{\beta}\|\}$ is a convex cone for each choice of $v \in \bar{\mathbb{R}}_+$. Accordingly, we denote

$$\begin{aligned}
\xi^{(2)}_{n,p,\kappa,v} &\coloneqq \xi^{(2)}_{n,p,\kappa}(\widetilde{\Theta}_\eta \cap \{(\gamma, z) : z \le v\gamma\}) \\
&= \min_{(\gamma,z)\in\widetilde{\Theta}_\eta \cap \{(\gamma,z): z\le v\gamma\}} \left\{ -z + \sqrt{\delta} \sqrt{\mathbb{E}\left(\sqrt{\gamma^2 + z^2} G + \kappa - \gamma \|\boldsymbol{\mu}\|\right)_+^2} \right\}
\end{aligned}$$

Then, by definition,

$$\xi_{n,p,\kappa} = \min_{v\in\bar{\mathbb{R}}_+} \xi_{n,p,\kappa,v}, \qquad \xi^{(2)}_{n,p,\kappa} = \min_{v\in\bar{\mathbb{R}}_+} \xi^{(2)}_{n,p,\kappa,v}.$$

Recall the definition of $\widehat{\kappa}$ and $\widehat{u}$ in Theorem 4.5.2, then

$$\widehat{\kappa} = \sup_{\kappa>0} \left\{\xi_{n,p,\kappa} = 0\right\}, \qquad \widehat{u} \in \left\{v : \ \xi_{n,p,\widehat{\kappa},v} = 0\right\}.$$

Step 1: Proving $\mathcal{E}_{n,p,\kappa^*,u^*+\varepsilon}$ and $\mathcal{E}^c_{n,p,\kappa^*,u^*-\varepsilon}$. For any $\varepsilon > 0$, we first show that

$$\lim_{n\to+\infty, n/p\to\delta} \mathbb{P}\left(\xi_{n,p,\kappa^*,u^*-\varepsilon} > 0\right) = 1.$$

By the result in Lemma C.3.8, we need to show that

$$\inf_{(\gamma,z)\in\widetilde{\Theta}_\eta \cap \{(\gamma,z): z\le (u^*-\varepsilon)\gamma\}} F_\delta(\gamma, z, \kappa^*) > 0,$$

up to the change of variables, which is equivalent to

$$\inf_{(\gamma,u)\in\widetilde{\Omega}\cap\{(\gamma,u): z\le u^*-\varepsilon\}} \widetilde{F}_\delta(\gamma, u, \kappa^*) > 0.$$

Recall the definition of u^* as the smallest minimizer of $f_\delta(u, \kappa^*, c)$ in (C.27). By Section C.3.2 and Lemma C.3.13, we have $\widetilde{F}_\delta((u^{*2}+c)^{-1/2}, u^*, \kappa^*) = 0$. Assume

$$\inf_{(\gamma,u)\in\widetilde{\Omega}\cap\{(\gamma,u): u\le u^*-\varepsilon\}} \widetilde{F}_\delta(\gamma, u, \kappa^*) \le 0,$$

which implies that there exists $v^* \le u^* - \varepsilon < u^*$ and γ_1 such that

$$\widetilde{F}_\delta(\gamma_1, v^*, \kappa^*) \le \widetilde{F}_\delta((u^{*2}+c)^{-1/2}, u^*, \kappa^*) = \inf_{(\gamma,u)\in\widetilde{\Omega}} \widetilde{F}_\delta(\gamma, u, \kappa^*),$$

then (γ_1, v^*) is also the minimizer to the unconstrained problem. This is a contradiction to the definition that u^* is the smallest minimizer, thus

$$\inf_{(\gamma,z)\in\widetilde{\Theta}_\eta \cap \{(\gamma,z): z\le (u^*-\varepsilon)\gamma\}} F_\delta(\gamma, z, \kappa^*) > 0 \qquad \Longrightarrow \qquad \lim_{n\to+\infty, n/p\to\delta} \mathbb{P}\left(\xi_{n,p,\kappa^*,u^*-\varepsilon} > 0\right) = 1.$$

In addition, since $\widetilde{\Theta}_\eta \cap \{(\gamma, z) : z \le u^*\gamma\} \subseteq \widetilde{\Theta}_\eta \cap \{(\gamma, z) : z \le (u^* + \varepsilon)\gamma\}$,

$$\begin{aligned}\inf_{(\gamma,z)\in\widetilde{\Theta}_\eta\cap\{(\gamma,z):z\le(u^*+\varepsilon)\gamma\}} F_\delta(\gamma, z, \kappa^*) &\le \inf_{(\gamma,z)\in\widetilde{\Theta}_\eta\cap\{(\gamma,z):z\le u^*\gamma\}} F_\delta(\gamma, z, \kappa^*) \le 0 \\ &= \inf_{(\gamma,u)\in\widetilde{\Omega}\cap\{(\gamma,u):u\le u^*\}} \widetilde{F}_\delta(\gamma, u, \kappa^*) \\ &= \inf_{u\in[0,u^*]} \widetilde{F}_\delta((u^{*2} + c)^{-1/2}, u^*, \kappa^*) \le 0.\end{aligned}$$

Then, for any $\varepsilon > 0$, we have

$$\mathcal{E}_{n,p,\kappa^*,u^*+\varepsilon} \qquad \text{and} \qquad \mathcal{E}^c_{n,p,\kappa^*,u^*-\varepsilon} \qquad \text{hold.}$$

Step 2: Proving $\widehat{u} \ge u^*$. Suppose the claim is wrong, then there exists $\varepsilon > 0$ such that $\widehat{u} \le u^* - \varepsilon < u^*$, which indicates that

$$\widehat{\boldsymbol{\beta}} \in \Theta \cap \mathbf{B}_{u^*-\varepsilon}.$$

Then, with $\widehat{\boldsymbol{\beta}}$,

$$y_i\langle \boldsymbol{h}_{0,i}, \widehat{\boldsymbol{\beta}}\rangle \ge \widehat{\kappa}, \qquad \text{for all } i \le n.$$

By Lemma C.3.8, we have

$$\inf_{0\le u\le u^*-\varepsilon} f_\delta(u, \widehat{\kappa}, c) \le 0.$$

As $\widehat{\kappa} \xrightarrow{p} \kappa^*$ and $f_\delta(u, \kappa, c)$, we have

$$\inf_{0\le u\le u^*-\varepsilon} f_\delta(u, \kappa^*, c) \le 0,$$

which implies that $\mathcal{E}_{n,p,\kappa^*,u^*-\varepsilon}$ holds. This draws the contradiction and proves that $\widehat{u} \ge u^*$.

Step 3: Proving $\widehat{u} \le u^* + 2\varepsilon$. For any $\varepsilon > 0$, as $\mathbf{B}_{u^*} \subseteq \mathbf{B}_{u^*+\varepsilon}$, we already have

$$\lim_{n\to+\infty, n/p\to\delta} \mathbb{P}\left(\xi_{n,p,\kappa^*,u^*+\varepsilon} = 0\right) = 1.$$

Denote by $\widehat{\boldsymbol{\beta}}^{\varepsilon}$ any maximizer to the optimization problem (C.25) with an additional constraint $\boldsymbol{\beta} \in \mathbf{B}_{u^*+\varepsilon}$, and let $\widehat{u}^{\varepsilon} = \|\mathbf{P}_{\boldsymbol{\mu}}^{\perp}\widehat{\boldsymbol{\beta}}^{\varepsilon}\|/\|\mathbf{P}_{\boldsymbol{\mu}}\widehat{\boldsymbol{\beta}}^{\varepsilon}\|$. Our next goal is to show that with probability approaching one, $\widehat{\boldsymbol{\beta}}$ satisfies $\|\mathbf{P}_{\boldsymbol{\mu}}^{\perp}\widehat{\boldsymbol{\beta}}\| \le (u^* + 2\varepsilon)\|\mathbf{P}_{\boldsymbol{\mu}}\widehat{\boldsymbol{\beta}}\|$. Once it is proved, then $\widehat{u} \xrightarrow{p} u^*$ since ε is arbitrary.

Consider any $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\|_{\boldsymbol{\Omega}} \le 1$ and $\boldsymbol{\beta} \notin \mathbf{B}_{u^*+2\varepsilon}$ where we recall

$$\mathbf{B}_{u^*+2\varepsilon} = \{\boldsymbol{x} \in \mathbb{R}^p : \|\mathbf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{x}\| \le (u^* + 2\varepsilon)\|\mathbf{P}_{\boldsymbol{\mu}}\boldsymbol{x}\|\}.$$

Since the region $\Theta = \{\boldsymbol{x} : \boldsymbol{x}^\top\boldsymbol{\Omega}\boldsymbol{x} \le 1\}$ is an ellipsoid, it must be strictly convex. For any $\boldsymbol{\beta} \in \Theta$, the interpolant

$$\boldsymbol{\beta}_\lambda := \lambda\widehat{\boldsymbol{\beta}}^{\varepsilon} + (1-\lambda)\boldsymbol{\beta}, \qquad \text{where } \lambda \in (0,1)$$

lies in Θ. Since $\|\widehat{\boldsymbol{\beta}}^{\varepsilon} - \boldsymbol{\beta}\| \ge c_\varepsilon > 0$ for certain constant c_ε, we can find (sufficiently small) constants $\lambda \in (0,1)$ and $\alpha > 0$ such that

$$\boldsymbol{\beta}_\lambda \in \mathbf{B}_{u^*+\varepsilon}, \quad \text{and} \quad (1+\alpha)\boldsymbol{\beta}_\lambda \in \Theta.$$

Let us use $\kappa(\boldsymbol{\beta}) = \min_{i\le n} y_i\langle \boldsymbol{h}_{0,i}, \boldsymbol{\beta}\rangle$ to denote the margin associated with $\boldsymbol{\beta}$. By the definition of $\widehat{\boldsymbol{\beta}}^{\varepsilon}$ and the fact that $(1+\alpha)\boldsymbol{\beta}_\lambda \in \mathbf{B}_{u^*+\varepsilon}$, we have

$$\kappa((1+\alpha)\boldsymbol{\beta}_\lambda) \le \kappa(\widehat{\boldsymbol{\beta}}^{\varepsilon}). \tag{C.34}$$

By linearity, we also have

$$\kappa((1+\alpha)\boldsymbol{\beta}_\lambda) \geq (1+\alpha)\lambda\kappa(\widehat{\boldsymbol{\beta}}^\varepsilon) + (1+\alpha)(1-\lambda)\kappa(\boldsymbol{\beta}). \tag{C.35}$$

Combining (C.34) and (C.35), we have

$$\kappa(\boldsymbol{\beta}) \leq \frac{1-(1+\alpha)\lambda}{(1+\alpha)(1-\lambda)}\kappa(\widehat{\boldsymbol{\beta}}^\varepsilon) < \kappa(\widehat{\boldsymbol{\beta}}^\varepsilon).$$

This proves that if $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}} \notin \mathbf{B}_{u^*+2\varepsilon}$, then $\widehat{\boldsymbol{\beta}} < \kappa(\widehat{\boldsymbol{\beta}}^\varepsilon)$, which is a contradiction. Then, we have shown that $\widehat{\boldsymbol{\beta}} \in \mathbf{B}_{u^*+2\varepsilon}$. Consequently, we have $\widehat{u} \in [u^*, u^*+2\varepsilon)$ for any $\varepsilon > 0$, thus

$$\widehat{u} = \frac{\|\mathbf{P}_{\boldsymbol{\mu}}^\perp \widehat{\boldsymbol{\beta}}\|}{\|\mathbf{P}_{\boldsymbol{\mu}} \widehat{\boldsymbol{\beta}}\|} \xrightarrow{p} u^*.$$

Proof of Lemma C.3.12

By the definition of $\kappa^*(c)$ and $u^*(c)$, it is easy to check that

$$\begin{cases} f_\delta(u^*(c), \kappa^*(c), c) & = 0, \\ \partial_u f_\delta(u^*(c), \kappa^*(c), c) & = 0. \end{cases}$$

In other words, for any c, $(u^*(c), \kappa^*(c), c)$ is the zero point of the systems of equations

$$\begin{cases} f(u, \kappa, c) & = 0, \\ \partial_u f(u, \kappa, c) & = 0, \end{cases} \tag{C.36}$$

where we identify f_δ with f to simply the notation.

Fix any c_0, and its corresponding $u^*(c_0), \kappa^*(c_0)$. We can compute that

$$\partial_c \partial_u f - \frac{\partial_c f}{\partial_\kappa f} \partial_\kappa \partial_u f \mid_{u^*(c_0),\kappa^*(c_0),c_0} < 0 \tag{C.37}$$

As a result, the implicit function theorem tells us that in an open neighborhood around $u^*(c_0)$, the zero point of (C.36) can be written as a continuously differentiable function $u \to \kappa^*(u), c^*(u)$. Moreover, implicit function theorem also tells us that

$$\left(\partial_c \partial_u f - \frac{\partial_c f}{\partial_\kappa f} \partial_\kappa \partial_u f \right) \mid_{u^*(c_0),\kappa^*(c_0),c_0} \cdot c'(u^*(c_0)) + \partial_u^2 f \mid_{u^*(c_0),\kappa^*(c_0),c_0} = 0$$

As $u^*(c_0)$ is a minimizer for $f(u, \kappa^*(c_0), c_0)$, we have $\partial_u^2 f \mid_{u^*(c_0),\kappa^*(c_0),c_0} \geq 0$. In all, we conclude that $c'(u^*(c_0)) \geq 0$.

We have the following lemma which shows the continuity of $u^*(c)$.

Lemma C.3.14. *Both $\kappa^*(c)$ and $u^*(c)$ are continuous in c.*

See the end of this section for the proof.

According to Lemma C.3.14, in an open neighborhood $\mathcal{O}$ of $(u_0^*, \kappa_0^*, c_0) = (u^*(c_0), \kappa^*(c_0), c_0)$, for any $c_1 > c_0$ such that $c_1 - c_0$ is sufficiently small, denote (u_1^*, κ_1^*) as the solution to (C.27) with $c = c_1$. Since both $u^*(c)$ and $\kappa^*(c)$ are continuous in c, we have $(u_0^*, \kappa_0^*, c_0), (u_1^*, \kappa_1^*, c_1) \in \mathcal{O}$. By the implicit function theorem, in the neighborhood $\mathcal{O}$, c can be written as a differentiable function $c(u^*)$ of u^* with $0 \leq c'(u^*) < +\infty$. Suppose we do not have $u_0^* < u_1^*$. If $u_0^* > u_1^*$, by the mean value theorem, then there exists $u_2^* \in (u_1^*, u_0^*)$ such that $c'(u_2^*) < 0$, which draws the contradiction. If $u_0^* = u_1^*$, then u_0^* is mapped to two distinct values and the function $c(u)$ is not well-defined. Therefore, we have $u_0^* < u_1^*$ and $u^*(c)$ is increasing in c.

Proof of (C.37)**.** Denote $Z(u) = \sqrt{1+u^2}G + \kappa\sqrt{u^2+c} - \|\boldsymbol{\mu}\|$. By direct calculations, we have

1. $\partial_\kappa f = 2\delta\sqrt{u^2+c}\mathbb{E}[Z_+(u)]$,

2. $\partial_c f = \frac{\kappa\delta}{\sqrt{u^2+c}}\mathbb{E}[Z_+(u)]$,

3. $\partial_u f = -2u + 2u\delta\mathbb{P}(Z(u)\geq 0) + 2\delta\frac{\kappa u}{\sqrt{u^2+c}}\mathbb{E}[Z_+(u)]$,

4. $\partial_\kappa\partial_u f = 2u\delta\frac{\sqrt{u^2+c}}{\sqrt{u^2+1}}\phi\left(\frac{\sqrt{u^2+c}}{\sqrt{u^2+1}}\kappa - \frac{\|\boldsymbol{\mu}\|}{\sqrt{u^2+1}}\right) + 2u\frac{\delta}{\sqrt{u^2+c}}\mathbb{E}[Z_+(u)] + 2u\delta\kappa\mathbb{P}(Z(u)\geq 0)$,

5. $\partial_c\partial_u f = u\delta\frac{\kappa}{\sqrt{u^2+1}}\frac{1}{\sqrt{u^2+c}}\phi\left(\frac{\sqrt{u^2+c}}{\sqrt{u^2+1}}\kappa - \frac{\|\boldsymbol{\mu}\|}{\sqrt{u^2+1}}\right) - u\delta\kappa\frac{1}{(u^2+c)^{3/2}}\mathbb{E}[Z_+(u)] + \delta\frac{\kappa^2}{u^2+c}\mathbb{P}(Z(u)\geq 0)$.

Denoting $\nu = \frac{\sqrt{u^2+c}}{\sqrt{u^2+1}}\kappa - \frac{\|\boldsymbol{\mu}\|}{\sqrt{u^2+1}}$, one then has

$$\frac{\partial_c f}{\partial_\kappa f}\partial_\kappa\partial_u f = \delta u\kappa\sqrt{\frac{1}{(u^2+1)(u^2+c)}}\phi(\nu) + \frac{\delta u\kappa}{(u^2+c)^{3/2}}\mathbb{E}[Z_+(u)] + \frac{\delta u\kappa^2}{u^2+c}\mathbb{P}(Z(u)\geq 0).$$

As the result, for any $0 < u < +\infty$, we have

$$\partial_c\partial_u f - \frac{\partial_c f}{\partial_\kappa f}\partial_\kappa\partial_u f = -\frac{2\delta u\kappa}{(u^2+c)^{3/2}}\mathbb{E}[Z_+(u)] < 0.$$

Proof of Lemma C.3.14. Denote $\boldsymbol{w} = (\gamma, u)$ and

$$\kappa(\boldsymbol{w}) = \sup_{\kappa\geq 0}\{\widetilde{F}_\delta(\gamma, u, \kappa)\leq 0\},$$

where $\kappa(\boldsymbol{w}) = 0$ if $\{\kappa\geq 0 : \widetilde{F}_\delta(\gamma, u, \kappa)\leq 0\} = \varnothing$, then $\kappa(\boldsymbol{w})$ is continuous in $\boldsymbol{w}$. Rewrite $U(c) = \{(\gamma, u)\in\mathbb{R}\times\mathbb{R}_+ : \gamma^2(u^2+c)\leq 1\}$, which satisfies that $U(c_1)\subseteq U(c_2)$ for $c_1\geq c_2$, and denote

$$\kappa^*(c) = \sup_{\boldsymbol{w}\in U(c)}\kappa(\boldsymbol{w})$$

Consider any $0 < c < +\infty$ and $\boldsymbol{w}^*(c)$ is the associated solution to (C.31).

Step 1: Continuity of $\kappa^*(c)$. We first show that $\kappa^*(c)$ is continuous in c. Otherwise, without loss of generality, there exists $\varepsilon > 0$ such that there exists an increasing sequence $\{c_m\}_{m\geq 1}$ with $c/2 < c_m < c$ and $c - c_m \leq \frac{1}{m}$ which satisfies that $|\kappa^*(c_m) - \kappa^*(c)| \geq \varepsilon$, which forms a non-increasing sequence $\{\kappa^*(c_m)\}_{m=1}^{\infty}$ and $\{\boldsymbol{w}^*(c_m)\}_{m=1}^{\infty} \subseteq U(c/2)$. Recall that $\kappa^*(c)$ is non-increasing in c, then $\kappa^*(c_m) \geq \kappa^*(c) + \varepsilon$ and $\kappa^*(c_m) \in [\kappa^*(c), \kappa^*(c/2)]$. Then, for the sequence in a compact set, there is a subsequence $\{c_{m_k}\}_{k=1}^{\infty} \subseteq \{c_m\}_{m=1}^{\infty}$ such that

$$\lim_{k\to+\infty} c_{m_k} = c, \qquad \lim_{k\to+\infty} \kappa^*(c_{m_k}) = \widetilde{\kappa}, \qquad \lim_{k\to+\infty} \boldsymbol{w}^*(c_m) = \widetilde{\boldsymbol{w}} \in \bigcap_{k=1}^{\infty} U(c_{m_k}) = U(c).$$

Thus, there exists $K \in \mathbb{N}$ such that for any $k \geq K$,

$$\kappa(\widetilde{\boldsymbol{w}}) \geq \kappa(\boldsymbol{w}^*(c_{m_k})) - \frac{\varepsilon}{2} = \kappa^*(c_{m_k}) - \frac{\varepsilon}{2} \geq \kappa^*(c) + \frac{\varepsilon}{2} > \kappa^*(c),$$

which violates the fact that $\boldsymbol{w}^*(c) = \underset{\boldsymbol{w}\in U(c)}{\operatorname{argmax}} \kappa(\boldsymbol{w})$. This proves the left-continuity of $\kappa^*(c)$ and the right-continuity follows the same technique.

Step 2: Continuity of $\boldsymbol{w}^*(c)$. Suppose $\boldsymbol{w}^*(c) = (\gamma^*(c), u^*(c))$ is not continuous in c, particularly for the second coordinate, then there exists $\varepsilon > 0$ such that for any $m \in \mathbb{N}$, there exists an increasing sequence $\{c_m\}_{m\geq 1}$ with $c/2 < c_m < c$ and $c - c_m \leq \frac{1}{m}$ such that $|u^*(c_m) - u^*(c)| \geq \varepsilon$, which forms a sequence $\{\boldsymbol{w}^*(c_m)\}_{m=1}^{\infty}$. Also, we have $\boldsymbol{w}^*(c_m) = (\gamma^*(c_m), u^*(c_m)) \in U(c_m) \subseteq U(c/2)$ for each m. As $U(c/2)$ is compact, then there is a subsequence $\{c_{m_k}\}_{k=1}^{\infty} \subseteq \{c_m\}_{m=1}^{\infty}$ such that

$$\lim_{k\to+\infty} c_{m_k} = c, \qquad \lim_{k\to+\infty} \boldsymbol{w}^*(c_{m_k}) = \widetilde{\boldsymbol{w}} = (\widetilde{\gamma}, \widetilde{u}) \in \bigcap_{k=1}^{\infty} U(c_{m_k}) = U(c).$$

Since $\widetilde{F}_\delta(\boldsymbol{w}, \kappa)$ is continuous in $(\boldsymbol{w}, \kappa)$ and for any k, $\widetilde{F}_\delta(\boldsymbol{w}^*(c_{m_k}), \kappa^*(c_{n_k})) = 0$, then $\widetilde{F}_\delta(\widetilde{\boldsymbol{w}}, \widetilde{\kappa}) = 0$ with $|\widetilde{u} - u^*(c)| \geq \varepsilon$. Denote

$$\mathcal{B}_k = \left\{ \boldsymbol{w} : \widetilde{F}_\delta(\boldsymbol{w}, \kappa^*(c_{m_k})) \leq 0 \right\}, \qquad \mathcal{B}_\infty = \left\{ \boldsymbol{w} : \widetilde{F}_\delta(\boldsymbol{w}, \kappa^*(c)) \leq 0 \right\}$$

Since $\kappa^*(c_{m_k})$ is non-increasing as k increases with $\lim_{k \to +\infty} \kappa^*(c_{m_k}) = \kappa^*(c)$, and $\widetilde{F}_\delta(\boldsymbol{w}, \kappa)$ is continuous and increasing in κ, we have

$$\mathcal{B}_k \subseteq \mathcal{B}_{k+1} \subseteq \mathcal{B}_\infty, \qquad \bigcup_{k \geq 1} \mathcal{B}_k = \mathcal{B}_\infty.$$

By definition, $u^*(c_{m_k}) = \inf\{u : (\gamma, u) \in \mathcal{B}_k\}$, then as $u^*(c_{m_k}) \to \widetilde{u}$,

$$\widetilde{u} \leq \inf\{u : (\gamma, u) \in \mathcal{B}_k\}, \qquad \text{for any } k.$$

Consequently, we have

$$\widetilde{u} \leq \inf_{k \geq 1} \inf \{u : (\gamma, u) \in \mathcal{B}_k\} = \inf \left\{ u : (\gamma, u) \in \bigcup_{k \geq 1} \mathcal{B}_k \right\} = \inf \{u : (\gamma, u) \in \mathcal{B}_\infty\}.$$

In addition, by definition, $u^*(c) = \inf \{u : (\gamma, u) \in \mathcal{B}_\infty\}$, we have $\widetilde{u} \leq u^*(c)$ and $\widetilde{u}, u^*(c) \in \{u : (\gamma, u) \in \mathcal{B}_\infty\}$ such that $|\widetilde{u} - u^*(c)| \geq \varepsilon > 0$, which draws the contradiction. Thus, $u^*(c)$ is continuous in c.

C.4 Proof for Section 4.6

C.4.1 Proof of Proposition 4.6.1

Recall that we assumed features and augmentations with inhomogeneous covariance take the following form.

$$\begin{aligned}
\boldsymbol{h}_{0,i} &\overset{\text{i.i.d.}}{\sim} \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_p) + \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_p) \\
\boldsymbol{h}_i, \boldsymbol{h}_i^+ \,|\, \boldsymbol{h}_{0,i} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{h}_{0,i}, \sigma_{\text{aug}}^2 \mathbf{A}).
\end{aligned}$$

where $\mathbf{A} \succeq \mathbf{I}_p$. The special case where $\mathbf{A} = \mathbf{I}_p$ is already analyzed before. Here, we can write

$$\begin{aligned}
&\boldsymbol{h}_i \overset{d}{=} \xi_i \boldsymbol{\mu} + (\mathbf{I}_p + \sigma_{\text{aug}}^2 \mathbf{A})^{1/2} \boldsymbol{g}_i, \qquad \text{where} \\
&\xi_i \perp \boldsymbol{g}_i, \qquad \xi \sim \text{Unif}(\{\pm 1\}), \qquad \boldsymbol{g}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p).
\end{aligned}$$

We have

$$\begin{aligned}
\mathbb{E}\big[\|\mathbf{W}\boldsymbol{h}_i\|^2\big] &= \|\mathbf{W}\boldsymbol{\mu}\|^2 + \text{Tr}\big(\mathbf{W}(\mathbf{I}_p + \sigma_{\text{aug}}^2 \mathbf{A})\mathbf{W}^\top\big) \\
&= \|\mathbf{W}\boldsymbol{\mu}\|^2 + \|\mathbf{W}\|_{\text{F}}^2 + \sigma_{\text{aug}}^2 \text{Tr}\big(\mathbf{W}\mathbf{A}\mathbf{W}^\top\big) =: \widetilde{\alpha}, \\
\mathbb{E}\big[\|\mathbf{W}\boldsymbol{h}_i^+ - \mathbf{W}\boldsymbol{h}_i\|^2\big] &= 2\sigma_{\text{aug}}^2 \text{Tr}\big(\mathbf{W}\mathbf{A}\mathbf{W}^\top\big) \\
\|\mathbf{W}\boldsymbol{h}_i - \mathbf{W}\boldsymbol{h}_i^-\|^2 &= \|\mathbf{W}(\xi_i - \xi_i^-)\boldsymbol{\mu} + \mathbf{W}(\mathbf{I}_p + \sigma_{\text{aug}}^2 \mathbf{A})^{1/2}(\boldsymbol{g}_i - \boldsymbol{g}_i^-)\|^2.
\end{aligned}$$

Consider the same infinite-sample loss $\mathcal{L}(\mathbf{W})$ in Section 4.4. The alignment loss can be written as

$$\mathcal{L}_{\text{align}}(\mathbf{W}) = \frac{1}{2\tau\widetilde{\alpha}} \mathbb{E}_{\boldsymbol{h},\boldsymbol{h}^+} \Big[\|\mathbf{W}\boldsymbol{h} - \mathbf{W}\boldsymbol{h}^+\|^2\Big] = \frac{\sigma_{\text{aug}}^2}{\tau\widetilde{\alpha}} \text{Tr}\big(\mathbf{W}\mathbf{A}\mathbf{W}^\top\big).$$

Further, the difference between the negative pair has the distribution

$$\boldsymbol{h} - \boldsymbol{h}^- \stackrel{d}{=} \frac{1}{2}\delta_0 + \frac{1}{4}\delta_{2\boldsymbol{\mu}} + \frac{1}{4}\delta_{-2\boldsymbol{\mu}} + \sqrt{2}\left(\mathbf{I} + \sigma_{\text{aug}}^2\mathbf{A}\right)^{1/2}\boldsymbol{g},$$

where $\boldsymbol{g} \sim \mathcal{N}(0, \mathbf{I})$. Let $\widetilde{\mathbf{W}} = \mathbf{W}\left(\mathbf{I} + \sigma_{\text{aug}}^2\mathbf{A}\right)^{1/2}$ and $\widetilde{\boldsymbol{\mu}} = \left(\mathbf{I} + \sigma_{\text{aug}}^2\mathbf{A}\right)^{-1/2}\boldsymbol{\mu}$. Then, we have

$$\mathbf{W}\boldsymbol{h} - \mathbf{W}\boldsymbol{h}^- \stackrel{d}{=} \frac{1}{2}\delta_0 + \frac{1}{4}\delta_{2\widetilde{\mathbf{W}}\widetilde{\boldsymbol{\mu}}} + \frac{1}{4}\delta_{-2\widetilde{\mathbf{W}}\widetilde{\boldsymbol{\mu}}} + \sqrt{2}\widetilde{\mathbf{W}}\boldsymbol{g},$$

which has the same form as in Section C.2.1. Therefore, with the SVD of $\widetilde{\mathbf{W}} = \sum_{j=1}^p \widetilde{\sigma}_j \widetilde{\boldsymbol{u}}_j \widetilde{\boldsymbol{v}}_j^\top$, we can calculate the following expectation

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{h}^-,\boldsymbol{h}}\left[\exp\left(-\frac{\|\widetilde{\mathbf{W}}\boldsymbol{h} - \widetilde{\mathbf{W}}\boldsymbol{h}^-\|_2^2}{2\tau\widetilde{\alpha}}\right)\right] \\
&= \frac{1}{2}\prod_{j=1}^p\left(1 + \frac{2\widetilde{\sigma}_j^2}{\tau\widetilde{\alpha}}\right)^{-1/2}\left(1 + \exp\left(-\sum_{j=1}^p \frac{2\widetilde{\sigma}_j^2}{2\widetilde{\sigma}_j^2 + \tau\widetilde{\alpha}}(\widetilde{\boldsymbol{v}}_j^\top\widetilde{\boldsymbol{\mu}})^2\right)\right).
\end{aligned}$$

As a result, we have decomposition $\mathcal{L}(\mathbf{W}) = \mathcal{L}_{\text{align}}(\mathbf{W}) + \mathcal{L}_{\text{unif}}(\mathbf{W})$, where

$$\begin{aligned}
\mathcal{L}_{\text{align}}(\mathbf{W}) &= \frac{\sigma_{\text{aug}}^2}{\tau\widetilde{\alpha}}\operatorname{Tr}\big(\mathbf{W}\mathbf{A}\mathbf{W}^\top\big) \\
\mathcal{L}_{\text{unif}}(\mathbf{W}) &= -\frac{1}{2}\sum_{j=1}^p \log\left(1 + \frac{2\widetilde{\sigma}_j^2}{\tau\widetilde{\alpha}}\right) + \log\left(1 + \exp\Big(-\sum_{j=1}^p \frac{2\widetilde{\sigma}_j^2\langle\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{v}}_j\rangle^2}{2\widetilde{\sigma}_j^2 + \tau\widetilde{\alpha}}\Big)\right) - \log 2.
\end{aligned}$$

Further, if we assume that $\widetilde{\alpha} \gg \|\mathbf{W}\|_{\text{F}}^2$, then $\mathcal{L}_{\text{unif}}(\mathbf{W})$ can be approximated by

$$\begin{aligned}
\widetilde{\mathcal{L}}_{\text{unif}}(\mathbf{W}) &= -\sum_{j=1}^p \frac{\widetilde{\sigma}_j^2}{\tau\widetilde{\alpha}} + \log\left(1 + \exp\left(-\sum_{j=1}^p \frac{2\widetilde{\sigma}_j^2\langle\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{v}}_j\rangle^2}{\tau\widetilde{\alpha}}\right)\right) - \log 2 \\
&= -\frac{\|\widetilde{\mathbf{W}}\|_{\text{F}}^2}{\tau\widetilde{\alpha}} + \log\left(1 + \exp\Big(-\frac{2\|\widetilde{\mathbf{W}}\widetilde{\boldsymbol{\mu}}\|^2}{\tau\widetilde{\alpha}}\Big)\right) - \log 2 \\
&= -\frac{1}{\tau\widetilde{\alpha}}\left(\|\mathbf{W}\|_{\text{F}}^2 + \sigma_{\text{aug}}^2\operatorname{Tr}\big(\mathbf{W}\mathbf{A}\mathbf{W}^\top\big)\right) + \log\left(1 + \exp\Big(-\frac{2\|\mathbf{W}\boldsymbol{\mu}\|^2}{\tau\widetilde{\alpha}}\Big)\right) - \log 2.
\end{aligned}$$

We can then introduce an approximate loss $\widetilde{\mathcal{L}}(\mathbf{W}) = \mathcal{L}_{\text{align}}(\mathbf{W}) + \widetilde{\mathcal{L}}_{\text{unif}}(\mathbf{W})$, where

$$\widetilde{\mathcal{L}}(\mathbf{W}) = -\frac{1}{\tau\widetilde{\alpha}}\|\mathbf{W}\|_{\text{F}}^2 + \log\left(1 + \exp\Big(-\frac{2\|\mathbf{W}\boldsymbol{\mu}\|^2}{\tau\widetilde{\alpha}}\Big)\right).$$

C.4.2 Proof of Theorem 4.6.2

Recall that the minimization problem is $\min_{\mathbf{W}} \widetilde{\mathcal{L}}(\mathbf{W})$, where

$$\widetilde{\mathcal{L}}(\mathbf{W}) = -\frac{1}{\tau\widetilde{\alpha}}\|\mathbf{W}\|_{\text{F}}^2 + \log\left(1 + \exp\left(-\frac{2\|\mathbf{W}\boldsymbol{\mu}\|^2}{\tau\widetilde{\alpha}}\right)\right),$$
$$\widetilde{\alpha} = (1+\sigma_{\text{aug}}^2)\|\mathbf{W}\|_{\text{F}}^2 + \|\mathbf{W}\boldsymbol{\mu}\|^2 + \sigma_{\text{aug}}^2\rho_{\text{aug}}\|\mathbf{W}\boldsymbol{v}_{\text{aug}}\|^2.$$

First, notice that the loss $\widetilde{\mathcal{L}}(\mathbf{W})$ is scale-invariant. Hence without loss of generality, we can assume $\|\mathbf{W}\|_{\text{F}}^2 = 1$. Then $\widetilde{\mathcal{L}}(\mathbf{W})$ only depends on the values of

$$T(\mathbf{W}) = \|\mathbf{W}\bar{\boldsymbol{\mu}}\|^2, \qquad \text{and} \qquad \|\mathbf{W}\boldsymbol{v}_{\text{aug}}\|^2.$$

Second, we can rewrite the optimization problem into a nested one:

$$\min_{T\in[0,1]}\Big[\min_{\mathbf{W}:T(\mathbf{W})=T}\widetilde{\mathcal{L}}(\mathbf{W})\Big]. \tag{C.38}$$

It is easy to see that given any fixed value of $T(\mathbf{W})$, the loss $\widetilde{\mathcal{L}}(\mathbf{W})$ is decreasing as $\|\mathbf{W}\boldsymbol{v}_{\text{aug}}\|^2$ decreases. Thus, the inner minimization problem is easy to solve: we only need to determine the smallest possible value of $\|\mathbf{W}\boldsymbol{v}_{\text{aug}}\|^2$, which will be the focus below.

Consider the Gram–Schmidt orthogonalization for $\bar{\boldsymbol{\mu}}, \boldsymbol{v}_{\text{aug}}$: Let $\boldsymbol{v}_\perp \in \text{span}(\bar{\boldsymbol{\mu}}, \boldsymbol{v}_{\text{aug}})$ be a unit vector orthogonal to $\bar{\boldsymbol{\mu}}$; define $a \in [-1, 1]$ to be coefficient in the orthogonal decomposition

$$\bar{\boldsymbol{\mu}} = r\boldsymbol{v}_{\text{aug}} + a\boldsymbol{v}_\perp,$$

where we recall that $r = \langle \bar{\boldsymbol{\mu}}, \boldsymbol{v}_{\text{aug}} \rangle \in [-1, 1]$. It is also clear that $a^2 + r^2 = 1$.

We split the inner minimization into two cases.

Case 1: $0 \le T \le a^2$**.** We claim that there exists some $\mathbf{W}$ such that $\|\mathbf{W}\|_{\text{F}}^2 = 1$, $\|\mathbf{W}\bar{\boldsymbol{\mu}}\|^2 = T$ and $\|\mathbf{W}\boldsymbol{v}_{\text{aug}}\| = 0$. To prove this claim, we will construct a $\mathbf{W}$ of the form

$$\mathbf{W} = \sum_{j=1}^{p} s_j \bar{\boldsymbol{u}}_j \bar{\boldsymbol{v}}_j^\top,$$

such that the required equalities hold. Let us set

$$\begin{aligned} \bar{\boldsymbol{v}}_p &= \boldsymbol{v}_{\text{aug}}, \qquad & s_p &= 0, \\ s_1^2 &= \frac{T}{a^2} \le 1, \qquad & \bar{\boldsymbol{v}}_1 &= \boldsymbol{v}_\perp. \end{aligned}$$

Choose $(\bar{\boldsymbol{v}}_j)_{2 \le j \le p-1}$ be any orthogonal basis in the orthogonal complement of $\text{span}(\bar{\boldsymbol{v}}_1, \bar{\boldsymbol{v}}_p)$, and $\bar{\boldsymbol{u}}_1, \ldots, \bar{\boldsymbol{u}}_p$ be any orthogonal basis in $\mathbb{R}^p$. Also, choose any $s_2, \ldots, s_{p-1} \ge 0$ such that $s_2^2 + \ldots + s_{p-1}^2 = 1 - T/a^2$.

Our construction in fact gives an SVD of $\mathbf{W}$, with $(s_p, \bar{\boldsymbol{v}}_p)$ being the bottom singular pair. So we have $\mathbf{W}\boldsymbol{v}_{\text{aug}} = s_p \boldsymbol{v}_{\text{aug}} = 0$. Moreover,

$$\begin{aligned} \|\mathbf{W}\bar{\boldsymbol{\mu}}\|^2 &= \sum_{j=1}^{p} s_j^2 \langle \bar{\boldsymbol{v}}_j, \bar{\boldsymbol{\mu}} \rangle^2 \\ &= \sum_{j=1}^{p-1} s_j^2 \langle \bar{\boldsymbol{v}}_j, \bar{\boldsymbol{\mu}} \rangle^2 \qquad \text{since } s_p = 0 \\ &= s_1^2 \langle \bar{\boldsymbol{v}}_1, \bar{\boldsymbol{\mu}} \rangle^2 \qquad \text{by orthogonality } \boldsymbol{\mu} \perp (\bar{\boldsymbol{v}}_j)_{2 \le j \le p-1} \\ &= s_1^2 a^2 = T. \end{aligned}$$

Thus the inner minimization of (C.38) is solved at $\|\mathbf{W}\boldsymbol{v}_{\text{aug}}\| = 0$.

In all, under the constraint $T \le a^2$, the problem of minimizing $\widetilde{\mathcal{L}}(\mathbf{W})$ boils down to

$$\min_{T \le a^2} \quad -\frac{1}{\tau\left[(1+\sigma_{\text{aug}}^2) + \|\boldsymbol{\mu}\|^2 T\right]} + \log\left(1 + \exp\left(-\frac{2\|\boldsymbol{\mu}\|^2 T}{\tau\left[(1+\sigma_{\text{aug}}^2) + \|\boldsymbol{\mu}\|^2 T\right]}\right)\right).$$

This objective is in fact the same as that in Section 4.4 via change of variables. Denote $t = \left[(1+\sigma_{\text{aug}}^2) + \|\boldsymbol{\mu}\|^2 T\right]^{-1} \in [(1+\sigma_{\text{aug}}^2 + \|\boldsymbol{\mu}\|^2 a^2)^{-1}, (1+\sigma_{\text{aug}}^2)^{-1}]$. Then the above objective function can be written as

$$h(t) = -\frac{t}{\tau} + \log\left(1 + \exp\left(-\frac{2}{\tau} + 2\frac{1+\sigma_{\text{aug}}^2}{\tau} t\right)\right),$$

for which

$$h'(t) = \frac{1+2\sigma_{\text{aug}}^2}{\tau} - 2\frac{1+\sigma_{\text{aug}}^2}{\tau} \frac{1}{1 + \exp\left(-\frac{2}{\tau} + 2\frac{1+\sigma_{\text{aug}}^2}{\tau} t\right)}.$$

As in the homogeneous case, note that $h'(t)$ is increasing in t and $h'((1+\sigma_{\text{aug}}^2)^{-1}) = \sigma_{\text{aug}}^2/\tau > 0$. Further, notice that

$$h'((1+\sigma_{\text{aug}}^2 + \|\boldsymbol{\mu}\|^2 a^2)^{-1}) > 0 \iff \tau > \frac{2a^2\|\boldsymbol{\mu}\|^2}{\log(1+2\sigma_{\text{aug}}^2)(1+\sigma_{\text{aug}}^2 + a^2\|\boldsymbol{\mu}\|^2)} = \tau_1^*.$$

Then, if $\tau > \tau_1^*$, then $h'((1+\sigma_{\text{aug}}^2 + \|\boldsymbol{\mu}\|^2 a^2)^{-1}) > 0$ and $h(t)$ is increasing in $t \in [(1+\sigma_{\text{aug}}^2 + \|\boldsymbol{\mu}\|^2 a^2)^{-1}, (1+\sigma_{\text{aug}}^2)^{-1}]$, so the minimum of $h(t)$ is achieved when t is the smallest in the interval, or equivalently $T = a^2$. If $\tau \le \tau_1^*$, the minimizer in the interval $[0, a^2]$ is determined by $h'(t) = 0$, which yields

$$t^* = \frac{1}{1+\sigma_{\text{aug}}^2}\left(1 - \frac{\tau}{2}\log(1+2\sigma_{\text{aug}}^2)\right).$$

As a result, we have

$$T^* = \frac{1}{\|\boldsymbol{\mu}\|^2}\left(\frac{1}{t^*} - (1+\sigma_{\text{aug}}^2)\right) = \frac{\tau(1+\sigma_{\text{aug}}^2)}{2\|\boldsymbol{\mu}\|^2}\log(1+2\sigma_{\text{aug}}^2)\left[1 - \frac{\tau}{2}\log(1+2\sigma_{\text{aug}}^2)\right]^{-1} \quad \text{(C.39)}$$

Case 2: $a^2 < T \leq 1$**.** Note that we implicitly assume $|a| < 1$ since $|a| = 1$ is always covered in Case 1. In order to solve the inner minimization problem of (C.38), it suffices to consider the semidefinite program

$$\begin{aligned} \min \quad & \boldsymbol{v}_{\text{aug}}^\top \mathbf{A} \boldsymbol{v}_{\text{aug}} \\ \text{subject to} \quad & \text{Tr}(\mathbf{A}) = 1,\ \bar{\boldsymbol{\mu}}^\top \mathbf{A} \bar{\boldsymbol{\mu}} = T,\ \mathbf{A} \succeq 0, \end{aligned} \quad \text{(C.40)}$$

where we define $\mathbf{A} = \mathbf{W}^\top \mathbf{W} \in \mathbb{R}^{p \times p}$. The square-root matrix of the minimizer gives $\mathbf{W}$ that obeys $\|\mathbf{W}\bar{\boldsymbol{\mu}}\|^2 = T$ and achieves the minimum of the inner minimization problem.

The dual problem of (C.40) is given by

$$\begin{aligned} \max_{\lambda_1, \lambda_2 \in \mathbb{R}} \quad & -\lambda_1 - \lambda_2 T \\ \text{subject to} \quad & \boldsymbol{v}_{\text{aug}} \boldsymbol{v}_{\text{aug}}^\top + \lambda_1 \mathbf{I}_p + \lambda_2 \bar{\boldsymbol{\mu}} \bar{\boldsymbol{\mu}}^\top \succeq 0. \end{aligned} \quad \text{(C.41)}$$

By duality (in particular, complementary slackness), we know that the optimal $\mathbf{A}$ and λ_1, λ_2 obey

$$\left\langle \boldsymbol{v}_{\text{aug}} \boldsymbol{v}_{\text{aug}}^\top + \lambda_1 \mathbf{I} + \lambda_2 \bar{\boldsymbol{\mu}} \bar{\boldsymbol{\mu}}^\top, \mathbf{A} \right\rangle = 0\,. \quad \text{(C.42)}$$

For now, we will assume that $a > 0$; the case $a = 0$ (i.e., $|\langle \bar{\boldsymbol{\mu}}, \boldsymbol{v}_{\text{aug}} \rangle| = 1$) will be discussed later. We consider two separate cases: (1) the optimal $\lambda_1 = 0$, and (2) the optimal $\lambda_1 > 0$.

First, when $\lambda_1 = 0$, it is easy to check that the optimal choice of λ_2 is also 0, which yields the optimal objective 0. Second, when $\lambda_1 > 0$, we claim that the optimal $\mathbf{A}$ is a rank-one

matrix, and it is given by

$$\mathbf{A} = (\sqrt{T}\,\bar{\boldsymbol{\mu}} - \sqrt{1-T}\,\bar{\boldsymbol{\mu}}_\perp)(\sqrt{T}\,\bar{\boldsymbol{\mu}} - \sqrt{1-T}\,\bar{\boldsymbol{\mu}}_\perp)^\top, \tag{C.43}$$

where $\bar{\boldsymbol{\mu}}_\perp = a\boldsymbol{v}_{\text{aug}} - r\boldsymbol{v}_\perp$. To see this, we note that $\boldsymbol{v}_{\text{aug}}\boldsymbol{v}_{\text{aug}}^\top + \lambda_1\mathbf{I}_p + \lambda_2\bar{\boldsymbol{\mu}}\bar{\boldsymbol{\mu}}^\top$ has rank at least $p-2$ since it is two rank-one updates of $\lambda_1\mathbf{I}_p$. In fact, we will show that this matrix has rank exactly $p-1$. Otherwise, we must have

$$\begin{aligned}
\boldsymbol{v}_{\text{aug}}^\top\Big(\boldsymbol{v}_{\text{aug}}\boldsymbol{v}_{\text{aug}}^\top + \lambda_1\mathbf{I}_p + \lambda_2\bar{\boldsymbol{\mu}}\bar{\boldsymbol{\mu}}^\top\Big)\boldsymbol{v}_{\text{aug}} = 0,\\
\bar{\boldsymbol{\mu}}^\top\Big(\boldsymbol{v}_{\text{aug}}\boldsymbol{v}_{\text{aug}}^\top + \lambda_1\mathbf{I}_p + \lambda_2\bar{\boldsymbol{\mu}}\bar{\boldsymbol{\mu}}^\top\Big)\bar{\boldsymbol{\mu}} = 0.
\end{aligned}$$

These are equivalent to

$$\begin{aligned}
1 + \lambda_1 + \lambda_2\langle\bar{\boldsymbol{\mu}}, \boldsymbol{v}_{\text{aug}}\rangle^2 = 0,\\
\lambda_1 + \lambda_2 + \langle\bar{\boldsymbol{\mu}}, \boldsymbol{v}_{\text{aug}}\rangle^2 = 0,
\end{aligned}$$

which implies $a^2\cdot(1-\lambda_2) = 0$. This further implies $\lambda_2 = 1$ since $a > 0$. However, we know that when $\lambda_1 > 0$, the optimal λ_2 would not be 1, as $\lambda_2 = 0$ is strictly better. Therefore, we have shown that $\boldsymbol{v}_{\text{aug}}\boldsymbol{v}_{\text{aug}}^\top + \lambda_1\mathbf{I}_p + \lambda_2\bar{\boldsymbol{\mu}}\bar{\boldsymbol{\mu}}^\top$ has rank $p-1$. By (C.42) primal optimal $\mathbf{A}$ is a rank-one matrix, which together with the primal constraint $\bar{\boldsymbol{\mu}}^\top\mathbf{A}\bar{\boldsymbol{\mu}} = T$ yields the conclusion that the optimal $\mathbf{A}$ is uniquely determined by (C.43).

As a result, the optimal objective value of (C.40) is

$$\|\mathbf{W}\boldsymbol{v}_{\text{aug}}\|^2 = \langle\sqrt{T}\,\bar{\boldsymbol{\mu}} - \sqrt{1-T}\,\bar{\boldsymbol{\mu}}_\perp, \boldsymbol{v}_{\text{aug}}\rangle^2 = \left(r\sqrt{T} - a\sqrt{1-T}\right)^2.$$

As this is always than 0 (the optimal objective when $\lambda_1 = 0$), we conclude that the maximal objective of (C.41) is $\langle\sqrt{T}\,\bar{\boldsymbol{\mu}} - \sqrt{1-T}\,\bar{\boldsymbol{\mu}}_\perp, \boldsymbol{v}_{\text{aug}}\rangle^2 = \left(r\sqrt{T} - a\sqrt{1-T}\right)^2$.

In the corner case $a = 0$, trivially, the objective is a constant: $\boldsymbol{v}_{\text{aug}}^\top \mathbf{A} \boldsymbol{v}_{\text{aug}} = \bar{\boldsymbol{\mu}}^\top \mathbf{A} \bar{\boldsymbol{\mu}} = T$, which is the same as $\left(r\sqrt{T} - a\sqrt{1-T}\right)^2$ in this case.

In all, under the constraint $a^2 < T \leq 1$, the original minimization problem (C.38) becomes

$$\begin{aligned}\min_{1 \geq T > a^2} \quad & -\frac{1}{\tau\left[(1+\sigma_{\text{aug}}^2) + \|\boldsymbol{\mu}\|^2 T + \rho_{\text{aug}}\sigma_{\text{aug}}^2 (r\sqrt{T} - a\sqrt{1-T})^2\right]} \\ & + \log\left(1 + \exp\left(-\frac{2\|\boldsymbol{\mu}\|^2 T}{\tau\left[(1+\sigma_{\text{aug}}^2) + \|\boldsymbol{\mu}\|^2 T + \rho_{\text{aug}}\sigma_{\text{aug}}^2 (r\sqrt{T} - a\sqrt{1-T})^2\right]}\right)\right).\end{aligned}$$

Step 3: combining two cases. Recall the notation $x_+ = \max\{x, 0\}$. Introduce the unified loss $\ell(T)$ that subsumes both cases:

$$\begin{aligned}\ell(T) = & -\frac{1}{\tau\left[(1+\sigma_{\text{aug}}^2) + \|\boldsymbol{\mu}\|^2 T + \rho_{\text{aug}}\sigma_{\text{aug}}^2 [(r\sqrt{T} - a\sqrt{1-T})_+]^2\right]} \\ & + \log\left(1 + \exp\left(-\frac{2\|\boldsymbol{\mu}\|^2 T}{\tau\left[(1+\sigma_{\text{aug}}^2) + \|\boldsymbol{\mu}\|^2 T + \rho_{\text{aug}}\sigma_{\text{aug}}^2 [(r\sqrt{T} - a\sqrt{1-T})_+]^2\right]}\right)\right).\end{aligned}$$

Notice that $\ell(T) \geq \ell^*(T)$ for each $T \in [0, 1]$ where

$$\ell^*(T) := -\frac{1}{\tau\left[(1+\sigma_{\text{aug}}^2) + \|\boldsymbol{\mu}\|^2 T\right]} + \log\left(1 + \exp\left(-\frac{2\|\boldsymbol{\mu}\|^2 T}{\tau\left[(1+\sigma_{\text{aug}}^2) + \|\boldsymbol{\mu}\|^2 T\right]}\right)\right),$$

and $\ell(T) = \ell^*(T)$ for $T \in [0, a^2]$. In addition, the two functions satisfy that

$$\ell^*(a^2) = \ell(a^2), \qquad (\ell^*)'(a^2) = \ell'(a^2).$$

If $\tau \le \tau_1^*$, we have seen in Step 1 that the minimizer T^* of $\ell^*(T)$ is given by (C.39) and $T^* \in [0, a^2]$. Since $\ell(T) \ge \ell^*(T)$ and equality holds if $T \le a^2$, it verifies that T^* is also the minimizer of $\ell(T)$.

If $\tau > \tau_1^*$, the analysis in Step 1 shows that for $T \in [0, a^2]$, $(\ell^*)(T)$ (and thus $\ell(T)$) is monotonically decreasing. Further, denote

$$h(T) = (1 + \sigma_{\text{aug}}^2) + \|\boldsymbol{\mu}\|^2 T + \rho_{\text{aug}} \sigma_{\text{aug}}^2 (r\sqrt{T} - a\sqrt{1-T})^2,$$

which is an increasing function of T when $T \ge a^2$. The derivative of $\ell(T)$ takes the following form:

$$\ell'(T) = \frac{1}{\tau} \frac{h'(T)}{h^2(T)} - \frac{2\|\boldsymbol{\mu}\|^2}{\tau} \frac{\exp\left(-\frac{2\|\boldsymbol{\mu}\|^2 T}{\tau h(T)}\right)}{1 + \exp\left(-\frac{2\|\boldsymbol{\mu}\|^2 T}{\tau h(T)}\right)} \frac{h(T) - T h'(T)}{h^2(T)}.$$

If $a \neq 0$ and $\rho_{\text{aug}} > 0$, we have $\lim_{T \to 1-} h'(T) = +\infty$, then $\ell'(1-) > 0$, which indicates that the minimizer is attained in the interval $(a^2, 1)$. (Note that $T^* = 1$ is impossible!)

Step 4: degenerate cases. If $\rho_{\text{aug}} = 0$, our analysis is reduced to the homogeneous case, where $T^* = 1$ if $\tau > \tau^*$.

If $a = 0$ and $\rho_{\text{aug}} > 0$, we have $\bar{\boldsymbol{\mu}} = \pm \boldsymbol{v}_{\text{aug}}$ and

$$\ell(T) = -\frac{1}{\tau\left[1 + \sigma_{\text{aug}}^2 + LT\right]} + \log\left(1 + \exp\left(-\frac{2\|\boldsymbol{\mu}\|^2 T}{\tau\left[1 + \sigma_{\text{aug}}^2 + LT\right]}\right)\right)$$

where $L = \|\boldsymbol{\mu}\|^2 + \rho_{\text{aug}} \sigma_{\text{aug}}^2$. Th analysis of the $\ell(T)$ is similar to the homogeneous case after a change of variables.

C.4.3 Proof of Corollary 4.6.3

As a consequence of Theorem 4.6.2, most of the statements in Corollary 4.6.3 are straightforward. Below we will focus on Item 2 and 3. Let $S_k = \sum_{j \le k} \langle \boldsymbol{v}_j, \bar{\boldsymbol{\mu}} \rangle^2$ for the k-th cumulative

score for $\boldsymbol{\mu}$, where $1 \le k \le p-1$.

First, by the ordering of singular values, we have

$$(k+1)\sigma_{k+1}^2 \le \|\mathbf{W}\|_F^2 = 1 \qquad \Rightarrow \qquad \sigma_{k+1} \le \frac{1}{\sqrt{k+1}}\,.$$

Therefore, we have

$$\sum_{j>k} \sigma_j^2 \langle \boldsymbol{v}_j, \bar{\boldsymbol{\mu}} \rangle^2 \le \sigma_{k+1}^2 \cdot \sum_{j>k} \langle \boldsymbol{v}_j, \bar{\boldsymbol{\mu}} \rangle^2 \le \frac{1}{k+1}(1 - S_k).$$

We thus obtain an upper bound on $\|\mathbf{W}\bar{\boldsymbol{\mu}}\|^2$:

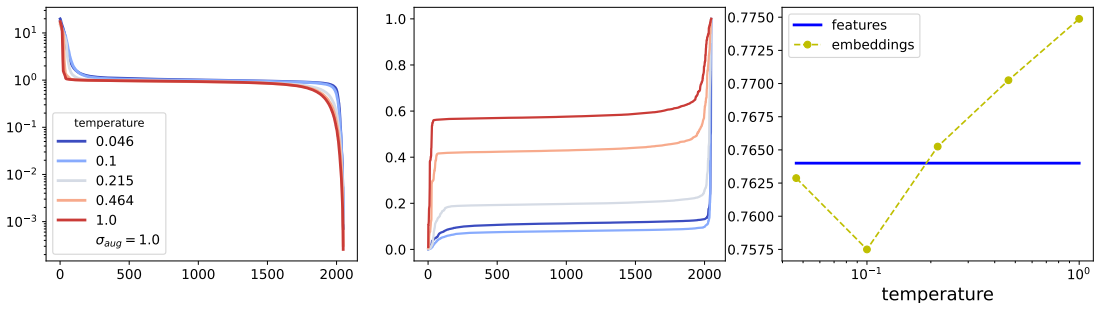
$$\begin{aligned} \|\mathbf{W}\bar{\boldsymbol{\mu}}\|^2 &= \sum_{j \le k} \sigma_j^2 \langle \boldsymbol{v}_j, \bar{\boldsymbol{\mu}} \rangle^2 + \sum_{j>k} \sigma_j^2 \langle \boldsymbol{v}_j, \bar{\boldsymbol{\mu}} \rangle^2 \\ &\le \sum_{j \le k} \langle \boldsymbol{v}_j, \bar{\boldsymbol{\mu}} \rangle^2 + \frac{1}{k+1}(1 - S_k) \\ &\le \frac{k}{k+1} S_k + \frac{1}{k+1}. \end{aligned}$$

If $\|\mathbf{W}\bar{\boldsymbol{\mu}}\|^2 \ge c$ for certain constant $c > 0$, we can choose $k \ge 1$ to the smallest integer with $k \ge (2c)^{-1} - 1$. Combining the upper bound and lower bound, we obtain $S_k \ge c/4$. This proves the first jump of the cumulative score. The second jump is because $S_p - S_{p-1} = \langle \boldsymbol{v}_p, \bar{\boldsymbol{\mu}} \rangle^2 = r^2 > 0$ due to nondegeneracy.

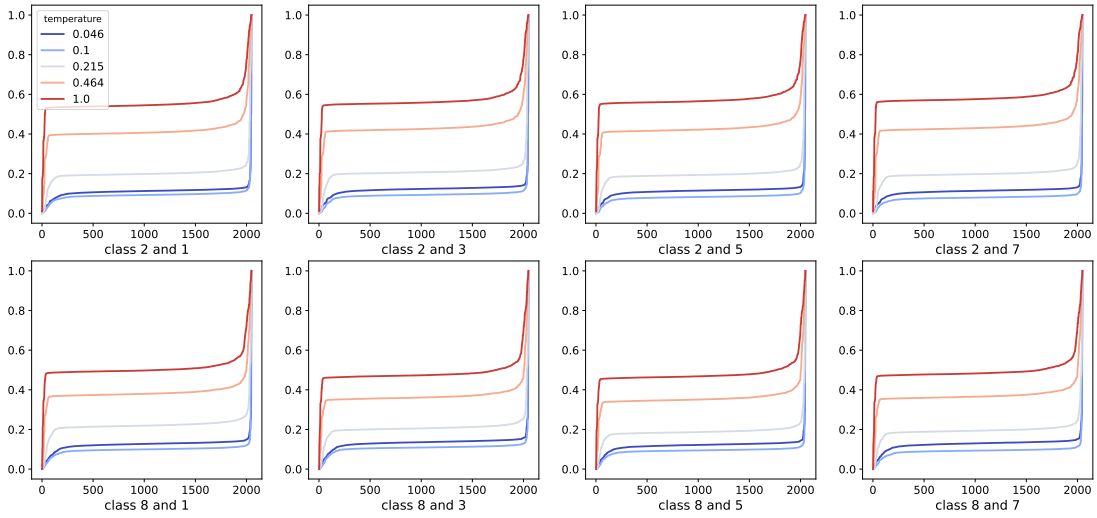
Moreover, if $\tau > \tau_1^*$, the top right singular vector of $\mathbf{W}$ is $\boldsymbol{v}_1 = \sqrt{T^*}\bar{\boldsymbol{\mu}} - \sqrt{1-T^*}\boldsymbol{\mu}_\perp$ where $T^* > \sqrt{1-r^2}$. Since

$$\langle \boldsymbol{v}_1, \boldsymbol{v}_{\text{aug}} \rangle = r\sqrt{T^*} - \sqrt{1-r^2}\,\sqrt{1-T^*}$$

is monotone increasing in T^*. The inner product must be positive if $\tau > \tau_1^*$. Also, $\langle \boldsymbol{v}_1, \bar{\boldsymbol{\mu}} \rangle = \sqrt{T^*} > 0$, so this inner product is also positive.



(a)



(b)

Figure C.2: Results with the pretrained encoder and a one-layer linear projector $\mathbf{W} \in \mathbb{R}^{2048\times2048}$ under the standard SimCLR loss (ResNet-50 and 10-class STL-10 dataset). **(a) Left**: singular values of $\mathbf{W}$ with varying temperature. **(a) Middle**: cumulative sums of alignment scores $\kappa_{j,c_1,c_2} = \langle \boldsymbol{v}_j, \boldsymbol{\mu}_{c_1,c_2}\rangle^2/\|\boldsymbol{\mu}_{c_1,c_2}\|^2$ for the pair $(c_1, c_2) = (7, 2)$. **(a) Right**: comparison between downstream task accuracy with features and embeddings for 10-class evaluation. **(b)** Cumulative sums of alignment scores $\kappa_{j,c_1,c_2} = \langle \boldsymbol{v}_j, \boldsymbol{\mu}_{c_1,c_2}\rangle^2/\|\boldsymbol{\mu}_{c_1,c_2}\|^2$ for different pairs.

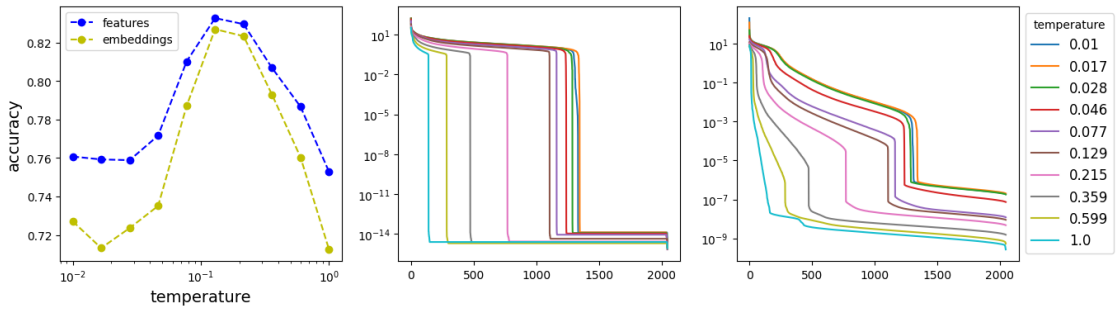


Figure C.3: Results for SimCLR models trained from scratch. **Left**: comparison between classification accuracy with features (before projection) and embeddings (after projection). **Middle**: singular values of feature matrix $\mathbf{H}_0 = [\boldsymbol{h}_{1,0}, \ldots, \boldsymbol{h}_{n,0}]$ with varying temperatures. **Right**: singular values of embedding matrix $\mathbf{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n]$ with varying temperatures.

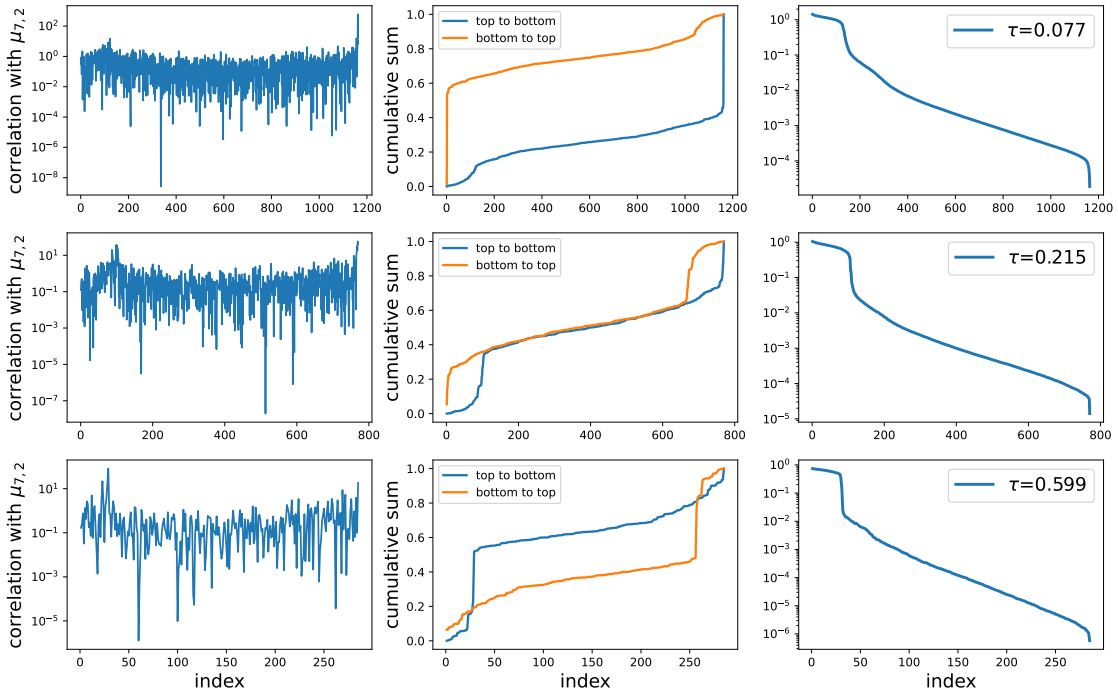


Figure C.4: Plots for train-from-scratch experiment with projector $\mathbf{W}^{(r)} = \mathbf{W}\mathbf{V}_r(\mathbf{H}_0) \in \mathbb{R}^{p\times r}$ and $\boldsymbol{\mu}^{(r)} = \mathbf{V}_r^\top(\mathbf{H}_0)\boldsymbol{\mu} \in \mathbb{R}^r$ which are both projected onto feature matrix $\mathbf{H}_0$'s top-r right singular subspace with $r = \text{argmax}k : s_k(\mathbf{H}_0) \geq 10^{-3}\}$. **Left**: unsquared alignment scores $\langle \boldsymbol{v}_j^{(r)}, \boldsymbol{\mu}_{c_1,c_2}^{(r)} \rangle / \|\boldsymbol{\mu}_{c_1,c_2}^{(r)}\|$ for the pair $(c_1, c_2) = (7, 2)$. **Middle**: cumulative sum of alignment scores. **Right**: singular values of $\mathbf{W}^{(r)}$.

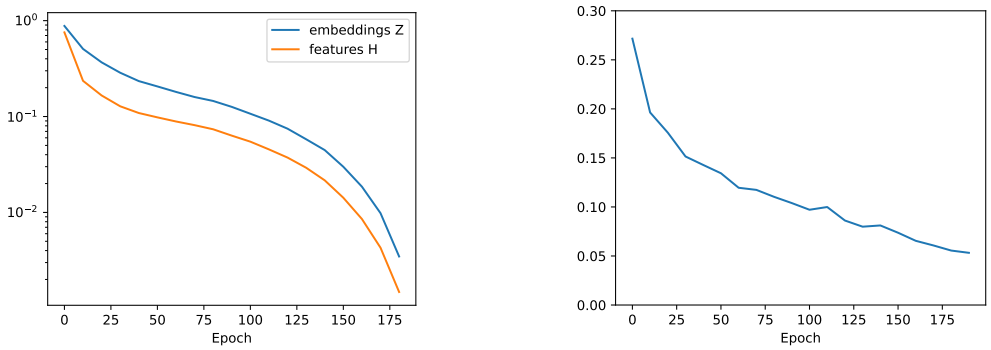


Figure C.5: **Left** the relative change of features and embeddings: $\|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}\|/\|\boldsymbol{\theta}^{(t)}\|$ and $\|\mathbf{W}^{(t+1)} - \mathbf{W}^{(t)}\|/\|\mathbf{W}^{(t)}\|$; **Right** the standardized distance between $\mathbf{W}^{(t)}$ and the minimizer $\widetilde{\mathbf{W}}^{(t)}$: $\|\widetilde{\mathbf{W}}^{(t)} - \mathbf{W}^{(t)}\|/\|\mathbf{W}^{(t)}\|$.

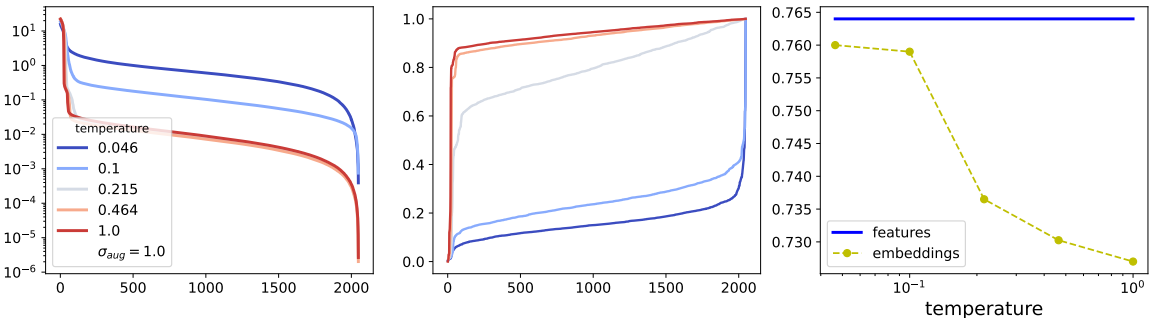


Figure C.6: Results with the pretrained encoder and a one-layer linear projector $\mathbf{W} \in \mathbb{R}^{2048\times 2048}$ under the standard SimCLR loss (ResNet-50, 10-class `STL-10` dataset and feature-level perturbation with $\sigma_{\text{aug}} = 1.0$). **Left**: singular values of $\mathbf{W}$ with varying temperature. **Middle**: cumulative sums of alignment scores $\kappa_{j,c_1,c_2} = \langle \boldsymbol{v}_j, \boldsymbol{\mu}_{c_1,c_2}\rangle^2/\|\boldsymbol{\mu}_{c_1,c_2}\|^2$ for the pair $(c_1, c_2) = (7, 2)$. **Right**: comparison between downstream task accuracy with features and embeddings for 10-class evaluation.